Hierarchical Reinforcement Learning, Sequential Behavior, and the Dorsal Frontostriatal System

Miriam Janssen*, Christopher LeWarne*, Diana Burk, and Bruno B. Averbeck

Abstract

■ To effectively behave within ever-changing environments, biological agents must learn and act at varying hierarchical levels such that a complex task may be broken down into more tractable subtasks. Hierarchical reinforcement learning (HRL) is a computational framework that provides an understanding of this process by combining sequential actions into one temporally extended unit called an option. However, there are still open questions within the HRL framework, including how options are formed and how HRL mechanisms might be realized within the brain. In this review, we propose that the existing human motor sequence literature can aid in understanding both of these questions. We give specific emphasis to visuomotor sequence learning tasks such as the discrete sequence production task and the M × N (M steps × N sets) task to understand how hierarchical learning and behavior manifest across sequential action tasks as well as how the dorsal cortical–subcortical circuitry could support this kind of behavior. This review highlights how motor chunks within a motor sequence can function as HRL options. Furthermore, we aim to merge findings from motor sequence literature with reinforcement learning perspectives to inform experimental design in each respective subfield. ■

INTRODUCTION

Reinforcement learning (RL) provides a theoretical framework for understanding how organisms learn to make choices to satisfy their needs (Averbeck & Murray, 2020). Within the standard RL framework, agents learn the values of actions through experience (Sutton & Barto, 1998). Substantial progress has been made toward understanding the computational mechanisms and neural circuits underlying RL (Averbeck & O'Doherty, 2021; Neftci & Averbeck, 2019; Averbeck & Costa, 2017; Lee, Seo, & Jung, 2012). Most studies of RL have used bandit tasks, in which probabilistic outcomes are associated with a limited set of two to three possible actions and a single action leads to an outcome (Costa, Tran, Turchi, & Averbeck, 2015; O'Doherty et al., 2004). However, in realistic situations where the set of available actions is large and multiple steps are required to reach a goal, standard RL is limited. Goal-directed behavior typically requires coordination of movement sequences that unfold over long timescales and that can be described at multiple levels of action granularity. For example, enjoying a cup of coffee can be described at multiple levels of abstraction from drinking coffee, at the highest level, to the complex sequence of context-specific muscle activations required to press a button, at a low level (Figure 1).

Learning the sequence of context- or state-dependent muscle activations required to make a cup of coffee would be complex with standard RL models, which would learn the value of each muscle activation in a state-dependent way. This increase in complexity at less abstract levels of task specification leads to a scaling problem often known as the curse of dimensionality, because the number of states and possible muscle activations is very large. Standard or flat RL models do not do well with complex or high-dimensional state spaces as they must experience each state many times to learn action values in each state. A promising way to deal with the scaling problem is to apply RL at different levels of abstraction. This approach is known as hierarchical RL (HRL). In HRL, levels of abstraction can be organized as options, which are sets of sequential actions that are grouped together (Barto & Mahadevan, 2003; Sutton, Precup, & Singh, 1999). The formation of options lowers the dimensionality of the problem. Learning can then be carried out at the level of options. Options, therefore, mitigate the scaling problem by reducing the computational demand required by standard RL. To do so, however, an HRL agent must learn which actions to group together as options—that is, they must solve the option discovery problem (Botvinick, Niv, & Barto, 2009).

HRL is useful for learning at higher levels of abstraction in real-world tasks, such as making a cup of coffee. These tasks have subtasks that exist at varying levels of abstraction. Each level of abstraction may affect the execution at other levels by reducing initiation costs, which can be defined by increased RTs. A recent finding points toward a mixed hierarchical model where superordinate

National Institute of Mental Health, Bethesda, MD
*These authors contributed equally.

Journal of Cognitive Neuroscience 34:8, pp. 1307–1325
https://doi.org/10.1162/jocn_a_01869

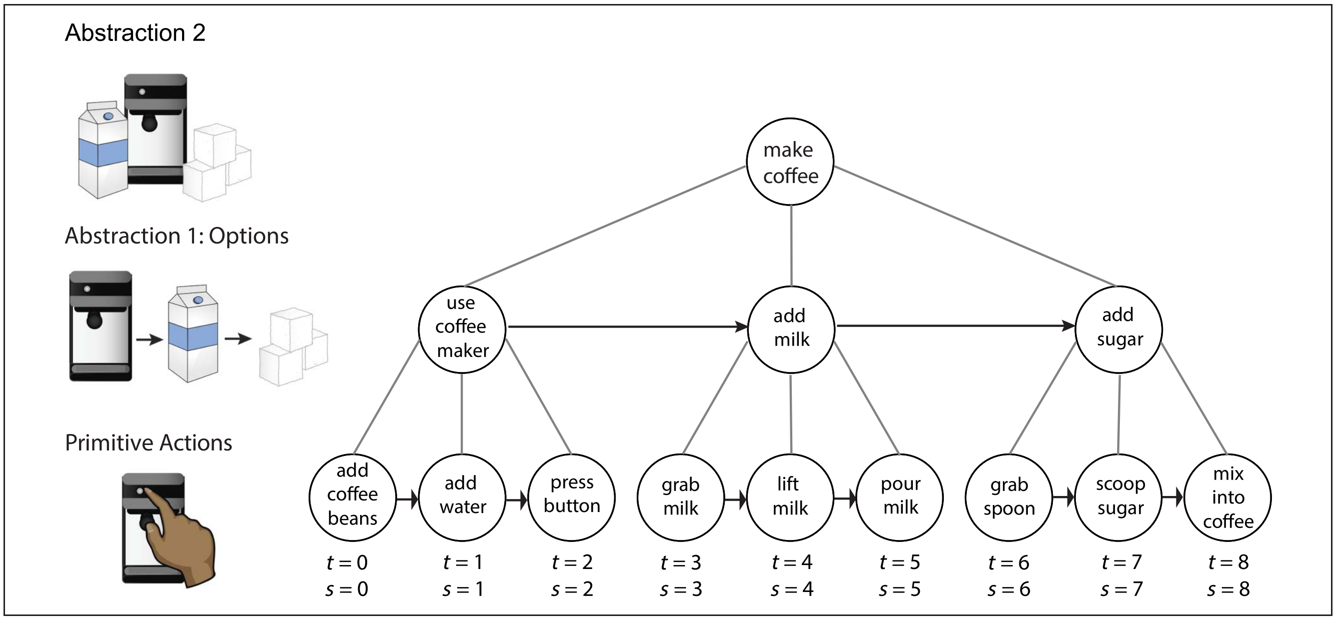

Figure 1. Level of abstractions across time. The top level, Abstraction 2, consists of one chunked action of making coffee. The second level, Abstraction 1, consists of less abstract chunked actions. The third level consists of examples of primitive actions. Although actions in Levels 2 and 3 consist of three actions in this figure, they can be composed of one or more actions. The primitive actions are executed over eight time steps and eight states.

tasks can influence lower-level actions and subordinate actions can influence superordinate tasks. For example, if options share motor sequence elements, improving performance within one option can transfer to the other option, via the overlapping elements.

Specifically, there is support for a strict or unidirectional relationship between superordinate task sequence goals (e.g., making coffee) and options (e.g., using a coffee maker). However, it seems there is a nonstrict relationship between options and primitive motor actions (e.g., grabbing milk) as well as motor actions and task goals (Trach, McKim, & Desrochers, 2021). There is evidence that these hierarchical control structures, such as options or task sequence goals, show improvements with motor-level practice (Trach et al., 2021; Yokoi & Diedrichsen, 2019). Despite the utility of this framework, understanding the complex behavior and neural underpinnings involved in discovering useful options is lacking.

There is considerable evidence that biological agents organize behavior hierarchically, often in the form of action sequences (Eckstein & Collins, 2020; Yokoi & Diedrichsen, 2019; Botvinick & Weinstein, 2014; Seidler, 2010; Botvinick et al., 2009; Rhodes, Bullock, Verwey, Averbeck, & Page, 2004; Marsden, 1982). Through such hierarchical and sequential organization of behavior, biological agents appear to solve the option discovery problem. Biological agents also utilize chunks, which are a series of actions abstracted into a single unit, to execute learned action sequences efficiently, similar to how options in HRL allow efficient traversal of complex state spaces. Chunks in behavior and options in RL both aid learning and may serve a similar function of breaking down complex problems. Therefore, by considering research on the sequential organization of behavior in biological systems and recasting it under the framework of HRL, we aim to answer two main questions. First, is it possible to recognize the concatenation of actions in sequence learning, such as chunks, in a way that is consistent with HRL, thus providing insight into the option discovery problem? Second, because the constituents of the dorsal frontostriatal system are likely involved in action selection and motor sequence learning (Averbeck & Murray, 2020; Martiros, Burgess, & Graybiel, 2018; Averbeck & Costa, 2017), does the dorsal system specialize in learning the hierarchical organization of action sequences that allows flexible execution of actions to achieve goals?

In this review, we first give an overview of HRL and discuss the behavioral literature on the sequential and hierarchical organization of behavior to elucidate the link between chunks of actions and HRL options. Previous reviews on the utility of hierarchical behavior for biological agents have emphasized the role of cortico-striatal feedback loops in hierarchical action control (Badre & Nee, 2018; Balleine, Dezfouli, Ito, & Doya, 2015). These loops, which incorporate secondary motor areas like the pre-SMA, may allow for overcoming issues of dimensionality while also explaining the formation of action sequences and chunks. Here, we expand upon these reviews by looking to motor sequence learning tasks to gain a deeper understanding of how biological agents learn and use action sequences and chunks. We also discuss how these chunks can aid in learning and performance and how they are conceptually analogous to options in HRL, which also facilitate learning (Garr, 2019; Botvinick et al., 2009). Furthermore, we characterize the involvement of cortico-striatal circuits using a recent anatomical, computational, and behavioral framework for motivated behavior (Averbeck & Murray, 2020). Using this framework, we present an augmented understanding of the details concerning cortico-striatal feedback loops and suggest specific functions for each brain region within them.

Next, we discuss neural evidence for option formation and representation of sequences as a hierarchical unit, as opposed to their constituent actions. We compare neural sequence task bracketing with option initiation and termination functions as well as neural outcome and cost representations with the option framework's pseudo-reward prediction errors (pseudo-RPEs). A pseudo-RPE is an option-specific reward prediction error (RPE) that updates an option-specific policy and state value. Third, we discuss the recent proposal that RL maps onto two large-scale neural systems (Averbeck & Murray, 2020), and we propose that their distinction might be clarified by incorporating HRL. Thus, we propose that the existing literature on sequential behaviors may both be understood in terms of HRL and be useful in understanding how biological agents solve the option discovery problem.

Hierarchical Reinforcement Learning

We begin with a brief overview of HRL (Box 1). For additional reviews on HRL in systems neuroscience that cover computational mechanisms in more detail, see Botvinick (2012) and Botvinick et al. (2009). Within the RL framework, learning is realized by optimizing a policy, which is a function that maps from states to actions. An optimal policy selects actions with the highest value in each state. States, or contexts, are specified by a combination of internal and external environmental variables that contain the information necessary to select an action.

Box 1. Computational HRL

HRL seeks to apply the well-known framework of RL—from machine learning, neuroscience, and psychology—to the diverse and complex actions that are routine for human learners. The objective for a standard RL agent is to maximize long-term reward given certain action possibilities. This entails a set of state spaces, a set of possible actions to take, a transition function representing a probability of changing from one state to another given a certain action, and a reward function that calculates the rewards obtained through a given transition. An RPE can be computed, which is the difference between the expected reward and the actual obtained reward. The RPE, $\delta(t)$, is given by

$$\delta(t) = R(t) - v_i(t) \tag{1}$$

where $R(t)$ is the reward received and $v_i(t)$ is the expected reward for action or state i. A highly useful algorithm incorporating RPEs for learning is the Rescorla–Wagner algorithm:

$$v_i(t+1) = v_i(t) + \rho\delta(t) \tag{2}$$

where the updated value estimate, $v_i(t + 1)$, is equal to the previous value estimate, $v_i(t)$, plus the scaled RPE with learning rate ρ.

These equations, however, do not distinguish between state and values. Temporal-difference (TD) RL algorithms, in contrast, differentiate state values, $v(s_t)$, and action values, $q(s_t, a)$. States are generally defined by the available information that can be used to predict upcoming rewards. In TD algorithms, the RPEs are computed as differences between accumulated rewards and the value of the next state compared to values of previous states.

$$\delta = r_{t+1} - v(s_t) + \gamma v(s_{t+1}) \tag{3}$$

where state values $v(s_t)$ represent the action with the largest value in each state.

$$v(s_t) = \max_{a\epsilon A_{s_t}}\{q(s_t, a)\} \tag{4}$$

An action value is the value of an action a that is taken in a state s_t, which is computed as the sum of immediate and discounted future rewards.

$$q_{t+1}(s_t, a) = R(s_t, a) + \gamma\sum_{j\epsilon S} p(j|s_t, a)v_{t+1}(j) \tag{5}$$

where $p(j\,|\,s_t, a)$ defines the probability of transitioning into state j from state s_t, given action a.

Although much progress has been made, both theoretically and in practice, there remain serious limitations to the current RL framework, particularly in its application to the learning of complex actions over vast timescales. As the number of state spaces and possible actions increases, the computational complexity required for the RL agent increases exponentially. This is known as the *scaling problem*.

Temporal abstraction across multiple actions and state spaces serves as a potential advance past this problem, where TD algorithms shown above are a representative step in the right direction. Whereas basic RL handles simple actions mapping across state spaces, constituting an agent's policy, HRL posits options, where an option entails groups of simple actions that jointly map across more abstract state spaces. Options are temporally abstract, meaning that the initiation and termination of the option spans across multiple time points. In doing so, rather than mapping each simple action to each state space, the HRL agent avoids such computational complexity by abstracting across multiple actions, state spaces, and time points.

Yet, an HRL agent must know how to group actions into options. This is known as the *option discovery problem*.

Formally, an agent may select option o with a probability $P(o)$ according to

$$P(o) = \frac{e^{w_{o_{\text{ctrl}}}(s_t, o)/\tau}}{\sum_{o' \epsilon O} e^{w_{o_{\text{ctrl}}}(s_t, o')/\tau}} \tag{6}$$

where $W_{o_{\text{ctrl}}}(s_t, o)$ denotes the weight of option o in state s_t, under the governing option of o_{ctrl}. A temperature parameter τ controls the tendency toward exploration in action selection.

The RPE for options, δ_o, is akin to that for actions, as seen in Equation (3).

$$\delta_o = r_{\text{cum}} - v_{o_{\text{ctrl}}}(s_{\text{init}}) + \gamma^{t_{\text{tot}}} v_{o_{\text{ctrl}}}(s_{t+1}) \tag{7}$$

where s_{init} is the state wherein the option was selected, t_{tot} is the number of time steps elapsed since s_{init}, and r_{cum} is the cumulative discounted reward for the duration of the option.

Given an RPE, an HRL learner can update option strengths and value functions as the following:

$$v_{o_{\text{ctrl}}}(s_{t_{\text{init}}}) \leftarrow v_{o_{\text{ctrl}}}(s_{t_{\text{init}}}) + \alpha_i \delta_o \tag{8}$$

$$W_{o_{\text{ctrl}}}(s_{t_{\text{init}}}, o) \leftarrow W_{o_{\text{ctrl}}}(s_{t_{\text{init}}}, o) + \alpha_i \delta_o \tag{9}$$

where α_i denotes a given learning rate.

Standard RL, which we refer to as flat RL, is a framework for learning to select actions in Markov decision processes (MDPs). MDPs are composed of an agent that learns a policy by interacting with its environment. Typically, MDPs use a discrete timescale, with actions and state transitions occurring sequentially in time (e.g., Action a_1 is followed by Action a_2). When an agent selects an action in a given state, it transitions to a new state and receives a reward (or not). It then selects an action in the new state, repeating the cycle. Although these methods have proven useful for many problems, flat-RL models are limited in their application because of the scaling problem: As the quantity of states and possible actions in the agent's environment increases, computing the optimal policy becomes computationally intractable (Barto & Mahadevan, 2003; Sutton, 1999). Because RL agents learn the available choice values in each state through experience, when the state space becomes very large, they cannot accumulate enough experience with the actions in each state to learn effectively (Gershman & Daw, 2017). However, in many situations, biological agents can nevertheless learn to execute effective action sequences. How is this achieved?

In HRL, agents solve the scaling problem by applying RL to grouped actions or state representations, which allows the agent to learn efficiently in environments that are too computationally demanding for standard RL (Xia & Collins, 2021; Sutton et al., 1999). The sets of actions that are grouped together in HRL are referred to as options. Once defined, an option groups together the actions entailed by the policy as a single unit (Sutton et al., 1999). An option consists of a policy, initiation conditions, and termination conditions. Initiation and termination conditions refer to the states that specify an option's beginning and ending. For example, the option of adding milk to coffee will initiate once there is a cup of coffee available and terminate once there is sufficient milk in one's coffee.

Although HRL facilitates planning in complex state spaces, this creates the problem of discovering useful options, known as the option discovery problem (Eckstein & Collins, 2021; Tomov, Yagati, Kumar, Yang, & Gershman, 2020; Botvinick et al., 2009; Barto & Mahadevan, 2003). Although biological agents can efficiently navigate complex state spaces, we lack an understanding of the neural underpinnings for this behavior. Recent work on option discovery focused on hierarchical organization of state spaces, or state abstraction (Tomov et al., 2020). Tomov et al. (2020) have proposed a Bayesian model of hierarchy discovery. Their algorithm defines clusters of high-level states by analyzing a graph of the environment. It is assumed within the algorithm that the agent knows all states and edge transitions. Although the edge transitions are already known, the value associated with each state is unknown, and hence optimal options still need to be learned. To do this, the algorithm minimizes a cost function, such that clusters of similar states are identified and efficiently traversed across the graph. Once these state clusters are identified, options that traverse the clusters can be planned and executed. For example, when one plans how to travel from their office to their home, it is useful to start by planning at a high level: leave the office, get on the freeway, exit toward the right neighborhood, and so forth. Importantly, each of these high-level actions is associated with several states that can be intuitively clustered together. When planning how to leave the office, we group all the different rooms, hallways, and staircases in our place of work into a useful simplification: the office. Although this simplification seems obvious, it is useful for planning and is hence leveraged by computational algorithms as well.

Graph environments can inform and account for a wealth of behavioral data from human and animal studies (Peer, Brunec, Newcombe, & Epstein, 2021; Kim et al., 2018). In a human task where, as opposed to the previously described algorithm, an underlying graph structure needs to be learned, participants can learn to detect states that constitute transitions between densely connected clusters or bottleneck transitions (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013), and they can learn to navigate between perceived communities of state clusters as well as find optimal paths to achieve the task at hand (Solway et al., 2014). Thus, humans appear to be capable of learning graph structures, which suggests that the model proposed by Tomov et al., which relies on knowledge of graph structure, can account for state abstraction and option discovery. Other graph partitioning methods, similar to the work done by Tomov et al., have also been developed to abstract across states by identifying useful clusters (Şimşek, Wolfe, & Barto, 2005; Mannor, Menache, Hoze, & Klein, 2004; Menache, Mannor, & Shimkin, 2002). Alternative hierarchical clustering models have also been successful in capturing human participants' ability to identify hierarchical structure within a task (Collins & Frank, 2016). However, models that rely on state abstraction will need to account for action sequence abstraction as well. Future work will need to accommodate both kinds of abstraction to fully understand option discovery.

Using HRL as a computational framework to understand human behavior in complex environments seems to have greater promise over flat RL. Although work is underway identifying the computational mechanisms that make efficient hierarchical planning possible, the option discovery problem needs to be understood biologically as well. Hence, we turn to behavioral studies of sequentially organized behavior. To understand these behavioral studies through the lens of HRL, we compare how biological and computational agents simplify complex action sequences and learn more efficiently through either motor chunks or options, respectively.

Sequential Organization of Behavior: Motor Chunks as HRL Options

Many tasks used to study RL in biological agents require only a single action to achieve a goal (Costa, Dal Monte, Lucas, Murray, & Averbeck, 2016; Ostlund, Winterbauer, & Balleine, 2009). These studies have not been designed to investigate the flexible use of action sequences to achieve a goal. Nevertheless, the motor sequence literature has explored a variety of sequential movement tasks, although only recently have they been examined from the perspective of RL (Balleine & Dezfouli, 2019; Garr, 2019; Desrochers, Amemori, & Graybiel, 2015; Dezfouli & Balleine, 2012; Desrochers, Jin, Goodman, & Graybiel, 2010). In these tasks, action sequences are simplified compared to everyday life because more abstract and complex tasks, such as making coffee or grabbing milk, are difficult to analyze in the laboratory. Thus, the motor sequence tasks tend to use sequences of simple button presses, whereas the abstractions in HRL are usually higher level, like making coffee. The motor sequence tasks used in human participants are more complex than the tasks used in primate or rodent studies. For this reason, we have focused on human participant tasks, although tasks designed for animal models may also provide insight into hierarchical organization. Below, we consider several of the motor sequence paradigms and how they may relate to options within HRL.

Motor Chunk Characterization and Evidence from Devaluation Experiments

Motor chunking is the behavioral process of grouping multiple actions into a single action unit. Sometimes, this happens in the context of a longer sequence, which is divided into multiple smaller chunks (Abrahamse, Ruitenberg, de Kleine, & Verwey, 2013; Verwey, 1996, 2001), and other times, it happens within the context of a shorter series of actions, which are grouped together into a single chunk (Ostlund et al., 2009). Alternatively, with extensive practice, short chunks can lengthen into longer ones for additional behavioral optimization (Ramkumar et al., 2016). These groupings have been referred to by many different terms, including chunks (Abrahamse et al., 2013; Verwey, 1996, 2001), macro-actions (Dezfouli & Balleine, 2012), habits (Martiros et al., 2018; Miller, Ludvig, Pezzulo, & Shenhav, 2018; Desrochers, Burk, Badre, & Sheinberg, 2016; Desrochers et al., 2015; Graybiel & Grafton, 2015; Smith & Graybiel, 2013; Jin & Costa, 2010), or options (Botvinick et al., 2009; Sutton et al., 1999). Behavioral and neural evidence suggests that chunks are treated as a unit, and therefore, chunks can be understood as one form of an option in HRL.

There are many ways to characterize chunks including unique RT patterns (Abrahamse et al., 2013; Verwey, 1996, 2001), decreased variability (Tremblay et al., 2009; Levesque et al., 2007), automaticity (Graybiel, 2008), development of stereotyped actions (Geddes, Li, & Jin, 2018; Desrochers et al., 2010, 2015), and insensitivity to action–outcome contingencies (Balleine & Dezfouli, 2019; Miller et al., 2018; Smith & Graybiel, 2016; Graybiel & Grafton, 2015; Dezfouli & Balleine, 2013; Daw, Gershman, Seymour, Dayan, & Dolan, 2011). With respect to insensitivity to action–outcome contingencies, one way that chunks can be evaluated is through devaluation experiments on operantly conditioned sequences (Garr, 2019; Killcross & Coutureau, 2003; Balleine, Garner, Gonzalez, & Dickinson, 1995). In these studies, rodents are trained to perform a task that involves a series of actions that lead to a reward. After training, the reward is devalued by, for example, feeding the animal to satiation on the reward before the experimental sessions (Balleine & Dickinson, 1992). Chunking is assessed by measuring whether the animals withhold the execution of the entire sequence or only the action preceding reward. One such study found chunking behavior in rats on a sequential lever press task (Ostlund et al., 2009). Interestingly, rats with lesions to the dorsomedial PFC did not show sequence-level suppression, suggesting that they were unable to represent the sequence as a chunk. This study suggested that rodents can chunk action sequences and that there are dedicated neural systems to support chunking. It has also been shown in rodents that the motor cortex is necessary for learning a motor sequence, but once the sequence has been learned, lesions to the motor cortex have no effect on sequence execution (Kawai et al., 2015). This seems to suggest a crucial role for subcortical regions in sequence storage, perhaps with signaling from the dorsomedial PFC also playing a role. However, to substantiate the claims of HRL with evidence from chunking and to show the role of chunks in complex behavior, tasks with more complicated action sequences are required.

Motor Learning and Sequence Chunking in Human Participants: Chunks, Like Options, Aid Efficient Learning

Many tasks have been used to study sequence learning and execution in human participants (Figure 2) including the serial RT (SRT; Nissen & Bullemer, 1987), discrete sequence production (DSP; Abrahamse et al., 2013; Verwey, 2001), and M × N tasks (where *M* denotes set length and *N* is hyperset or sequence; Sakai, Kitaguchi, & Hikosaka, 2003; Hikosaka, Rand, Miyachi, & Miyashita, 1995). Although the SRT task provides important insights on procedural or skill learning, chunking is often not observed unless it is induced through uniform segmentation (Jiménez, Méndez, Pasquali, Abrahamse, & Verwey, 2011) or unless training is very extensive (Verstynen et al., 2012), which is thought to be because of the lack of sufficient practice with a repeated, discrete sequence (Abrahamse et al., 2013). The task, however, allows for the isolation of stimulus–response (S-R) latency and accuracy improvements without chunking. In the SRT task, average RT decreases, whereas response accuracy increases across training with fixed sequences (Nissen & Bullemer, 1987).

The differences between the tasks allow for understanding which conditions lead to chunks (e.g., DSP and M × N) and which typically do not (e.g., SRT). In these tasks, motor chunks are often reflected in less variable interresponse intervals (IRIs) within chunks than between chunks (Sakai et al., 2003; Verwey, 1996, 2001). In addition, decreased RT and decreased total execution time have been cited as evidence for hierarchical organization of movements, such as chunks (Rhodes et al., 2004). Even after 10 days of training, human participants can use chunks to continuously decrease RTs across training (Verstynen et al., 2012). Hence, chunking actions together allows for an agent to act quickly and efficiently across a complex set of movements as is seen in HRL (Nachum, Gu, Lee, & Levine, 2018).

M × N Task: Demonstrating That Sequence Chunks Can Aid Learning

The M × N task has provided evidence for the formation of motor chunks. The M × N task (Hikosaka et al., 1995) involves trial-and-error-based S-R learning (Figure 2B). Typically, participants are presented with two stimuli (*M* = 2) simultaneously on a button pad and must learn the correct pressing order to complete the set. The number of sets in a hyperset is typically 10 (*n* = 10), amounting to many actions that the participant needs to learn. Although RTs across all sets decrease during training, patterns of RTs can emerge for contiguous sets, suggesting chunking (Sakai et al., 2003).

It has further been shown that participants use chunks within the M × N task to learn more efficiently and perform more accurately. After participants had formed chunks on a certain hyperset, they were given a novel hyperset to learn. Some participants undertook hypersets that contained contiguous sets from the previous hyperset that had been chunked, whereas others did not. Unsurprisingly, those participants with hypersets containing previously learned chunks learned more efficiently (Sakai et al., 2003). Alternatively, some participants were given the same sets they learned in the original hyperset, but they were shuffled so that their previously learned chunks would no longer be useful. This led to significantly diminished performance, suggesting that the use of chunks allowed for greater performance. The observed performance benefits from retaining learned chunks are similar to efficiency gains found with options in HRL (Botvinick et al., 2009; Barto & Mahadevan, 2003; Sutton et al., 1999). As chunks allow for more efficient

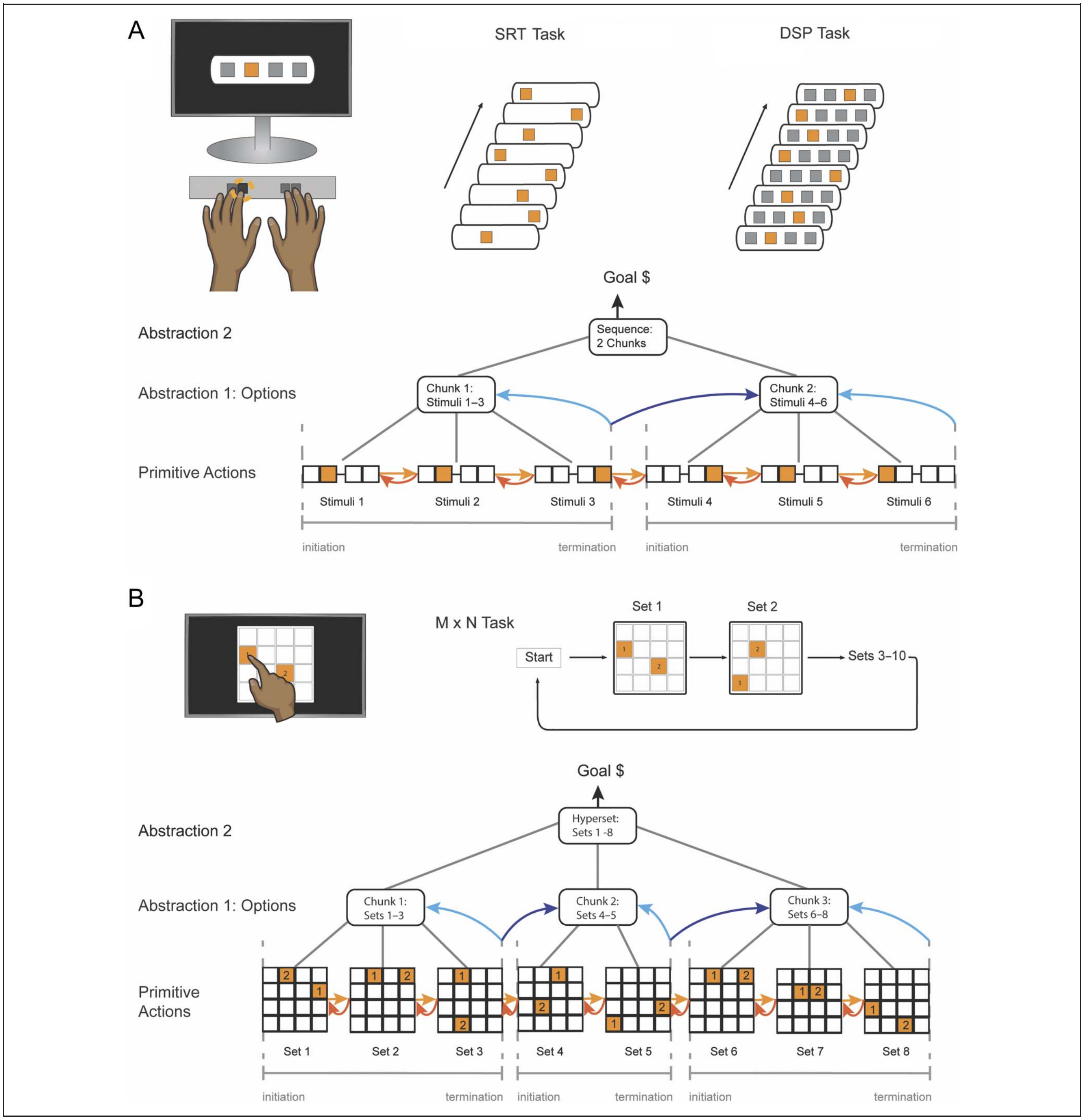

Figure 2. Common task designs for motor sequence learning. (A) SRT and DSP tasks. Participants are presented with visual stimuli, and they respond by pressing the associated key or button on a response pad. In SRT, one stimulus is presented at a time, and sequences can last for a varying number of stimuli across experiments. Sequences can be fixed, where participants see the same sequence repeatedly, or random, in which a new sequence is seen on each trial. In DSP, stimuli are presented by illuminating one of the placeholders presented on the screen. Sequences are typically six to eight elements in length and repeated 500–1000 times per sequence. The schematic below depicts a hierarchical diagram for the DSP task (Abrahamse et al., 2013) using the HRL framework. Each sequence consists of two subsequences or chunks, which concatenate through an option-specific policy (dark blue arrow) and are indicated by initiation and termination states. The value of the option-specific policy is updated by a pseudo-RPE (light blue arrow). The light orange arrow represents a standard policy, and the dark orange arrow represents an RPE. The goal or reward earned at the end of this task is money. (B) M × N task: M = number of items in a set, N = number of sets. Participants typically press on simultaneously presented, illuminated squares on a 4 × 4 button pad (Hikosaka et al., 1995). Participants learn response order through trial and error. The schematic below depicts a hierarchical diagram for the M × N task (Sakai et al., 2003) using the HRL framework. Each hyperset consists of three subsets or chunks, which concatenate through an option-specific policy (dark blue arrow), and each option is indicated by initiation and termination states. The value of the option-specific policy is updated by a pseudo-RPE (light blue arrow). The light orange arrow represents a standard policy, and the dark orange arrow represents an RPE. The goal or reward earned at the end of this task is money.

learning and performance, we can see that options and chunks serve synonymous functions for computational or biological agents, respectively.

DSP Task: Demonstrating That Sequence Chunks Can Aid Efficiency

The DSP task was developed to study chunking. In this task, participants typically learn to execute a short motor sequence as fast as possible. The DSP task differs from other tasks because it uses shorter sequences, extensive training, and spatially defined key-specific stimuli (Abrahamse et al., 2013). Because of the limited sequence length and extensive training (500–1000 repetitions), it is thought that it allows for preparatory mechanisms, hierarchical control, and sequence chunking to be studied more clearly than in the SRT task (Rhodes et al., 2004). With this amount of training, there is time for multistep movements to concatenate together to form chunked movement patterns (Verwey, 1996), reflected by decreased RTs and also decreased IRIs between movements that are part of a chunk and longer RTs between chunks. This can also co-occur with S-R learning (Verwey & Wright, 2014; Verwey & Abrahamse, 2012). Chunking from the DSP task, as in the M × N task, allows for participants to perform more efficiently (Abrahamse et al., 2013). The DSP task also suggests that simple, nonhierarchical S-R learning can occur alongside the learning of chunks. Still, concatenating action sequences into chunks yields faster performance. Interestingly, one study using the DSP task instructed participants to use predesignated, suboptimal chunks throughout the task (Popp, Yokoi, Gribble, & Diedrichsen, 2020). After performing the task, these participants understood that these chunk patterns were suboptimal and developed their own idiosyncratic chunk patterns to perform more effectively. In other words, participants were able to discover a more useful option through training. An understanding of this process from a neural perspective will be crucial for understanding the option discovery problem more generally.

Chunks as Options

Motor chunks have been observed across multiple experimental paradigms, and they allow for greater performance when complex action sequences are required (Verwey & Wright, 2014; Abrahamse et al., 2013; Verwey & Abrahamse, 2012; Botvinick et al., 2009; Rhodes et al., 2004; Barto & Mahadevan, 2003; Sakai et al., 2003; Sutton, 1999). In the M × N task, it was shown that using chunks allows for quicker learning and more efficient performance of a new sequence. In the DSP task, chunks served a similar role by allowing quicker performance. The faster performance suggests that the component actions have been grouped together as a unit, similar to an option in HRL. In summary, the following parallels can be drawn between options and chunks. First, chunks and options are both treated as single action units (Sakai et al., 2003; Sutton et al., 1999), and therefore, they reduce the computational complexity of the agent's behavior (Ramkumar et al., 2016; Sutton et al., 1999). Second, this reduction in complexity makes it easier to learn. Hence, from behavioral evidence discussed in this review, we see that chunks and options serve highly analogous roles for agents navigating complex state spaces. Furthermore, when looking at the neural underpinnings of chunking behavior, additional parallels can be seen between chunks and a computational understanding of HRL.

Neural Systems Underlying Sequence Learning and Execution

The neural basis of sequence learning is an active area of research (Janacsek et al., 2020; Verwey, Jouen, Dominey, & Ventre-Dominey, 2019; Desrochers et al., 2010, 2015, 2016; Wymbs, Bassett, Mucha, Porter, & Grafton, 2012; Tremblay et al., 2009, 2010; Bo, Langan, & Seidler, 2008; Levesque et al., 2007; Rhodes et al., 2004). Here, we review previous work in selected human, nonhuman primate, and rodent studies from the perspective of HRL. Furthermore, we frame the discussion of neural results with a recently proposed distinction between dorsal and ventral frontostriatal circuitry.

In general, motor sequence execution and learning are thought to depend on frontal cortico-striato-pallido-thalamo-cortical circuits (Alexander, DeLong, & Strick, 1986). This hypothesis, which was an early neural theory of hierarchical control, developed from proposals that the BG, which receives much of its input from the frontal cortex, is important for the automated execution of learned motor plans (Marsden, 1982). Recent work has focused more on dynamics in specific parts of this circuitry, but the ways in which neural responses in these areas coordinate to give rise to observable behavior are not yet clear (Desrochers et al., 2016). HRL provides a useful framework for synthesizing results from behavior experiments that may help answer some of the open questions. Existing work provides broad insight into the neural substrates involved with hierarchical control (Badre & Nee, 2018; Rasmussen, Voelker, & Eliasmith, 2017; Balleine et al., 2015), where cortico-BG-thalamocortical loops have been suggested to underlie RL as well as HRL processes (Rasmussen et al., 2017; Samejima & Doya, 2007; Haruno & Kawato, 2006). Rasmussen et al. used an HRL model to predict that, within cortico-BG-thalamocortical loops, the ventral striatum (VS) represents previous state–action information and the dorsal striatum (DS) represents current state–action information (Rasmussen et al., 2017). Others have described a functional dissociation between medial and lateral regions of cortical-BG circuits (Balleine et al., 2015). Specifically, the medial loop integrates information from the dorsolateral PFC (dlPFC) and pre-SMA, which project to the caudate (dorsomedial striatum [DMS]) to mediate hierarchical action selection, and the lateral loop, which involves the SMA and putamen (dorsolateral striatum), implements execution of action chunks. Both of these loops feed back to the cortex through the BG and thalamus.

We build on this previous research by using a recent anatomical, computational, and behavioral framework (Averbeck & Murray, 2020) to explain how individual regions within these cortico-BG-thalamo-cortico loops may underlie hierarchical behavior. In this framework, neural systems are organized in two networks, a dorsal network and a ventral network. The dorsal network represents information necessary to compute actions that achieve goals, and the ventral network represents state value learning and goal identification. This proposal overlaps with the previous proposals. However, our hypothesis suggests a minimal role for the ventral circuit in action specification. In many circumstances, however, it may be difficult to distinguish between goals and actions if they have not been specifically dissociated. We also describe behavioral and neural findings relating to human motor sequence learning tasks, which allows for the controlled study of hierarchical learning and performance.

If chunks are comparable to HRL options, then the HRL framework provides predictions on the effect of lesion and inactivations of the brain regions introduced above as well as predictions on what may be represented by neural activity. Recall that an option has three components: an option-specific policy and sets of initiation and termination conditions. Hence, we predict that each of these components should be reflected in the brain. Given the involvement of the dlPFC in hierarchical control, we may predict that this region will be crucial for tracking an option-specific policy and a flat policy, by both planning for upcoming movements and representing the sequential structure of the task at hand. As an RPE updates flat policy and state values, we should also expect to find evidence of an RPE that would update option-specific policies.

In addition, a lesion or inactivation of the rodent DMS has been shown to lessen goal-directed behavior, and the same manipulation of the dorsolateral striatum reduces the selection of abstract chunked actions in rats (Balleine et al., 2015). Because the dlPFC is also involved with goal-directed behavior, we expect a similar result as with DMS manipulation. Signaling the initiation and termination of an option will need to occur as well. Given literature on the role of the DS in habit formation (Graybiel, 2008; Yin, Knowlton, & Balleine, 2004), we may expect this signal to occur in the DS in the form of task bracketing, a well-known finding occurring during sequential tasks (Martiros et al., 2018; Desrochers et al., 2015; Graybiel & Grafton, 2015; Smith & Graybiel, 2013; Jin & Costa, 2010).

Dorsal vs. Ventral Circuitry: Hierarchical Action Selection vs. Value Learning?

Recent work has proposed that goal-directed behavior is in large part orchestrated by two separate circuits: the dorsal and ventral systems (Averbeck & Murray, 2020). The dorsal system is a neural circuit composed of multiple cortical and subcortical nodes (Figure 3). Within cortex, the dorsal system is composed of dlPFC and Area 7a within the inferior parietal cortex. Both areas project to the DS. The DS then projects to the globus pallidus (GP) internal segment, which then projects to the lateral portion of the medial dorsal thalamus. The medial dorsal thalamus completes the circuit by projecting back to the dlPFC. The ventral system largely mirrors the projections from the dorsal system in ventral regions and, in the context of RL, is largely responsible for value representation, whereas the dorsal system is mainly responsible for action selection (Murray, Wise, & Graham, 2017; Andersen, Snyder, Bradley, & Xing, 1997; Bindra, 1978).

Although there are interactions between the dorsal and ventral systems at the cortical level, connections within respective systems are stronger than those between systems (Barbas & Pandya, 1989). The dorsal system is responsible for computing actions given a certain environment. The ventral system identifies relevant state information, such as objects and their values. Hence, most learning about how to satisfy one's needs takes place in the ventral circuitry—namely, learning the values of states and how to satisfy needs within them. To learn optimal behavior and how to execute it, these two systems must interact—that is, an agent's goal must be paired with how to achieve it. Some work has been done to better understand this connection (Tang, Bartolo, & Averbeck, 2021; Rasmussen et al., 2017).

There is reason to think that learning how to hierarchically orchestrate motor behavior takes place in the dorsal circuitry. The dlPFC is thought to be the top of the frontal hierarchy controlling behavior (Badre & Nee, 2018). In addition, cortical areas, such as the dlPFC and the pre-SMA, which is densely connected to dorsal regions (Guenther, Tourville, & Bohland, 2015; Luppino, Matelli, Camarda, & Rizzolatti, 1993), have been suggested to be involved with sequential planning, execution, and control (Badre & Nee, 2018; Balleine et al., 2015; Nachev, Kennard, & Husain, 2008; Averbeck, Chafee, Crowe, & Georgopoulos, 2002). The dlPFC projects to the DS, which is thought to have a role in orchestrating abstract behavior to achieve goals (Averbeck & Murray, 2020). Further supporting the connection between dlPFC and DS, neural activity was found to represent selected action stronger and earlier in the dlPFC than in the DS (Seo, Lee, & Averbeck, 2012). The role of the DS in chunking behavior has also been well established (Jin, Tecuapetla, & Costa, 2014; Jin & Costa, 2010; Graybiel, 2008), which is necessary for the dorsal stream to orchestrate abstract actions. This is further supported by the dorsal system's inputs from parietal cortical areas, leaving it well poised to represent action metrics such as object number, distance, and action duration (Averbeck & Murray, 2020; Andersen et al., 1997).

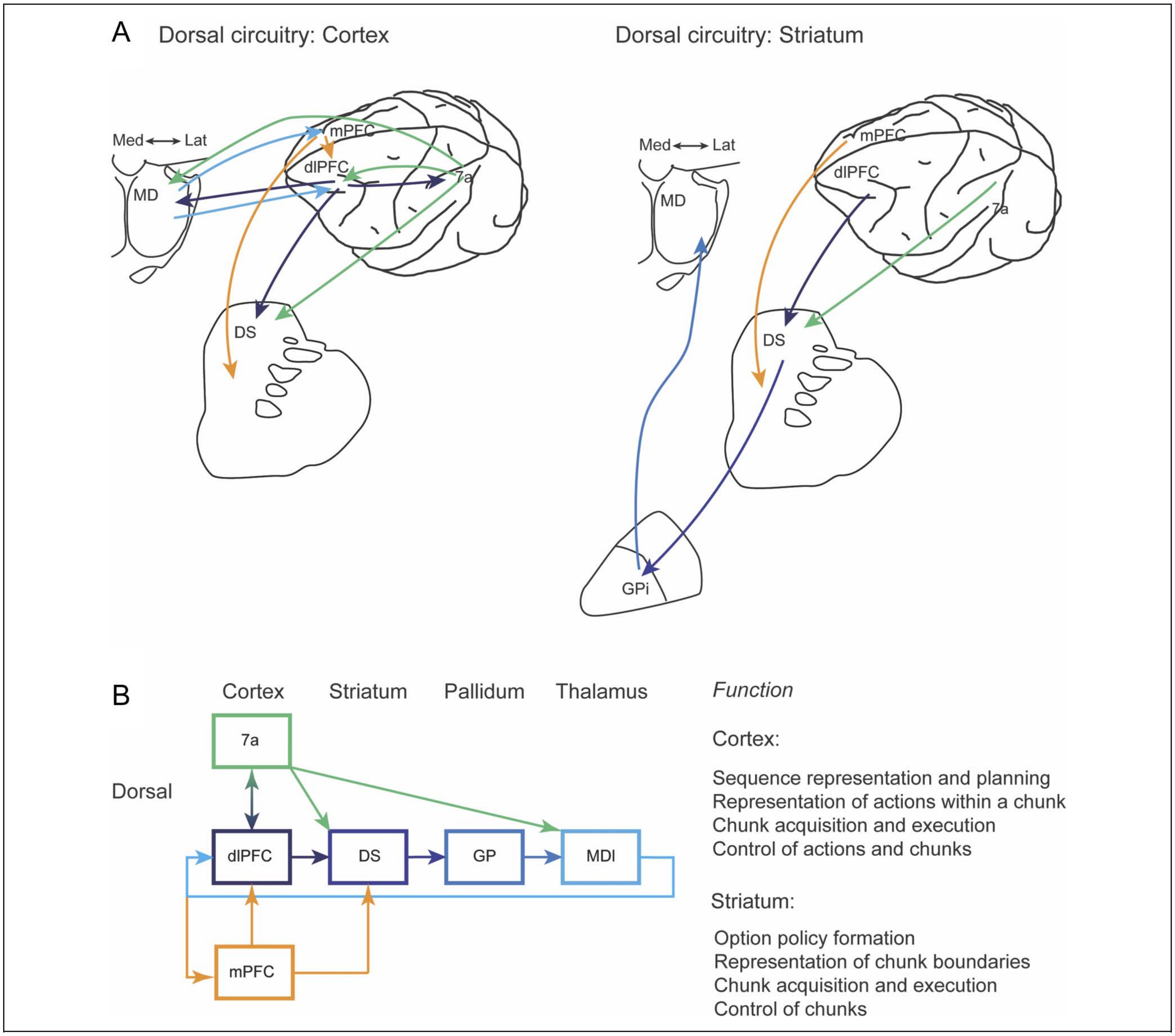
A
Dorsal circuitry: Cortex
Dorsal circuitry: Striatum
Med ⟷ Lat
MD
mPFC
dlPFC
7a
DS
Med ⟷ Lat
MD
mPFC
dlPFC
7a
DS
GPi
B
Cortex
Striatum
Pallidum
Thalamus
Function
Dorsal
7a
dlPFC
DS
GP
MDl
mPFC
Cortex:
Sequence representation and planning
Representation of actions within a chunk
Chunk acquisition and execution
Control of actions and chunks
Striatum:
Option policy formation
Representation of chunk boundaries
Chunk acquisition and execution
Control of chunks

Figure 3. Dorsal circuitry and functions for hierarchical learning and behavior. (A) Dorsal system consists of the dlPFC that projects to Area 7a within the inferior parietal cortex, both of which project to the DS. The DS then projects to the GPi, which then projects to the lateral portion of the MD thalamus. The MD thalamus completes the circuit by projecting back to the dlPFC. Additionally, the mPFC projects to the dlPFC and striatum and receives projections from the MD thalamus. Figure adapted from Averbeck and Murray (2020). (B) A simplified schematic of dorsal circuitry with corresponding functions of cortical and striatal areas is shown. Arrows denote directional projections between areas. GPe = GP external segment; GPi = GP internal segment; MDl = medial dorsal thalamus; mPFC = medial PFC including the pre-SMA and SMA; 7a = inferior parietal cortex or Area 7a.

The ventral striatal circuitry, which learns value-based information, a critical feature of hierarchical learning, and other networks must interact at a system level for complex, adaptable behavior (Badre & Nee, 2018). As mentioned earlier, studies using graph structures have furthered our understanding of how hierarchical structure can be learned (Tomov et al., 2020). Ventral circuitry, given its involvement in state learning (Averbeck & Murray, 2020), is likely to be involved in this process. Studies in primates, however, have not found substantial action representations in the ventral circuitry (Tang, Costa, Bartolo, & Averbeck, 2022; Costa, Mitz, & Averbeck, 2019). Hence, the ventral circuitry may underlie aspects of hierarchical learning and behavior, and perhaps the organization of subgoals, without specifying the action sequences necessary to achieve those goals (Averbeck & Murray, 2020; Murray & Rudebeck, 2018). Given the extensive work suggesting the dorsal circuitry's role in hierarchical action, we will focus on dorsal circuitry and how it contributes to hierarchical action control. In what follows, we review the available evidence from the sequence learning literature previously discussed as well as other relevant findings that may support the dorsal system's role in hierarchical learning. To do so, we go through each node of the dorsal circuit and discuss evidence for its involvement.

The dlPFC Contains Sequence Representations at Multiple Levels of Abstraction

The dlPFC is a critical node in the dorsal system with various connections to dorsal regions across cortical and subcortical areas. Recently, the entire PFC has been proposed to consist of an action hierarchy to orchestrate actions at various levels of abstraction (Fine & Hayden, 2022), and the rostrolateral PFC has been known to support sequential action (Desrochers, Collins, & Badre, 2019). Findings from the dlPFC specifically provide further evidence of this view. The dlPFC represents movement sequences at multiple levels of abstraction (Xie et al., 2022). Neurophysiology experiments have shown that chunked sequences may be planned before execution and represented in the dlPFC (Averbeck et al., 2002). In these experiments, monkeys were trained to draw shapes, which were composed of sequences of movements. Neural activity represented all movements in the sequence before the sequence began. Furthermore, the serial order of the elements of the sequence were represented by the strength of the representation in dlPFC. Examination of coding in single cells in this experiment showed that neural activity simultaneously represented information about the shape being drawn, the movement currently being executed, and metrics of the movement: all relevant to executing an option-specific policy (Averbeck, Chafee, Crowe, & Georgopoulos, 2003). This result was consistent with previous modeling and behavioral work (Rhodes et al., 2004; Grossberg, 1978, 1982) and has also been replicated in human participants (Kornysheva et al., 2019). An additional study found a similar result in the lateral PFC of nonhuman primates, and it was noted that neurons dorsal to the principal sulcus were most likely to participate in planning of abstract sequences (Shima, Isoda, Mushiake, & Tanji, 2007). Additional work using an eye-movement variant of the sequence learning task found that when neural activity in dlPFC represented the wrong sequence, before it was executed, errors in the sequence could be predicted up to three movements into the future (Averbeck & Lee, 2007). Thus, the dlPFC appears to play a key role in the neural representation and planning of motor sequences, which may suggest an involvement of the dlPFC in tracking metrics relevant to option-specific policies and planning for their execution.

The dlPFC is strongly connected with the inferior parietal cortex (Selemon & Goldman-Rakic, 1988), which is another region within the dorsal system shown to be important for sequence learning and chunking. Representation of both sequences and motor chunks has been found in parietal areas during a finger press task (Yokoi & Diedrichsen, 2019; Wiestler & Diedrichsen, 2013). In addition, sequence segmentation, a necessary step in chunk formation within long sequences, was found to correlate with activation in dorsal–frontoparietal areas during a variant of the DSP task (Wymbs et al., 2012). Activity in the posterior parietal cortex has also been seen using the standard DSP task (Verwey et al., 2019), although, in a variant of a motor sequence learning task, activity in parietal areas has been shown to decrease across extended periods of learning (Berlot, Popp, & Diedrichsen, 2020). Furthermore, the parietal reach region has been shown to encode future sequence movements before initiation of a reach sequence (Baldauf, Cui, & Andersen, 2008). As these parietal cortical areas provide inputs to the dlPFC, this finding further suggests that the organization of hierarchical action depends on this region as well as its downstream regions throughout the dorsal system (Averbeck & Murray, 2020).

The results of these studies and others suggest that dlPFC has multiple functions with respect to sequence learning and hierarchical planning (Lee & Seo, 2007). Activity in this region seems to represent chunks at multiple levels of abstraction: both at the level of behavioral metrics and at the sequence level. Furthermore, regions with a strong connection to the dlPFC support the chunking process, with activity in these areas reflecting information relevant to sequence segmentation. With respect to HRL, the dlPFC may be a critical node potentially representing option preparation and execution, which communicates motor information throughout the dorsal stream. Furthermore, activity in the dlPFC and the associated frontoparietal areas may be crucial for chunk formation, which suggests that study of these regions may lend future insight into option discovery.

The DS Is Involved in Option Formation, Execution, and Update Signals

Studies have also found that the DS is involved in sequence learning and chunking processes. Recall that options are defined by three components: an option-specific policy that maps actions to states, initiation conditions, and termination conditions. There is also an option-specific reward function that attaches rewards to option completion, serving as an abstract form of the well-studied standard RPE (Schultz, Dayan, & Montague, 1997; Montague, Dayan, & Sejnowski, 1996). DS activity and dopamine signaling throughout sequence learning may show a role for this dorsal node in these HRL processes.

Functional imaging studies during the SRT and DSP tasks have shown activation in the DS during sequence learning (Wymbs et al., 2012; Poldrack et al., 2005; Grafton, Hazeltine, & Ivry, 1995). The DS is composed of the caudate and putamen (Averbeck & Murray, 2020). Studies have suggested that the caudate is important for learning S-R associations (Poldrack et al., 2001) and working memory processes (Owen et al., 1998) in sequential tasks. More recently, a functional neuroanatomical meta-analysis of SRT studies was conducted to investigate which brain regions have been consistently activated across experiments. This study found significant convergence in the caudate, GP, anterior putamen, and VS (Janacsek et al., 2020). Activity in the sensorimotor putamen has also been associated with the sequential binding of chunks, which were again shown through decreased variability in IRIs between movement patterns (Wymbs et al., 2012). Furthermore, distinct activity representing trained sequences has been seen in the DS during extensive training of a varied motor learning task (Berlot et al., 2020). These data support the idea that the DS may play a key role in mapping hierarchically organized actions together, a crucial aspect of policy formation.

Learning to select rewarding actions in the DS may depend on dopamine inputs (Kwak, Bohnen, Müller, Dayalu, & Seidler, 2013; O'Doherty, 2004). Studies using variants of the DSP task have evaluated the effects of disrupting dopaminergic signaling on chunking (Levesque et al., 2007; Matsumoto, Hanakawa, Maki, Graybiel, & Kimura, 1999). In these experiments, IRIs were slower and more variable for previously learned (and chunked) sequences when dopaminergic signaling was disrupted. This suggests that blocking dopaminergic signaling, specifically in the striatum, can lead to deficits in sequence chunking. Although some have asserted that disrupting dopamine affects movement kinematics and not the hierarchical representation of the sequence (Desmurget & Turner, 2010), human studies in Parkinson's disease have shown consistent support for the role of dopamine in chunking behavior. Patients off levodopa medication had slower and more variable RTs for sequential actions and did not chunk with training (Tremblay et al., 2010). This suggests that parsing of sequences into chunks, or the formation of options, likely depends on striatal dopamine signaling.

Other studies have shown that the DS is involved in task bracketing. Task bracketing is neural activity that brackets execution of sequences. Therefore, task bracketing signals the initiation and termination of a sequence. For example, it has been shown in rats and monkeys that, after extensive training, populations of striatal neurons signal the "beginning" and "end" of action sequences (Martiros et al., 2018; Desrochers et al., 2015; Graybiel & Grafton, 2015; Smith & Graybiel, 2013; Jin & Costa, 2010). These task-bracketing signals may serve as representations of initiation and termination conditions for HRL options, which notify an agent when to begin and end a selected option. This signaling appears to emerge with training (Smith & Graybiel, 2013). A similar task-bracketing activity has been observed in the infralimbic cortex, macaque PFC (Fujii & Graybiel, 2003), and caudate (Desrochers et al., 2015), suggesting that the representation of chunked sequences of actions is, unsurprisingly, distributed across a network of dorsal areas. Notably, caudate neurons with outcome and cost representations exhibited responses that were more tightly linked to the end of the sequence when sequences were overlearned. This may provide the update signal necessary for learning habit-like stereotyped sequences (Desrochers et al., 2015). Even further, this outcome and cost representation could provide the update signals necessary for option formation. Pseudo-RPEs are generated once the agent reaches the state at which the option terminates and a reward is received. Therefore, these representations could be analagous to HRL pseudo-RPEs.

Furthermore, recent work has shown that the DS may serve a critical role in control over chunked action sequences, which is particularly important when multiple options are simultaneously available (Geddes et al., 2018). Here, recordings in the rodent DS targeted distinct populations of spiny projection neurons (SPNs): those signaling through the direct pathway (dSPNs) or the indirect pathway (iSPNs). The rodents were tasked with pressing two levers in the proper sequence—two presses of the left lever (left subsequence) followed by two presses of the right lever (right subsequence)—and then moving to a magazine for reward. After training, the rodents formed a chunk for each subsequence, verified by reduced variability in response times between chunks and an overall increase in speed (Geddes et al., 2018). Single-cell recordings showed that dSPNs in the striatum were preferentially active for sequence-level start and stop-related activity, whereas iSPNs were preferentially active during the transitions between subsequences. In other words, dSPNs were primarily active as the rodent initiated or ended the full movement sequence, whereas iSPNs were primarily active while the rodent transitioned between subsequences, possibly options.

Optogenetic stimulation of either dSPNs or iSPNs was then performed as the rodents performed each element of the sequence. Stimulation of iSPNs during the very first lever press eliminated the following action in the sequence and evoked a behavioral transition to the next subsequence. In contrast, optogenetic stimulation of dSPNs caused the animals to repeat their previous lever press and omit the subsequent lever press of the next subsequence. This suggests that the mice were maintaining a numerical structure of the sequences as even when they repeated a lever press within a subsequence, they maintained the same number of lever presses within the entire sequence. Furthermore, this suggests that an option-specific policy may have been employed by the rodents. The policy under consideration contained two options, one for each subsequence. dSPNs and iSPNs jointly supported navigating through the option, and upon stimulation, the policy was disrupted. Taken together, these results show dSPNs and iSPNs in the DS are involved with control of actions and possibly option-specific policies.

As has been shown, the DS is involved in all aspects of an HRL option including formation of chunks and mapping them together (option-specific policy) as well as signaling the beginning and end of a given chunk (initiation and termination conditions). These findings further support the claim that the dorsal system is heavily involved in hierarchical sequence learning as well as the utility of considering chunks as HRL options.

Medial Frontal Cortex: Sequential Control

An additional region in the dorsal system is the dorsomedial frontal cortex, which projects to the dlPFC, DS, and GP (Figure 3). There is evidence in support of its role in acquiring new sequences and control of familiar sequences. This function is critical within HRL, particularly with respect to discovering useful options and implementing option-specific policies in behavior. Furthermore, given the aforementioned role of the dlPFC and DS in hierarchical behavior and the anatomical connectivity between these structures and the medial frontal cortex, the medial frontal cortex is well positioned to further support such behavior such as chunking and sequential control.

The medial frontal cortex, which includes the SMA, pre-SMA, ACC, and medial frontal pole, is known to be involved in sequential behavior (Amodio & Frith, 2006). Both the SMA and pre-SMA participate in a cortical–subcortical loop (Akkal, Dum, & Strick, 2007; Inase, Tokuno, Nambu, Akazawa, & Takada, 1999; Parthasarathy, Schall, & Graybiel, 1992). The pre-SMA and SMA regions are distinguished, however, both anatomically and functionally. The SMA, with strong connections to motor cortex and the putamen, is likely to be more involved in motor execution, whereas the pre-SMA with strong connections to regions of PFC and the caudate may be more involved in abstract functions (Guenther et al., 2015). The pre-SMA projects to the dlPFC (Luppino et al., 1993). Processing between the pre-SMA and dlPFC may send sequence-level information to downstream subcortical areas, such as the DS, where the proper action routine can be implemented.

Neural activity in the SMA and pre-SMA has been seen in the DSP and M × N tasks (Verwey et al., 2019; Nakamura, Sakai, & Hikosaka, 1998), suggesting that these regions may play a role in chunking of well-learned sequences. In an fMRI study using the DSP task, researchers found that the pre-SMA was activated for familiar and unfamiliar sequences, but SMA was specifically activated for familiar sequences that had been chunked by participants (Verwey et al., 2019). Similarly, SMA neurons in monkeys have been shown to be preferentially active before certain movement sequences (Tanji & Shima, 1994). Single neuron activity in SMA and pre-SMA has also been studied using the M × N paradigm in monkeys (Nakamura et al., 1998). Nakamura et al. found that, as monkeys learned a sequence of actions, some pre-SMA neurons were active throughout the sequence but decreased their activity as the sequence became well learned. This suggests that the pre-SMA was more important for acquiring new sequences than executing previously learned sequences, which is noteworthy for understanding option discovery. Other neurons in the pre-SMA respond to the rank order of actions within a sequence (Shima & Tanji, 2000; Clower & Alexander, 1998), and inactivation of pre-SMA or SMA during performance of a learned motor sequence results in errors, suggesting that these areas also support the tracking of sequential steps (Shima & Tanji, 1998).

These results suggest a critical role for both the SMA and pre-SMA in the control of chunked sequences. Although both regions seem to be important, there may be a slightly greater role for the pre-SMA than the SMA (Nachev et al., 2008). Regardless, these regions participate in the dorsal circuitry through their connections with the dlPFC, DS, and downstream projections to the GP and thalamus (Haber, 2016; Figure 3). Furthermore, the medial frontal cortex's involvement with dorsal signaling and sequence learning suggests its role in hierarchical learning and behavior. As mentioned, the pre-SMA may be critical for acquiring new options, whereas the SMA may be more relevant to option control.

There are also fMRI and EEG studies that provide evidence that ACC may have a role in encoding pseudo-RPEs (Ribas-Fernandes et al., 2011); however, in later studies, there was insignificant activation (Chiang & Wallis, 2018). The study from Ribas-Fernandes et al. was unsuccessfully replicated by their own further work (Ribas-Fernandes, Shahnazian, Holroyd, & Botvinick, 2019). ACC is also involved with error processing (Seidler, Kwak, Fling, & Bernard, 2013), but further work is required to understand if this region is involved with HRL processes, such as pseudo-RPEs.

In summary, research on the neural systems underlying sequence learning have extensively implicated frontal cortical-BG circuitry. First, the dlPFC seems to have a role in representing sequences at multiple levels of abstraction, sequence learning, and action planning. The medial PFC, which has connections to the dorsal circuitry, may have a role in acquiring new options and tracking and controlling sequences. Finally, there is evidence of the DS's role in option or chunk formation and execution through an update signal. A common thread throughout the findings reviewed here is that different regions within or with connections to the dorsal system may support similar functions with respect to sequence learning and hierarchical action execution. Further work will be needed to clarify the contributions of each region. In the remainder of this review, we will outline a few suggestions for future research that may allow for such clarification.

Future Directions

The work reviewed here continues to develop a framework for integrating RL and motor sequence learning from the perspective of HRL (Balleine & Dezfouli, 2019; Geddes et al., 2018). Future experimental work would benefit from increasingly sophisticated sequence tasks. In the standard sequence learning experiments (e.g., SRT, DSP, and M × N), participants execute simple sequences of actions such as key presses or taps. Some studies have used tasks that require multiple goal-driven steps (Ribas-Fernandes et al., 2011, 2019; Diuk, Tsai, Wallis, Botvinick, & Niv, 2013). In these tasks, a subgoal must be achieved, such as picking up a virtual package with a controller-guided truck, before completion of the trial, in which the package is delivered. This kind of design allows for a study of how different goals may be hierarchically arranged and how this arrangement influences behavior.

Furthermore, future tasks should explore the flexible achievement of goals. Studying how a biological agent devises a complex series of actions to achieve reward will be critical in understanding goal-directed behavior and, more specifically, the option discovery problem. Most behavioral tasks remove the flexibility required in real-world environments. Designing experiments that allow for flexible completion of goals and constant learning that are simultaneously controlled well enough to interpret is challenging. However, incremental steps toward naturalistic tasks are achievable. For example, experiments might allow for multiple, quantifiable strategies and look at neural correlates for the different strategies taken by participants, rather than constrain the task to one allowable strategy (Balleine & Dezfouli, 2019; Garr, 2019; Desrochers et al., 2010, 2015; Dezfouli & Balleine, 2012). Designing an experiment to encourage behavioral variance in an interpretable way is currently contrarian to typical experimental approaches, but its value in understanding complex behavior cannot be understated.

Conclusion

Understanding motor sequence learning through the lens of HRL is a promising route to gaining a better understanding of complex behavior. The sequence literature has provided a basic understanding of how simple chunked action routines are acquired and executed as well as what conditions are ideal for producing them. By bringing in the HRL framework, we can provide context for understanding chunks as options executed by an agent learning in a hierarchical manner. Although computational methods are continually advancing on the option discovery problem, an understanding of how this is accomplished by the brain is still lacking. Nevertheless, a dorsal cortical–subcortical network of brain areas is largely supporting both hierarchical behavior and learning, and a deeper understanding of this circuitry is likely to bring an understanding of how complex behavior is orchestrated in real-world environments.

Acknowledgments

This work was supported by the intramural research program of National Institute of Mental Health (ZIA MH002928).

Reprint requests should be sent to Bruno B. Averbeck, Laboratory of Neuropsychology, National Institute of Mental Health/National Institutes of Health, Building 49 Room 1B80, 49 Convent Drive MSC 4415, Bethesda, MD 20892-4415, or via e-mail: bruno.averbeck@nih.gov.

Author Contributions

Miriam Janssen: Conceptualization; Investigation; Writing—Original draft; Writing—Review & editing. Christopher LeWarne: Conceptualization; Investigation; Writing—Original draft; Writing—Review & editing. Diana Burk: Resources; Supervision; Visualization; Writing—Review & editing. Bruno B. Averbeck: Conceptualization; Resources; Supervision; Writing—Original draft; Writing—Review & editing.

Funding Information

Bruno B. Averbeck, National Institute of Mental Health (https://dx.doi.org/10.13039/100000025), grant number: ZIA MH002928.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this article report its proportions of citations by gender category to be as follows: M/M = .716, W/M = .116, M/W = .074, and W/W = .095.

REFERENCES

Abrahamse, E. L., Ruitenberg, M. F., de Kleine, E., & Verwey, W. B. (2013). Control of automated behavior: Insights from the discrete sequence production task. *Frontiers in Human Neuroscience*, *7*, 82. https://doi.org/10.3389/fnhum.2013.00082, PubMed: 23515430

Akkal, D., Dum, R. P., & Strick, P. L. (2007). Supplementary motor area and presupplementary motor area: Targets of basal ganglia and cerebellar output. *Journal of Neuroscience*, *27*, 10659–10673. https://doi.org/10.1523/jneurosci.3134-07.2007, PubMed: 17913900

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. https://doi.org/10.1146/annurev.ne.09.030186.002041, PubMed: 3085570

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*, 268–277. https://doi.org/10.1038/nrn1884, PubMed: 16552413

Andersen, R. A., Snyder, L. H., Bradley, D. C., & Xing, J. (1997). Multimodal representation of space in the posterior parietal cortex and its use in planning movements. *Annual Review of Neuroscience*, *20*, 303–330. https://doi.org/10.1146/annurev.neuro.20.1.303, PubMed: 9056716

Averbeck, B. B., Chafee, M. V., Crowe, D. A., & Georgopoulos, A. P. (2002). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *99*, 13172–13177. https://doi.org/10.1073/pnas.162485599, PubMed: 12242330

Averbeck, B. B., Chafee, M. V., Crowe, D. A., & Georgopoulos, A. P. (2003). Neural activity in prefrontal cortex during copying geometrical shapes: I. Single cells encode shape, sequence, and metric parameters. *Experimental Brain Research*, *150*, 127–141. https://doi.org/10.1007/s00221-003-1416-6, PubMed: 12669170

Averbeck, B. B., & Costa, V. D. (2017). Motivational neural circuits underlying reinforcement learning. *Nature Neuroscience*, *20*, 505–512. https://doi.org/10.1038/nn.4506, PubMed: 28352111

Averbeck, B. B., & Lee, D. (2007). Prefrontal neural correlates of memory for sequences. *Journal of Neuroscience*, *27*, 2204–2211. https://doi.org/10.1523/jneurosci.4483-06.2007, PubMed: 17329417

Averbeck, B. B., & Murray, E. A. (2020). Hypothalamic interactions with large-scale neural circuits underlying reinforcement learning and motivated behavior. *Trends in Neurosciences*, *43*, 681–694. https://doi.org/10.1016/j.tins.2020.06.006, PubMed: 32762959

Averbeck, B. B., & O'Doherty, J. P. (2021). Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology*, *47*, 147–162. https://doi.org/10.1038/s41386-021-01108-0, PubMed: 34354249

Badre, D., & Nee, D. E. (2018). Frontal cortex and the hierarchical control of behavior. *Trends in Cognitive Sciences*, *22*, 170–188. https://doi.org/10.1016/j.tics.2017.11.005, PubMed: 29229206

Baldauf, D., Cui, H., & Andersen, R. A. (2008). The posterior parietal cortex encodes in parallel both goals for double-reach sequences. *Journal of Neuroscience*, *28*, 10081–10089. https://doi.org/10.1523/JNEUROSCI.3423-08.2008, PubMed: 18829966

Balleine, B. W., & Dezfouli, A. (2019). Hierarchical action control: Adaptive collaboration between actions and habits. *Frontiers in Psychology*, *10*, 2735. https://doi.org/10.3389/fpsyg.2019.02735, PubMed: 31920796

Balleine, B. W., Dezfouli, A., Ito, M., & Doya, K. (2015). Hierarchical control of goal-directed action in the cortical–basal ganglia network. *Current Opinion in Behavioral Sciences*, *5*, 1–7. https://doi.org/10.1016/j.cobeha.2015.06.001

Balleine, B., & Dickinson, A. (1992). Signalling and incentive processes in instrumental reinforcer devaluation. *Quarterly Journal of Experimental Psychology, Series B, Comparative and Physiological Psychology*, *45*, 285–301. PubMed: 1475401

Balleine, B. W., Garner, C., Gonzalez, F., & Dickinson, A. (1995). Motivational control of heterogeneous instrumental chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*, 203–217. https://doi.org/10.1037/0097-7403.21.3.203

Barbas, H., & Pandya, D. N. (1989). Architecture and intrinsic connections of the prefrontal cortex in the rhesus monkey. *Journal of Comparative Neurology*, *286*, 353–375. https://doi.org/10.1002/cne.902860306, PubMed: 2768563

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, *13*, 341–379. https://doi.org/10.1023/A:1025696116075

Berlot, E., Popp, N. J., & Diedrichsen, J. (2020). A critical re-evaluation of fMRI signatures of motor sequence learning. *eLife*, *9*, e55241. https://doi.org/10.7554/eLife.55241, PubMed: 32401193

Bindra, D. (1978). How adaptive behavior is produced: A perceptual–motivational alternative to response reinforcements. *Behavioral and Brain Sciences*, *1*, 41–52. https://doi.org/10.1017/S0140525X00059380

Bo, J., Langan, J., & Seidler, R. D. (2008). Cognitive neuroscience of skill acquisition. In A. S. Benjamin, J. S. De Belle, B. Etnyre, & T. A. Polk (Eds.), *Advances in psychology* (Vol. 139, pp. 101–112). North Holland, The Netherlands: Elsevier. https://doi.org/10.1016/S0166-4115(08)10009-7

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current Opinion in Neurobiology*, *22*, 956–962. https://doi.org/10.1016/j.conb.2012.05.008, PubMed: 22695048

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*, 262–280. https://doi.org/10.1016/j.cognition.2008.08.011, PubMed: 18926527

Botvinick, M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *369*, 20130480. https://doi.org/10.1098/rstb.2013.0480, PubMed: 25267822

Chiang, F. K., & Wallis, J. D. (2018). Neuronal encoding in prefrontal cortex during hierarchical reinforcement learning. *Journal of Cognitive Neuroscience*, *30*, 1197–1208. https://doi.org/10.1162/jocn_a_01272, PubMed: 29694261

Clower, W. T., & Alexander, G. E. (1998). Movement sequence-related activity reflecting numerical order of components in supplementary and presupplementary motor areas. *Journal of Neurophysiology*, *80*, 1562–1566. https://doi.org/10.1152/jn.1998.80.3.1562, PubMed: 9744961

Collins, A. G. E., & Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition*, *152*, 160–169. https://doi.org/10.1016/j.cognition.2016.04.002, PubMed: 27082659

Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A., & Averbeck, B. B. (2016). Amygdala and ventral striatum make distinct contributions to reinforcement learning. *Neuron*, *92*, 505–517. https://doi.org/10.1016/j.neuron.2016.09.025, PubMed: 27720488

Costa, V. D., Mitz, A. R., & Averbeck, B. B. (2019). Subcortical substrates of explore–exploit decisions in primates. *Neuron*, *103*, 533–545. https://doi.org/10.1016/j.neuron.2019.05.017, PubMed: 31196672

Costa, V. D., Tran, V. L., Turchi, J., & Averbeck, B. B. (2015). Reversal learning and dopamine: A bayesian perspective. *Journal of Neuroscience*, *35*, 2407–2416. https://doi.org/10.1523/jneurosci.1989-14.2015, PubMed: 25673835

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027, PubMed: 21435563

Desmurget, M., & Turner, R. S. (2010). Motor sequences and the basal ganglia: Kinematics, not habits. *Journal of Neuroscience*, *30*, 7685–7690. https://doi.org/10.1523/jneurosci.0163-10.2010, PubMed: 20519543

Desrochers, T. M., Amemori, K., & Graybiel, A. M. (2015). Habit learning by naive macaques is marked by response sharpening of striatal neurons representing the cost and outcome of acquired action sequences. *Neuron*, *87*, 853–868. https://doi.org/10.1016/j.neuron.2015.07.019, PubMed: 26291166

Desrochers, T. M., Burk, D. C., Badre, D., & Sheinberg, D. L. (2016). The monitoring and control of task sequences in human and non-human primates. *Frontiers in Systems Neuroscience*, *9*, 185. https://doi.org/10.3389/fnsys.2015.00185, PubMed: 26834581

Desrochers, T. M., Collins, A. G. E., & Badre, D. (2019). Sequential control underlies robust ramping dynamics in the rostrolateral prefrontal cortex. *Journal of Neuroscience*, *39*, 1471–1483. https://doi.org/10.1523/jneurosci.1060-18.2018, PubMed: 30578340

Desrochers, T. M., Jin, D. Z., Goodman, N. D., & Graybiel, A. M. (2010). Optimal habits can develop spontaneously through sensitivity to local cost. *Proceedings of the National Academy of Sciences, U.S.A.*, *107*, 20512–20517. https://doi.org/10.1073/pnas.1013470107, PubMed: 20974967

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*, 1036–1051. https://doi.org/10.1111/j.1460-9568.2012.08050.x, PubMed: 22487034

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Computational Biology*, *9*, e1003364. https://doi.org/10.1371/journal.pcbi.1003364, PubMed: 24339762

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *Journal of Neuroscience*, *33*, 5797–5805. https://doi.org/10.1523/jneurosci.5445-12.2013, PubMed: 23536092

Eckstein, M. K., & Collins, A. G. E. (2020). Computational evidence for hierarchically structured reinforcement learning in humans. *Proceedings of the National Academy of Sciences, U.S.A.*, *117*, 29381–29389. https://doi.org/10.1073/pnas.1912330117, PubMed: 33229518

Eckstein, M. K., & Collins, A. G. E. (2021). How the mind creates structure: Hierarchical learning of action sequences. *Cognitive Science Society*, *43*, 618–624. PubMed: 34964045

Fine, J. M., & Hayden, B. Y. (2022). The whole prefrontal cortex is premotor cortex. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *377*, 20200524. https://doi.org/10.1098/rstb.2020.0524, PubMed: 34957853

Fujii, N., & Graybiel, A. M. (2003). Representation of action sequence boundaries by macaque prefrontal cortical neurons. *Science*, *301*, 1246–1249. https://doi.org/10.1126/science.1086872, PubMed: 12947203

Garr, E. (2019). Contributions of the basal ganglia to action sequence learning and performance. *Neuroscience and Biobehavioral Reviews*, *107*, 279–295. https://doi.org/10.1016/j.neubiorev.2019.09.017, PubMed: 31541637

Geddes, C. E., Li, H., & Jin, X. (2018). Optogenetic editing reveals the hierarchical organization of learned action sequences. *Cell*, *174*, 32–43. https://doi.org/10.1016/j.cell.2018.06.012, PubMed: 29958111

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128. https://doi.org/10.1146/annurev-psych-122414-033625, PubMed: 27618944

Grafton, S. T., Hazeltine, E., & Ivry, R. (1995). Functional mapping of sequence learning in normal humans. *Journal of Cognitive Neuroscience*, *7*, 497–510. https://doi.org/10.1162/jocn.1995.7.4.497, PubMed: 23961907

Graybiel, A. M. (2008). Habits, rituals, and the evaluative brain. *Annual Review of Neuroscience*, *31*, 359–387. https://doi.org/10.1146/annurev.neuro.29.051605.112851, PubMed: 18558860

Graybiel, A. M., & Grafton, S. T. (2015). The striatum: Where skills and habits meet. *Cold Spring Harbor Perspectives in Biology*, *7*, a021691. https://doi.org/10.1101/cshperspect.a021691, PubMed: 26238359

Grossberg, S. (1978). Behavioral contrast in short term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology*, *17*, 199–219. https://doi.org/10.1016/0022-2496(78)90016-0

Grossberg, S. (1982). A theory of human memory: Self-organization and performance of sensory–motor codes, maps, and plans. In S. Grossberg (Ed.), *Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control* (pp. 498–639). Dordrecht, The Netherlands: Springer Netherlands. https://doi.org/10.1007/978-94-009-7758-7_13

Guenther, F. H., Tourville, J. A., & Bohland, J. W. (2015). Speech production. In A. W. Toga (Ed.), *Brain mapping* (pp. 435–444). Waltham, MA: Academic Press. https://doi.org/10.1016/B978-0-12-397025-1.00265-7

Haber, S. N. (2016). Corticostriatal circuitry. *Dialogues in Clinical Neuroscience*, *18*, 7–21. https://doi.org/10.31887/DCNS.2016.18.1/shaber, PubMed: 27069376

Haruno, M., & Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus–action–reward association learning. *Journal of Neurophysiology*, *95*, 948–959. https://doi.org/10.1152/jn.00382.2005, PubMed: 16192338

Hikosaka, O., Rand, M. K., Miyachi, S., & Miyashita, K. (1995). Learning of sequential movements in the monkey: Process of learning and retention of memory. *Journal of Neurophysiology*, *74*, 1652–1661. https://doi.org/10.1152/jn.1995.74.4.1652, PubMed: 8989401

Inase, M., Tokuno, H., Nambu, A., Akazawa, T., & Takada, M. (1999). Corticostriatal and corticosubthalamic input zones from the presupplementary motor area in the macaque monkey: Comparison with the input zones from the supplementary motor area. *Brain Research*, *833*, 191–201. https://doi.org/10.1016/s0006-8993(99)01531-0, PubMed: 10375694

Janacsek, K., Shattuck, K. F., Tagarelli, K. M., Lum, J. A. G., Turkeltaub, P. E., & Ullman, M. T. (2020). Sequence learning in the human brain: A functional neuroanatomical meta-analysis of serial reaction time studies. *Neuroimage*, *207*, 116387. https://doi.org/10.1016/j.neuroimage.2019.116387, PubMed: 31765803

Jiménez, L., Méndez, A., Pasquali, A., Abrahamse, E., & Verwey, W. (2011). Chunking by colors: Assessing discrete learning in a continuous serial reaction-time task. *Acta Psychologica*, *137*, 318–329. https://doi.org/10.1016/j.actpsy.2011.03.013, PubMed: 21514547

Jin, X., & Costa, R. M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, *466*, 457–462. https://doi.org/10.1038/nature09263, PubMed: 20651684

Jin, X., Tecuapetla, F., & Costa, R. M. (2014). Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nature Neuroscience*, *17*, 423–430. https://doi.org/10.1038/nn.3632, PubMed: 24464039

Kawai, R., Markman, T., Poddar, R., Ko, R., Fantana, A. L., Dhawale, A. K., et al. (2015). Motor cortex is required for learning but not for executing a motor skill. *Neuron*, *86*, 800–812. https://doi.org/10.1016/j.neuron.2015.03.024, PubMed: 25892304

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, *13*, 400–408. https://doi.org/10.1093/cercor/13.4.400, PubMed: 12631569

Kim, J. Z., Soffer, J. M., Kahn, A. E., Vettel, J. M., Pasqualetti, F., & Bassett, D. S. (2018). Role of graph architecture in controlling dynamical networks with applications to neural systems. *Nature Physics*, *14*, 91–98. https://doi.org/10.1038/nphys4268, PubMed: 29422941

Kornysheva, K., Bush, D., Meyer, S. S., Sadnicka, A., Barnes, G., & Burgess, N. (2019). Neural competitive queuing of ordinal structure underlies skilled sequential action. *Neuron*, *101*, 1166–1180. https://doi.org/10.1016/j.neuron.2019.01.018, PubMed: 30744987

Kwak, Y., Bohnen, N. I., Müller, M. L. T. M., Dayalu, P., & Seidler, R. D. (2013). Striatal denervation pattern predicts levodopa effects on sequence learning in Parkinson's disease. *Journal of Motor Behavior*, *45*, 423–429. https://doi.org/10.1080/00222895.2013.817380, PubMed: 23971968

Lee, D., & Seo, H. (2007). Mechanisms of reinforcement learning and decision making in the primate dorsolateral prefrontal cortex. *Annals of the New York Academy of Sciences*, *1104*, 108–122. https://doi.org/10.1196/annals.1390.007, PubMed: 17347332

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*, 287–308. https://doi.org/10.1146/annurev-neuro-062111-150512, PubMed: 22462543

Levesque, M., Bedard, M. A., Courtemanche, R., Tremblay, P. L., Scherzer, P., & Blanchet, P. J. (2007). Raclopride-induced motor consolidation impairment in primates: Role of the dopamine type-2 receptor in movement chunking into integrated sequences. *Experimental Brain Research*, *182*, 499–508. https://doi.org/10.1007/s00221-007-1010-4, PubMed: 17653704

Luppino, G., Matelli, M., Camarda, R., & Rizzolatti, G. (1993). Corticocortical connections of area F3 (SMA-proper) and area F6 (pre-SMA) in the macaque monkey. *Journal of Comparative Neurology*, *338*, 114–140. https://doi.org/10.1002/cne.903380109, PubMed: 7507940

Mannor, S., Menache, I., Hoze, A., & Klein, U. (2004). Dynamic abstraction in reinforcement learning via clustering. *Paper presented at the Proceedings of the Twenty-First International Conference on Machine Learning*, Banff, Alberta, Canada. https://doi.org/10.1145/1015330.1015355

Marsden, C. D. (1982). The mysterious motor function of the basal ganglia: The Robert Wartenberg lecture. *Neurology*, *32*, 514–539. https://doi.org/10.1212/wnl.32.5.514, PubMed: 7200209

Martiros, N., Burgess, A. A., & Graybiel, A. M. (2018). Inversely active striatal projection neurons and interneurons selectively delimit useful behavioral sequences. *Current Biology*, *28*, 560–573. https://doi.org/10.1016/j.cub.2018.01.031, PubMed: 29429614

Matsumoto, N., Hanakawa, T., Maki, S., Graybiel, A. M., & Kimura, M. (1999). Nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner. *Journal of Neurophysiology*, *82*, 978–998. https://doi.org/10.1152/jn.1999.82.2.978, PubMed: 10444692

Menache, I., Mannor, S., & Shimkin, N. (2002). Dynamic discovery of sub-goals in reinforcement learning. In *Proceedings of the 13th European conference on machine learning* (pp. 295–306).

Miller, K. J., Ludvig, E. A., Pezzulo, G., & Shenhav, A. (2018). Realigning models of habitual and goal-directed decision-making. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-directed decision making: Computations and neural circuits* (pp. 407–428). Philadelphia: Elsevier Academic Press. https://doi.org/10.1016/B978-0-12-812098-9.00018-8

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947. https://doi.org/10.1523/jneurosci.16-05-01936.1996, PubMed: 8774460

Murray, E. A., & Rudebeck, P. H. (2018). Specializations for reward-guided decision-making in the primate ventral prefrontal cortex. *Nature Reviews Neuroscience*, *19*, 404–417. https://doi.org/10.1038/s41583-018-0013-4, PubMed: 29795133

Murray, E. A., Wise, S. P., & Graham, K. S. (2017). *The evolution of memory systems: Ancestors, anatomy, and adaptations*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199686438.001.0001

Nachev, P., Kennard, C., & Husain, M. (2008). Functional role of the supplementary and pre-supplementary motor areas. *Nature Reviews Neuroscience*, *9*, 856–869. https://doi.org/10.1038/nrn2478, PubMed: 18843271

Nachum, O., Gu, S., Lee, H., & Levine, S. (2018). Data-efficient hierarchical reinforcement learning. *Paper presented at the Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montréal, Canada. https://doi.org/10.48550/arXiv.1805.08296

Nakamura, K., Sakai, K., & Hikosaka, O. (1998). Neuronal activity in medial frontal cortex during learning of sequential procedures. *Journal of Neurophysiology*, *80*, 2671–2687. https://doi.org/10.1152/jn.1998.80.5.2671, PubMed: 9819272

Neftci, E. O., & Averbeck, B. B. (2019). Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence*, *1*, 133–143. https://doi.org/10.1038/s42256-019-0025-4

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, *19*, 1–32. https://doi.org/10.1016/0010-0285(87)90002-8

O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology*, *14*, 769–776. https://doi.org/10.1016/j.conb.2004.10.016, PubMed: 15582382

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*, 452–454. https://doi.org/10.1126/science.1094285, PubMed: 15087550

Ostlund, S. B., Winterbauer, N. E., & Balleine, B. W. (2009). Evidence of action sequence chunking in goal-directed instrumental conditioning and its dependence on the dorsomedial prefrontal cortex. *Journal of Neuroscience*, *29*, 8280–8287. https://doi.org/10.1523/jneurosci.1176-09.2009, PubMed: 19553467

Owen, A. M., Stern, C. E., Look, R. B., Tracey, I., Rosen, B. R., & Petrides, M. (1998). Functional organization of spatial and nonspatial working memory processing within the human lateral frontal cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *95*, 7721–7726. https://doi.org/10.1073/pnas.95.13.7721, PubMed: 9636217

Parthasarathy, H. B., Schall, J. D., & Graybiel, A. M. (1992). Distributed but convergent ordering of corticostriatal projections: Analysis of the frontal eye field and the supplementary eye field in the macaque monkey. *Journal of Neuroscience*, *12*, 4468–4488. https://doi.org/10.1523/JNEUROSCI.12-11-04468.1992, PubMed: 1279139

Peer, M., Brunec, I. K., Newcombe, N. S., & Epstein, R. A. (2021). Structuring knowledge with cognitive maps and cognitive graphs. *Trends in Cognitive Sciences*, *25*, 37–54. https://doi.org/10.1016/j.tics.2020.10.004, PubMed: 33248898

Poldrack, R. A., Clark, J., Paré-Blagoev, E. J., Shohamy, D., Creso Moyano, J., Myers, C., et al. (2001). Interactive memory systems in the human brain. *Nature*, *414*, 546–550. https://doi.org/10.1038/35107080, PubMed: 11734855

Poldrack, R. A., Sabb, F. W., Foerde, K., Tom, S. M., Asarnow, R. F., Bookheimer, S. Y., et al. (2005). The neural correlates of motor skill automaticity. *Journal of Neuroscience*, *25*, 5356–5364. https://doi.org/10.1523/JNEUROSCI.3880-04.2005, PubMed: 15930384

Popp, N., Yokoi, A., Gribble, P., & Diedrichsen, J. (2020). The effect of instruction on motor skill learning. *Journal of Neurophysiology*, *124*, 1449–1457. https://doi.org/10.1152/jn.00271.2020, PubMed: 32997556

Ramkumar, P., Acuna, D., Berniker, M., Grafton, S., Turner, R., & Kording, K. (2016). Chunking as the result of an efficiency computation trade-off. *Nature Communications*, *7*, 12176. https://doi.org/10.1038/ncomms12176, PubMed: 27397420

Rasmussen, D., Voelker, A., & Eliasmith, C. (2017). A neural model of hierarchical reinforcement learning. *PLoS One*, *12*, e0180234. https://doi.org/10.1371/journal.pone.0180234, PubMed: 28683111

Rhodes, B. J., Bullock, D., Verwey, W. B., Averbeck, B. B., & Page, M. P. (2004). Learning and production of movement sequences: Behavioral, neurophysiological, and modeling perspectives. *Human Movement Science*, *23*, 699–746. https://doi.org/10.1016/j.humov.2004.10.008, PubMed: 15589629

Ribas-Fernandes, J. J. F., Shahnazian, D., Holroyd, C. B., & Botvinick, M. M. (2019). Subgoal- and goal-related reward prediction errors in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *31*, 8–23. https://doi.org/10.1162/jocn_a_01341, PubMed: 30240308

Ribas-Fernandes, J. J., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., et al. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*, 370–379. https://doi.org/10.1016/j.neuron.2011.05.042, PubMed: 21791294

Sakai, K., Kitaguchi, K., & Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Experimental Brain Research*, *152*, 229–242. https://doi.org/10.1007/s00221-003-1548-8, PubMed: 12879170

Samejima, K., & Doya, K. (2007). Multiple representations of belief states and action values in corticobasal ganglia loops. *Annals of the New York Academy of Sciences*, *1104*, 213–228. https://doi.org/10.1196/annals.1390.024, PubMed: 17435124

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*, 486–492. https://doi.org/10.1038/nn.3331, PubMed: 23416451

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599. https://doi.org/10.1126/science.275.5306.1593, PubMed: 9054347

Seidler, R. D. (2010). Neural correlates of motor learning, transfer of learning, and learning to learn. *Exercise and Sport Sciences Reviews*, *38*, 3–9. https://doi.org/10.1097/JES.0b013e3181c5cce7, PubMed: 20016293

Seidler, R. D., Kwak, Y., Fling, B. W., & Bernard, J. A. (2013). Neurocognitive mechanisms of error-based motor learning. *Advances in Experimental Medicine and Biology*, *782*, 39–60. https://doi.org/10.1007/978-1-4614-5465-6_3, PubMed: 23296480

Selemon, L. D., & Goldman-Rakic, P. S. (1988). Common cortical and subcortical targets of the dorsolateral prefrontal and posterior parietal cortices in the rhesus monkey: Evidence for a distributed neural network subserving spatially guided behavior. *Journal of Neuroscience*, *8*, 4049–4068. https://doi.org/10.1523/jneurosci.08-11-04049.1988, PubMed: 2846794

Seo, M., Lee, E., & Averbeck, B. B. (2012). Action selection and action value in frontal–striatal circuits. *Neuron*, *74*, 947–960. https://doi.org/10.1016/j.neuron.2012.03.037, PubMed: 22681697

Shima, K., Isoda, M., Mushiake, H., & Tanji, J. (2007). Categorization of behavioral sequences in the prefrontal cortex. *Nature*, *445*, 315–318. https://doi.org/10.1038/nature05470, PubMed: 17183266

Shima, K., & Tanji, J. (1998). Both supplementary and presupplementary motor areas are crucial for the temporal organization of multiple movements. *Journal of Neurophysiology*, *80*, 3247–3260. https://doi.org/10.1152/jn.1998.80.6.3247, PubMed: 9862919

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. https://doi.org/10.1152/jn.2000.84.4.2148, PubMed: 11024102

Şimşek, Ö., Wolfe, A. P., & Barto, A. G. (2005). Identifying useful subgoals in reinforcement learning by local graph partitioning. *Paper presented at the Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany. https://doi.org/10.1145/1102351.1102454

Smith, K. S., & Graybiel, A. M. (2013). A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron*, *79*, 361–374. https://doi.org/10.1016/j.neuron.2013.05.038, PubMed: 23810540

Smith, K. S., & Graybiel, A. M. (2016). Habit formation. *Dialogues in Clinical Neuroscience*, *18*, 33–43. https://doi.org/10.31887/DCNS.2016.18.1/ksmith, PubMed: 27069378

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., et al. (2014). Optimal behavioral hierarchy. *PLoS Computational Biology*, *10*, e1003779. https://doi.org/10.1371/journal.pcbi.1003779, PubMed: 25122479

Sutton, R. S. (1999). *Reinforcement learning: Past, present and future*. Berlin/Heidelberg, Germany: Springer-Verlag. https://doi.org/10.1007/3-540-48873-1_26

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. https://doi.org/10.1016/S0004-3702(99)00052-1

Tang, H., Bartolo, R., & Averbeck, B. B. (2021). Reward-related choices determine information timing and flow across macaque lateral prefrontal cortex. *Nature Communications*, *12*, 894. https://doi.org/10.1038/s41467-021-20943-9, PubMed: 33563989

Tang, H., Costa, V. D., Bartolo, R., & Averbeck, B. B. (2022). Differential coding of goals and actions in ventral and dorsal corticostriatal circuits during goal-directed behavior. *Cell Reports*, *38*, 110198. https://doi.org/10.1016/j.celrep.2021.110198, PubMed: 34986350

Tanji, J., & Shima, K. (1994). Role for supplementary motor area cells in planning several movements ahead. *Nature*, *371*, 413–416. https://doi.org/10.1038/371413a0, PubMed: 8090219

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2020). Discovery of hierarchical representations for efficient planning. *PLoS Computational Biology*, *16*, e1007594. https://doi.org/10.1371/journal.pcbi.1007594, PubMed: 32251444

Trach, J. E., McKim, T. H., & Desrochers, T. M. (2021). Abstract sequential task control is facilitated by practice and embedded motor sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*, 1638–1659. https://doi.org/10.1037/xlm0001004, PubMed: 34516207

Tremblay, P. L., Bedard, M. A., Langlois, D., Blanchet, P. J., Lemay, M., & Parent, M. (2010). Movement chunking during sequence learning is a dopamine-dependant process: A study conducted in Parkinson's disease. *Experimental Brain Research*, *205*, 375–385. https://doi.org/10.1007/s00221-010-2372-6, PubMed: 20680249

Tremblay, P.-L., Bedard, M.-A., Levesque, M., Chebli, M., Parent, M., Courtemanche, R., et al. (2009). Motor sequence learning in primate: Role of the D2 receptor in movement chunking during consolidation. *Behavioural Brain Research*, *198*, 231–239. https://doi.org/10.1016/j.bbr.2008.11.002, PubMed: 19041898

Verstynen, T., Phillips, J., Braun, E., Workman, B., Schunn, C., & Schneider, W. (2012). Dynamic sensorimotor planning during long-term sequence learning: The role of variability, response chunking and planning errors. *PLoS One*, *7*, e47336. https://doi.org/10.1371/journal.pone.0047336, PubMed: 23056630

Verwey, W. B. (1996). Buffer loading and chunking in sequential keypressing. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 544–562. https://doi.org/10.1037/0096-1523.22.3.544

Verwey, W. B. (2001). Concatenating familiar movement sequences: The versatile cognitive processor. *Acta Psychologica*, *106*, 69–95. https://doi.org/10.1016/s0001-6918(00)00027-5, PubMed: 11256340

Verwey, W. B., & Abrahamse, E. L. (2012). Distinct modes of executing movement sequences: Reacting, associating, and chunking. *Acta Psychologica*, *140*, 274–282. https://doi.org/10.1016/j.actpsy.2012.05.007, PubMed: 22705631

Verwey, W. B., Jouen, A. L., Dominey, P. F., & Ventre-Dominey, J. (2019). Explaining the neural activity distribution associated with discrete movement sequences: Evidence for parallel functional systems. *Cognitive, Affective & Behavioral Neuroscience*, *19*, 138–153. https://doi.org/10.3758/s13415-018-00651-6, PubMed: 30406305

Verwey, W. B., & Wright, D. L. (2014). Learning a keying sequence you never executed: Evidence for independent associative and motor chunk learning. *Acta Psychologica*, *151*, 24–31. https://doi.org/10.1016/j.actpsy.2014.05.017, PubMed: 24929277

Wiestler, T., & Diedrichsen, J. (2013). Skill learning strengthens cortical representations of motor sequences. *eLife*, *2*, e00801. https://doi.org/10.7554/eLife.00801, PubMed: 23853714

Wymbs, N. F., Bassett, D. S., Mucha, P. J., Porter, M. A., & Grafton, S. T. (2012). Differential recruitment of the sensorimotor putamen and frontoparietal cortex during motor chunking in humans. *Neuron*, *74*, 936–946. https://doi.org/10.1016/j.neuron.2012.03.038, PubMed: 22681696

Xia, L., & Collins, A. G. E. (2021). Temporal and state abstractions for efficient learning, transfer, and composition in humans. *Psychological Review*, *128*, 643–666. https://doi.org/10.1037/rev0000295, PubMed: 34014709

Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X.-J., et al. (2022). Geometry of sequence working memory in macaque prefrontal cortex. *Science*, *375*, 632–639. https://doi.org/10.1126/science.abm0204, PubMed: 35143322

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, *19*, 181–189. https://doi.org/10.1111/j.1460-9568.2004.03095.x, PubMed: 14750976

Yokoi, A., & Diedrichsen, J. (2019). Neural organization of hierarchical motor sequence representations in the human neocortex. *Neuron*, *103*, 1178–1190. https://doi.org/10.1016/j.neuron.2019.06.017, PubMed: 31345643