CHAPTER 5

c0005 Learning Structures Through Reinforcement

Anne Collins
UC Berkeley, Berkeley, CA, United States

s0010 INTRODUCTION

[AU1]

p0010 The flexible and efficient decision-making that characterizes human behavior requires quick adaptation to changes in the environment and good use of gathered information. Thus, investigating the mechanisms by which humans learn complex behaviors is critical to understanding goal-directed decision-making. In the past 20 years, cognitive neuroscience has progressed immensely in understanding how humans learn from rewards and punishment, particularly for simpler behaviors shared in common with other mammals, such as learning simple associations between stimuli and actions. Reinforcement learning (RL) theory (Sutton & Barto, 1998) has provided a crucial theoretical framework explaining how humans learn to represent the value of choices and/or make decisions that are more likely to lead to rewards than to punishments. However, both cognitive neuroscience and artificial intelligence fields struggle with explaining more complex, and more characteristically human, learning behaviors, such as rapid learning in completely new and complex environments.

p0015 This chapter discusses the use of the RL framework to understand many complex learning behaviors, focusing specifically on model-free RL algorithms for learning values of or policies over states and actions, since we have a good understanding of how cortico—basal ganglia loops use dopaminergic input to implement an approximate [AU2] form of this computation. We will show that many forms of complex human RL can be framed by applying this RL computation, provided that we model the inputs and outputs of the algorithm appropriately. Specifically, we argue that by better defining the state and action spaces for which humans learn values or policies, we can broadly widen the types of behaviors for which RL can account. We support this statement with examples from the literature showing how the brain may be performing the same computations for different types of inputs/outputs and how this can account for complex behavior, such as hierarchical RL (HRL), structure learning, generalization, and transfer.

p0020 We will first provide a short introduction to RL, both from a computational point of view, highlighting the limitations and difficulties encountered by this algorithm, and from a cognitive neuroscience point of view, mapping these computations to neural

Goal-Directed Decision Making
ISBN 978-0-12-812098-9, https://doi.org/10.1016/B978-0-12-812098-9.00005-X

10005-MORRIS-9780128120989

mechanisms. We will then attempt to unify multiple frameworks from the human learning literature, such as representation learning (Wilson & Niv, 2012), HRL (Botvinick, Niv, & Barto, 2009), rule learning (Collins & Koechlin, 2012), and structure learning (Collins & Frank, 2013), into a single framework, whereby the brain uses a single mechanistic computation—defined by a model-free RL mechanism—and applies it to different input and output spaces, notably, state and action spaces. We will first focus on how we can mitigate the curse of dimensionality by altering how we define state spaces, leading to more complex and efficient learning. We will then show that assuming different action spaces, in particular, by introducing temporal abstraction or rule abstraction, leads to faster learning and to an ability to generalize information. Last, we will show that humans sometimes create latent state or action spaces, which seemingly makes learning problems more complicated but comes with a number of behavioral advantages. Finally, we will conclude by broadening to other open questions in flexible learning: the role of the reward function in RL, the various algorithms other than model-free RL that may also contribute to efficient learning, and the roles of models of the environment in learning.

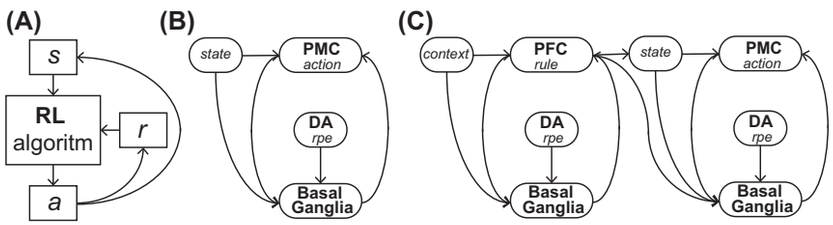
s0015

REINFORCEMENT LEARNING

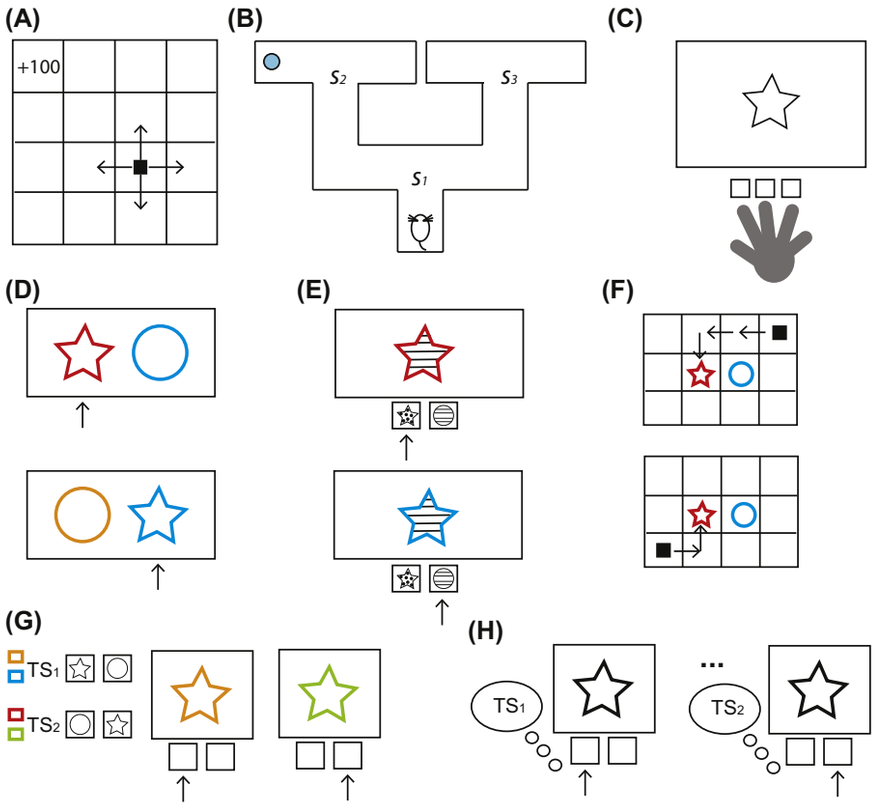
s0020

Reinforcement learning algorithms

p0025 RL models are a class of algorithms designed to solve specific kinds of learning problems for an agent interacting with an environment that provides rewards and/or punishments (Fig. 5.1A). The following type of "grid world" problem exemplifies an archetypical RL problem (Fig. 5.2A). The agent (black square) sits in one of the cells of a grid environment and can navigate through the grid by choosing one of four actions (up, down,



f0010 **Figure 5.1** ***Schematic of reinforcement learning (RL) systems.*** (A) RL algorithms observe a state ***s*** as input and select an action ***a*** as output. The environment provides reinforcement ***r***, which is used to update the RL algorithm and transitions to the next state. (B) An approximation of these computations is performed in the cortico–basal ganglia loop (Frank et al., 2004). For example, a sensory observation leads to preactivation of possible actions in the premotor cortex (PMC); the PMC–basal ganglia loops allow gating of one action; dopamine (DA) signals a reward prediction error (RPE) signal that reinforces corticostriatal synapses, allowing the gating mechanism to select the actions most likely to lead to reward. (C) This learning process occurs at multiple hierarchical levels in the brain in parallel (Collins & Frank, 2013). For example, loops involving the prefrontal cortex allow learning to occur between abstract contexts and high-level rules, which then constrains the lower-level learning loop.



f0015 **Figure 5.2** ***Examples of reinforcement learning (RL) problems.*** (A) Grid world: The artificial agent navigates between cells using one of four directions. (B) Animals navigate in a maze to obtain reward; the states s_i are physical locations. (C) Instrumental learning task: Participants use reward feedback to learn to select the correct button for each possible stimulus (e.g., shapes). (D) Representation learning task: Participants need to select one of two patterns; only one dimension matters (here, shape matters, with the star being the most rewarding of the two shapes). (E) Hierarchical learning: Participants learn that for one color (red—top), the shape of the input determines the correct action, but for the other color (blue—bottom), the texture determines the correct action. (F) Options framework or hierarchical RL (HRL): In both cases, participants select the same high-level action (or option): go to the star. This constrains a different sequence of low-level actions. (G) Structure learning: Participants learn to select one high-level abstract action (a rule, or task set, TS1) for some colors and another (TS2) for other colors; in parallel, they learn to associate low-level actions (button presses) to stimuli (here, shapes) for each of the high-level abstract rules. (H) Latent rule learning: Participants learn high-level rules as in (G) but do not observe the contexts. Instead, they infer the latent context from their observations of the outcomes to their choices.

left, or right). The agent can collect points by selecting some actions or by entering some cells (e.g., the top-left corner in Fig. 5.2A). The goal of the agent is to maximize points earned. Defined more technically, an RL problem is characterized by a state space $\boldsymbol{S}$ (here, the cells in the grid world), an action space $\boldsymbol{A}$ (here, the four available actions),

a transition function $T(s,a,s') = p(s'|s,a)$ that controls the probability of the next state s' given that the agent chose action a in a state s and a reward function $R(s,a,s')$. The goal of the agent is to optimize the expected sum of future discounted rewards and, specifically, to find a policy $\pi(s,a) = p(a|s)$ that maximizes this sum. One way to achieve this goal is to estimate the expected value of each state or of each state and action under the optimal policy (where value is the expected sum of discounted future rewards). If one can do this, the optimal policy falls out by selecting the action with the highest value.

p0030 There are many different algorithms that propose solutions to this problem and offer guarantees of convergence. We focus on a simple class of algorithms, called model-free because they do not require a model of the environment (i.e., knowledge of the transition function and the reward function). We focus on model-free RL algorithms, such as temporal difference learning, Q-learning, SARSA, and actor-critic algorithms (Sutton & Barto, 1998), because they have been extremely helpful in understanding animal behavior and neural correlates of learning. Model-free RL algorithms use a key quantity, called the reward prediction error, to learn to estimate values of states or of state–action pairs. At each trial t, the reward prediction error is defined as the difference between what is expected for future discounted reward after taking an action (the sum of reward r and discounted value of next step $\gamma V(s_{t+1})$, where γ is the discount factor) and what was expected prior to taking that action ($V(s_t)$). Using the reward prediction error $rpe = r + \gamma V(s_{t+1}) - V(s_t)$ to update the previous estimate of $V(s_t)$ by a small increment of the error $\alpha.rpe$ (where α is the learning rate) is a good algorithm under certain assumptions and constraints (Sutton & Barto, 1998).

s0025 Reinforcement learning in the brain

p0035 Through this model-free RL algorithm, an artificial agent can learn the optimal way to attain a reward in the simple grid world of the example in Fig. 5.2A, after many attempts to solve this problem (Sutton & Barto, 1998). It is a good model of behavior for an animal learning to find its way toward a reward in a maze (Fig. 5.2B). Further, this algorithm has been a critical source of progress in the cognitive neuroscience of learning because it provides a useful model of the neural correlates of RL. Specifically, researchers have discovered that dopaminergic neurons fire in a pattern that is consistent with a reward prediction error signal: Their firing increases phasically with unexpected reward, decreases phasically with missed expected reward or with unexpected punishment, and stays at the tonic level for expected rewards (Montague, Dayan, & Sejnowski, 1996). Dopamine release in the striatum follows parametrically what would be expected for a bidirectional reward prediction error signal (Hart, Rutledge, Glimcher, & Phillips, 2014). Furthermore, dopamine signaling in the striatum modulates plasticity of corticostriatal synapses, with increased dopamine strengthening associations in the pathway facilitating action selection and decreasing them in the pathway blocking it; decreased dopamine has the opposite effects in these pathways (Adamantidis et al., 2011; Hamid

et al., 2015; Kravitz, Tye, & Kreitzer, 2012; Tai, Lee, Benavidez, Bonci, & Wilbrecht, 2012). Cortico–basal ganglia loops act as a gate for action selection that is dependent on the strength of these two corticostriatal pathways (Collins & Frank, 2014; Frank, Seeberger, & O'Reilly, 2004). Thus, there is strong evidence that cortico–basal ganglia loops implement a model-free RL computation, with dopamine reward prediction errors training corticostriatal associations to help select choices that lead to reward and avoid those that lead to punishment (Fig. 5.1B).

s0030 Limitations

p0040 Model-free RL algorithms are thus very successful at explaining animal learning because they capture many behaviors well, including for example, probabilistic reward learning (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007), and they have a plausible mechanistic implementation in the brain. Using these computational models to link between brain and behavior has increased understanding of individual differences in RL, of learning deficit in some pathologies (e.g., Parkinson) and of the effect of dopaminergic drugs on learning (Frank, 2005). However, model-free RL also has a number of limitations that have led cognitive neuroscience and artificial intelligence researchers to look at other algorithms to better model human learning and enhance artificial agents, respectively. One major limitation of RL is that it suffers from the curse of dimensionality: While RL can be relatively efficient in small problem spaces, learning with this algorithm in relatively bigger problem spaces would take an enormous amount of practice, making it extremely inefficient. In contrast, humans can often learn new behaviors very quickly (e.g., how to drive a car). Another limitation is that model-free RL is inflexible: When the environment changes (e.g., the position of the reward in the grid world), model-free RL algorithms need to slowly unlearn. By contrast, humans (and animals) are sensitive to changes in the environment and can quickly alter their behavior toward their goal. To solve these and other limitations of model-free RL algorithms, researchers in artificial intelligence and cognitive neuroscience have proposed new algorithms. For example, model-based RL algorithms offer some solutions to the inflexibility problem by proposing a different way of computing expected values that integrates knowledge about the model of the world. However, we will show here that we can understand many complex human behaviors in the framework of the same simple model-free algorithm, with its grounding in a well-understood neural implementation, by carefully considering the state and action spaces over which model-free computations of estimated values or policies are performed.

s0035 Framing the problem

p0045 What are state and action spaces when modeling human behavior? This modeling choice is often dictated by the experimental design and is assumed away as obvious. We give some examples in Fig. 5.2B and C. The most direct translation from original RL

algorithms, such as grid worlds, is the modeling of spatial learning tasks in which animals need to learn to find a reward in a maze. States are modeled as discrete places in the maze at which a decision is needed, and actions are modeled as choices of direction (e.g., left or right; Fig. 5.2B). Note that making different choices for the state/action space could lead to a very different model (e.g., with more discrete places in the maze, actions could include stop, groom, etc.). For human behavior, state spaces are often replaced with sets of stimuli, and actions are replaced with simple choices, such as key presses (Fig. 5.2C); this modeling choice retains a fairly unambiguous interpretation of the environment. Probabilistic reward learning tasks offer a good example of the ambiguity of defining state/action spaces. In these tasks (e.g., Davidow, Foerde, Galvan, & Shohamy, 2016; Frank et al., 2004), subjects may be asked to choose between two shapes (e.g., Fig. 5.2D). There is some ambiguity in how this task should be modeled. Is the state the current pair of stimuli? Is this pair dependent or independent of their left/right position? More generally, this task tends to be modeled as a single state and two actions: "picking the star" or "picking the circle." It is important to note that (1) this action state is much more abstract than "press the left/right button," as it does not map to a single set of motor commands, and (2) a different choice, for example, "pick left" versus "pick right," would be unable to capture behavior in this task, since left and right are not informative about reward. Despite the abstraction of this action space, the model-free RL algorithm excels at capturing the behavior and neural effects in this task (Davidow et al., 2016). We show here that we can capture many behaviors of higher complexity in the model-free RL framework by carefully considering the state and action spaces over which the computations occur. Table 5.1 shows in pseudocode how this can be done in the examples of Fig. 5.2. We will show that developing appropriate states and action spaces overcomes many issues thought of as classic limitations of model-free RL.

s0040

STATE SPACES

s0045

Simplifying the state space

p0050 Figuring out an appropriate state space over which RL operates can dramatically improve RL performance by reducing the curse of dimensionality. Learning to drive is a task, which teenagers may accomplish in a few hours but which many top artificial intelligence researchers and companies have been unable to get an artificial agent to perform without major issues. How do we use our experience from 15 years of life to accomplish such fast learning? Taking all visual inputs into account would be overwhelming to a learning agent, as we essentially never see the same scene twice when driving. However, if one can discern that the relevant information for making a decision whether to stop or to go at an intersection is the color of the light (red, yellow, or green) then one part of the problem is suddenly reduced to a one-dimensional, two-feature state space. Niv and colleagues investigated such state space learning in a series of studies (Leong et al., 2017;

t0010 **Table 5.1** Pseudocode for learning examples in Fig. 5.2

A) Grid world	RL(S={all (x_i,y_j)}, A={up, down, left, right})
B) Maze	RL(S={all (x_i,y_j)}, A={forward, left, right})
C) Instrumental learning	RL(S={star, circle}, A={button1, button2, button3})
D) Representation learning	RL(S={all (shape, color, texture)}, A={left, right}) RL(S={star, circle}, A={left, right})
E) Hierarchical reinforcement learning	RL(S={all (shape, color, texture)}, A={left, right}) RL(S_1={colors}, A_1={attend(texture) = RL(S_2={texture}, A_2={left, right}), attend(shape) = RL(S_3={shape}, A_2={left, right})}))
F) Options— hierarchical reinforcement learning	RL(S={all (x_i,y_j)}, A={up, down, left, right}) RL(S={all (x_i,y_j)}, A_1={go to circle = RL(S={all (x_i,y_j)}, A_2=A), go to star = RL(S={all (x_i,y_j)}, A_2=A)})})
G) Structure learning	RL(S={all (color, shape)}, A={button1, button2}) RL(S_1={colors}, A_1={policy1 = RL(S_2={shapes}, A_2={button1, button2}), policy2 = RL(S_2={shapes}, A_2={button1, button2})})
H) Latent rule learning	RL(S={shapes}, A={button1, button2}) RL(S_1={context1, context2, ...}, A_1={policy1 = RL(S_2={shapes}, A_2={button1, button2}), policy2 = RL(S_2={shapes}, A_2={button1, button2})})

RL represents a single learning algorithm producing a policy over given state or action spaces ***S, A***. Light blue is the "naïve" or flat modeling of a problem, using the simplest state spaces for inputs and action spaces for outputs. Black models structure learning, as observed in participants.

Niv et al., 2015; Wilson & Niv, 2012; see also Chapter 12 by Shuck, Wilson, and Niv), and a simplified example is schematized in Fig. 5.2D. At each trial, participants were shown three items and needed to choose one item to try to win points. Each item had three dimensions (shape, color, and texture), and each dimension had three features (e.g., red, blue, and green). In a learning problem, only one feature from one dimension (e.g., the star) had a high likelihood of leading to reward; thus, if participants were able to learn that the other two dimensions did not matter and that they should learn to represent the problem as an RL problem concerned only with shapes, they could significantly simplify the dimensionality of the problem and thus improve their performance (Wilson & Niv, 2012). Results showed that behavior was best explained by a

process where participants learned to focus their attention on a single dimension and applied simple RL to features of this dimension. Thus, they effectively created a relevant, smaller state space, over which an RL algorithm was run (Table 5.1D); indeed, reward prediction error signals in the striatum were better explained by assuming RL happened over the state space defined by the focus of attention than by other models. This is one of the most direct examples of how humans define nonobvious state spaces over which to learn values or policies with RL. An important question is how we create the state space itself; in the example given here, how do we learn the feature on which we should focus our attention? A study by Leong et al. (2017) showed that creation of state space can be performed using reward feedback, such that there is a bidirectional interaction: Attention told subjects over which dimensions they should perform RL, and reward prediction errors helped participants direct their attention to the correct dimension and thus create the state space over which to operate RL.

s0050 Multiple state spaces

p0055 A state space that is appropriate for one goal may not be appropriate for another. Consider our driving example with the traffic light: If you are in the lane to go straight, the main round lights are relevant to your decision to stop or go, but if you are in the lane to turn left, you should ignore these lights and instead pay attention to the left arrow lights. Said differently, your state space should be conditioned on an additional aspect of the environment: which lane you are in. Being able to create multiple state spaces and knowing the one to which you should apply RL would allow significantly more complex learning behavior. Indeed, it would allow a hierarchical contextualization of learning by context. A series of studies (Badre & Frank, 2011; Badre, Kayser, & Esposito, 2010; Frank & Badre, 2011) has shown that healthy young adults are able to hierarchically contextualize the learning space and that it strongly improves their learning. Participants saw a single three-dimensional item on the screen and had to learn which of three actions to pick to receive points. In a flat condition, all the three dimensions were needed to figure out the correct action for an item, leading to three-dimensional state spaces with an overwhelming 18 items. In a hierarchical condition, one of the dimensions (color) controlled which of the other two dimensions was relevant for learning (e.g., if the item was red, only the shape mattered, but if it was blue, only the texture mattered; Fig. 5.2E). Thus, participants could essentially build two small state spaces (one corresponding to three textures and another to three shapes) and at each trial determine which state space to use based on the color of the item (Table 5.1E). Badre et al. (2010) showed that participants did learn this way, as evidenced by much more efficient learning in the hierarchical condition than in the flat condition. Further, studies (Badre & Frank, 2011; Frank & Badre, 2011) have shown that this method of learning could be computationally understood as RL computations happening over two hierarchical loops and different state (and action—see below) spaces (Fig. 5.1B): The top loop learned

through RL which of the two state spaces to select for a given color, while the bottom loop learned which key press to select for either of the two simpler state spaces.

p0060 This example highlights a number of important points. First, RL computations may happen over multiple state spaces in the same learning problem, with other signals serving as a contextualizing factor. Second, they may happen simultaneously over multiple state spaces (in the previous example, learning which state space to select for a color state and which key to press for a given shape or texture). This latter point implies two further important features: (1) a notion of hierarchy, whereby the choice from one of the RL loops has an influence over the learning and decision of a "lower-level" loop and (2) the choice in the higher hierarchical loop is more abstract than the one at the lower level—indeed, in this example, RL in the top loop happens not only on a subpart of the original state space (the color dimension) but also on a new abstract action space, indicating the dimension on which a subject must focus attention. Below, we will come back to hierarchical representations in RL and to the importance of learning action spaces, in addition to state spaces.

s0055 ACTION SPACES

s0060 Abstract hierarchical action spaces

p0065 The study by Frank and Badre (2011), discussed above, showed that learning the hierarchical structure of the environment, which simplifies a large unstructured state space into two smaller state spaces selected conditionally on a context, can facilitate learning. It introduced the need to operate RL not only over multiple state spaces but also over an abstract action space, where the action is the decision of which lower-level state space to use. More generally, other complex learning behavior can be obtained by this combination of two characteristics: (1) RL at multiple hierarchical levels simultaneously and (2) RL over abstract higher-level action spaces that control lower-level decisions. In that sense, the higher-level actions are themselves policies mapping lower-level stimuli to lower-level actions. A body of work extended the previous notion of HRL by showing that such abstract actions could be more than just attentional filters (i.e., the dimension of the input to which I should focus my attention for making my decision), and could instead be abstract policies, also called "rules" or task sets (Collins & Frank, [AU3] 2016a,b, 2013; Collins & Koechlin, 2012). Specifically, similarly to the studies of Badre and colleagues, these studies showed that participants learned to make a choice at a higher level in response to a feature of the environment (e.g., a color) and that the higher-level choice constrained answers to other features of the environment. However, in this case, the higher-level choice was not that one should focus on one dimension and neglect another dimension—indeed stimuli were only two-dimensional. Rather, the higher-level choice constrained the correct set of choices for the features of the second dimension (Fig. 5.2G).

p0070 Going back to the driving example, whether you are in France or in the United Kingdom, you need to pay attention to all the same visual signals to drive correctly. However, the actions you take in answer to these signals depend on the context: arriving at a circle in France requires you to turn right, but the same in the United Kingdom requires you to turn left. Thus, more complex behavior sometimes requires us not only to use context to determine where to focus our attention but also to determine how to respond to the focus of our attention. We showed that participants create such high-level abstract choice spaces, where choices correspond to this high-level policy choice; we call them rules or task sets (Collins & Frank, 2016a,b, 2013; Collins & Koechlin, 2012, Table 5.1G). Creating rules that one selects in response to a context, but that are not bound or equated to that context, is a critical factor in flexible, efficient learning. Indeed, because participants created these choice spaces, they were also able to try these choices in new contexts; this means that they were able to generalize a high-level policy to a new context (for example, the rules of driving in France apply mostly as a whole to driving in Germany). Furthermore, the new associations were stored by the policy learned at the lower level, constrained by the higher-level choice, without being tied to the context in which it was learned. Thus, participants were able to transfer knowledge learned in one context to other contexts that required selecting the same rule (for example, after having observed that driving is similar in Boston and Berkeley, learning how to handle a four-way stop in one location would immediately transfer to the other).

p0075 Creating an abstract action space (where actions are rules or task sets and can be viewed as a policy over another state action space) greatly increases the flexibility and efficiency of learning because it allows generalization and transfer. It also provides some form of *divide and conquer*, whereby a complicated decision over a large state space (all possible input features) is transformed into a series of simpler, hierarchical decisions: first selecting a rule in response to a context; then, given that rule, selecting an action in response to a stimulus. We showed with computational modeling and electroencephalography (Collins, Cavanagh, & Frank, 2014; Collins & Frank, 2016a,b, 2013) that this process can be performed in a model that applies RL computations in hierarchical cortico–basal ganglia loops (Fig. 5.1C). Thus, such hierarchical structure learning can also be understood as RL over appropriate state (at multiple hierarchical levels) and action (at multiple abstraction levels) spaces.

s0065 Temporally abstract actions

p0080 The previously described form of RL is clearly hierarchical: It consists of selecting a higher-level rule, which is really a policy in that it constrains selection of actions at the lower level. This feature allows us to draw a parallel to a specific class of algorithms that are known in the literature as "HRL," also called the "options framework." The

options framework also seeks to improve on simple RL mechanisms by building a more complex action space and, specifically, by introducing options. Options can be seen as local policies or hierarchical actions (Table 5.1F). In the simplest case, options correspond to a class of sequences of simple actions that lead to a subgoal. For example, reaching the door of a room in a grid world is a high-level option and may define a local policy (how to reach a door from any point in the room or the star in the example of Fig. 5.2F). In the driving example, an example of a high-level option is shifting gears. You may learn at the high level when to shift or not to shift gears, but then once you select that option, it requires a series of lower-level actions (engage the clutch, shift the gear, then release the clutch) over which you can also learn.

p0085 Using options can partially solve the curse of dimensionality by facilitating exploration (Botvinick et al., 2009). Indeed, a single higher-level choice may lead an agent to explore further and more efficiently. Options also capture an important feature of human sequential behavior, which often includes hierarchical sequences of actions. A few studies have shown evidence of human learning and neural computations being well explained by the options framework, whereby learning happens hierarchically, both for the option itself and for the actions within the option (Diuk, Tsai, Wallis, Botvinick, & Niv, 2013; Ribas-Fernandes et al., 2011; Solway et al., 2014). In these studies, participants made choices in sequential environments that provided a possibility for HRL. Further, these studies showed evidence in the brain for reward prediction errors corresponding to learning over both action spaces (within the option policy and at the higher hierarchical level).

s0070 LATENT STATE AND ACTION SPACES

p0090 We have shown that many complex learning behaviors can be explained as applying a simple model-free RL algorithm to the correct state and action space, or sometimes as applying more than one RL computation to multiple appropriate state and action spaces in parallel. An interesting feature is that in hierarchical forms of RL (structure learning, options framework, and hierarchical rule learning), the higher-level action space is abstract in the form of a policy. In particular, it cannot be described as a concrete motor action. Here, we show that abstraction in the state space can also help understand more complex learning behaviors. In particular, assuming unobservable, or latent, states can greatly enhance the flexibility of the learning agent (Gershman, Norman, & Niv, 2015). For example, if you are driving in winter, you might not be able to see that the road is icy, but if you observe that your usual actions lead to undesirable consequences (slipping), you might deduce that the latent cause in the environment is the weather and adapt your behavior based on this latent cause. This example captures some of the important features for which RL over latent states or causes can better explain human learning: when the contingencies of the environment change suddenly but not in an observable

way (e.g., in reversal learning experiments (Hampton, Bossaerts, & O'Doherty, 2006)), an RL agent operating over the observable state space needs to unlearn previous associations before being able to learn new associations. By contrast, humans may identify a change point, infer a new unobservable context or latent cause, and learn over this state. Several studies (Gershman, Blei, & Niv, 2010; Gershman et al., 2015; Soto, Gershman, & Niv, 2014) have shown how this assumption can explain a number of learning phenomena, such as extinction and compound generalization.

p0095 Latent spaces enrich the state representations over which RL operates. In combination with other previously described mechanisms, such as abstract action spaces (rules) that hierarchically constrain simultaneous learning over other state and action spaces, the mechanism of creating latent spaces provides an explanation for additional aspects of human fast and flexible learning. One behavioral study (Collins & Koechlin, 2012) had participants learn associations between one-dimensional stimuli and actions (task sets) using probabilistic reward feedback (Fig. 5.2H). The task sets changed periodically without warning and, unbeknownst to participants, could be reused as a whole later in the experiment.

p0100 Results showed that participants were able to create both an abstract action space of task sets and an abstract state space of latent temporal contexts (Table 5.1H); they identified the current temporal context as a state in which a given task set was to be selected, constraining RL over association between an observable state space (stimuli) and actions (key presses). Furthermore, when they identified a new temporal context (after an inferred switch in the environmental contingencies), they explored in the abstract action space of task sets, reselecting previously learned strategies as a whole, rather than exploring only in the low-level state space (Collins & Koechlin, 2012; Donoso, Collins, & Koechlin, 2014). This strategy allowed participants to transfer task sets to new contexts and thus to adapt more quickly than they would have otherwise.

p0105 The examples given above show that much of complex human learning does not require any learning algorithm more complex than model-free RL, provided that the latter algorithm is applied to the right inputs and outputs (state and action spaces). This process may require (1) running this algorithm over more than one set of spaces in parallel, a task for which the cortico–basal ganglia loops are well configured (Alexander & DeLong, 1986), and (2) using hierarchical influence of one output over another input, for which the prefrontal cortex is well organized (Badre, 2008; Koechlin, Ody, & Kouneiher, 2003; Koechlin & Summerfield, 2007; Nee & D'Esposito, 2016). These features enable much more efficient and flexible learning than was originally thought possible with a simple model-free algorithm for RL value estimation. Specifically, they allow for fast and efficient exploration, improvement of performance by massive simplification of problems, and fast learning in new environments by generalization and transfer of information.

[**AU4**]

s0075 HOW DO WE CREATE THE STATE/ACTION SPACES?

p0110 Efficiently modeling complex human learning with model-free RL crucially relies on operating over the right state and action spaces. Using inappropriate spaces instead strongly impairs learning, as shown, for example, by Botvinick et al. (2009) in simulations where using incorrect options lead to slowed exploration. The question of how we acquire the appropriate state and action spaces for our current environments remains largely open, although the previous examples do suggest some potential mechanisms.

p0115 For learning state spaces when the optimal state space is a subspace of the full sensory space, some studies (Leong et al., 2017; Niv et al., 2015) suggest that we use a frontoparietal mechanism to focus attention specifically on that subspace and that we learn to do so using reinforcement. Frank and Badre (2011) suggest that the gating mechanisms of the prefrontal cortex–basal ganglia loops may learn which aspects of the environment to keep in working memory, as well as which items should be allowed to influence other loops, thus also using the simple RL mechanism to create the ad hoc state spaces required for HRL. Collins and Frank (2013) also showed that such mechanisms enabled the creation of abstract action spaces. Furthermore, there seems to be a strong bias toward learning occurring hierarchically. Specifically, some studies (Badre & Frank, 2011; Badre et al., 2010) have shown that participants engaged anterior portions of the prefrontal cortex a priori initially, even in problems that could not be simplified. Further, other studies (Collins & Frank, 2013; Collins et al., 2014) have shown that participants built a hierarchical abstract rule structure even in environments that did not immediately benefit from it, highlighting a more general drive toward this kind of organization. This bias toward hierarchical learning could be due to a prior belief that hierarchical structures are useful (Collins & Frank, 2016a,b) or to constraints that result from the way our hierarchical cortico–basal ganglia loops evolved from motor cortex–originating loops (Collins & Frank, 2016a,b), or, more likely, it could be due to both.

p0120 A series of models from Alexander, Brown, and colleagues (Alexander & Brown, 2011, 2014, 2015) also point out the potential importance of medial prefrontal cortex in learning rules for cognitive control. Their models assume that the medial prefrontal cortex learns to represent errors of prediction at various hierarchical levels, thus teaching the lateral prefrontal cortex to represent useful state and action spaces to minimize such errors of prediction. These models also resonate with work by Holroyd and colleagues (Holroyd & McClure, 2015; Holroyd & Yeung, 2012), which points out the importance of the anterior cingulate cortex (ACC) in extended motivated behavior. Specifically, they argue that the ACC enables HRL (in the sense of the options framework), whereby the hierarchy is in the choice of higher-level actions that constrain sequences of lower-level actions.

p0125 This HRL/options framework also raises the question of how the action space is created, or the "options discovery" problem: How do we create options that take us

to the doors of the room rather than to the windows? Theoretical work suggests that using pseudoreward when reaching a subgoal and using RL with this pseudoreward to learn the option may help option creation (Botvinick et al., 2009), and there is some evidence that such a mechanism may occur in the brain (Diuk et al., 2013; Ribas-Fernandes et al., 2011; Solway et al., 2014). However, how do we determine useful subgoals? Work by Schapiro and colleagues (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013; Schapiro, Turk-Browne, Botvinick, & Norman, 2016) has shown that humans are able to identify bottlenecks in the environments we navigate and, if given a chance, create options with these bottlenecks as subgoals, which might be one mechanism for creating a useful action space in the framework of options.

p0130 Interestingly, some methods for learning useful state and action spaces require a model of the environment. For example, creating useful options may require identifying bottlenecks in a mental map of the environment. In the case of latent state spaces, in particular, a model of the environment consists of a likelihood function, defining expected outcomes (for example rewards) in response to interactions with the environment under a given latent space (Collins & Koechlin, 2012; Gershman et al., 2015). Using this likelihood function allows both an inference about the current hidden state and the online creation of what the latent state space is (Collins & Frank, 2013; Gershman et al., 2010). It is important to note that these models are used to create a state space and to infer a state but that, despite this use of a model, the learning algorithm in operation may still [**AU5**] be a model-free RL algorithm. This highlights the blurry line between what we should label as model-free and model-based learning (see also Chapter 18 by Miller, Ludvig, Pezzulo, and Shenhav); most learning may use a model of the environment, even in the absence of a mechanism of forward planning, as is usually defined in formal model-based RL algorithms (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). RL with a model can reach many more types of behaviors than those usually understood by model-based RL.

s0080 OPEN QUESTIONS

p0135 We have shown that thinking of human learning as a simple computation occurring over well-tailored state and action spaces can explain many feats of flexible and efficient decision-making. However, many open questions remain, one of which we have already discussed: how these state and action spaces are built. Two other classes of questions also merit further research to better understand human learning. Learning from reinforcement requires four elements: a state and action space, a reward function, and an algorithm to learn a policy. We have focused here on the role of the state and action spaces and have just assumed a simple model-free RL algorithm and reward function. However, both learning algorithms and reward functions should be further investigated.

s0085 Reward function

p0140 Most RL experiments use primary or secondary rewards or punishments, such as food, pain, points, and money (gains or losses), as reinforcers. However, other features might also contribute to the reward function. For instance, theoretical and experimental results have suggested various "bonuses" to the reward function, related for example to novelty (Kakade & Dayan, 2002) and information (Bromberg-Martin & Hikosaka, 2009); these and other influences may be reflected in the dopamine reward prediction error signal (see also , Chapter 11 by Sharpe and Schoenbaum). Other results have shown costs in the form of mental effort and conflict (Cavanagh, Masters, Bath, & Frank, 2014; Kool & Botvinick, 2014; Westbrook & Braver, 2015; see also Chapter 7 by Kool, Cushman, and Gershman). Furthermore, the movement of gamification relies on the notion that learners are motivated by nonrewarding outcomes (e.g., stars) that mark the attainment of subgoals (Deterding, Dixon, Khaled, & Nacke, 2011; Hamari, Koivisto, & Sarsa, 2014). This notion relates to pseudoreward, which may be useful for learning options in the HRL framework: Maintaining motivation over extended behaviors when real reward is infrequent might require us to consider intermediary, symbolic subgoals as rewarding (Diuk et al., 2013; Ribas-Fernandes et al., 2011; Lieder & Griffiths, n.d.). Theoretical work has shown that this notion could tremendously improve learning in complex situations (Lieder & Griffiths, n.d.). Thus, future research in human learning should aim to better understand what outcomes contribute to the reward function used by the RL algorithm for learning and to determine whether humans manipulate this reward function beyond normal reward to create better representations of the learning problem.

s0090 Algorithms

p0145 Separating the algorithm of learning from its inputs and outputs—the state and action spaces—enables us to better understand how a rich collection of human learning behaviors can be explained with this framework. However, this argument should not be taken to mean that we propose the brain uses only the learning algorithm presented and exactly this algorithm to learn to make decisions from reward information. In fact, much remains poorly understood about the computations performed by the brain to learn policies. For the model-free RL algorithm, we understand that the cortico–basal ganglia loops with dopamine reward prediction errors approximate it, but many precise aspects of this computation remain under debate. For example, the direct and indirect pathways apparently have redundant roles in learning (Collins & Frank, 2014; Dunovan & Verstynen, 2016); more research is needed to better understand their distinct contributions to model-free RL.

p0150 Furthermore, it is very likely that the brain also uses, in parallel, other algorithms to learn policies from reward. One method is simple memorization of associations in

working memory, which accounts for part of learning from rewards in simple associative learning tasks (Collins & Frank, 2012; Collins, Ciullo, Frank, & Badre, 2017). Similarly, by allowing us to sample from past events, episodic memory may play an important role in policies learned from reward (Bornstein & Norman, 2017; Bornstein, Khaw, Shohamy, & Daw, 2017). Furthermore, there is also ample evidence that humans also perform model-based planning RL in parallel to model-free RL (Daw et al., 2011; Doll, Duncan, Simon, Shohamy, & Daw, 2015). Exactly how this prospective planning occurs, especially many steps ahead, is not well understood—it may depend on the use of heuristics to simplify the forward search (Huys et al., 2015) or inferential processes (Chapter 3 by Solway and Botvinick). Thus, much remains unknown about the algorithms themselves.

s0095 CONCLUSION

p0155 Human learning is incredibly efficient and flexible and does much to promote human intelligence and goal-directed behavior. In this chapter, we explored how a very simple family of algorithms—that we know are approximately implemented by a precise neural circuitry in the brain—can explain a surprisingly wide array of complex learning, unifying literature on HRL, the options framework, structure learning, and representation learning. Specifically, we show that this simple computation of expected value (or policy weight), obtained by incremental updates with reward prediction errors, can lead to very efficient learning, exploring, transfer, and generalization when applied to useful state and action spaces. Understanding how we construct these useful spaces and how we interlock multiple computational loops in parallel to learn at multiple levels simultaneously is a future challenge. One important point is that finding useful spaces is not simply a matter of simplifying the sensory and motor space by factoring it into lower-dimensional or discrete subspaces but can rather also involve making the spaces more complex—creating new states that are not a subspace of sensory and motor space but are abstract states and actions carrying more information about the structure of the problem. These state and action spaces of higher complexity can counter-intuitively lead to an eventual improvement in behavior by rendering decision-making more flexible and by providing useful subpolicies that achieve subgoals or other generalizable chunks of behavior.

REFERENCES

Adamantidis, A. R., Tsai, H.-C., Boutrel, B., Zhang, F., Stuber, G. D., Budygin, E. A., … de Lecea, L. (2011). Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. *The Journal of Neuroscience: the Official Journal of the Society for Neuroscience, 31*(30), 10829–10835. http://doi.org/10.1523/JNEUROSCI.2246-11.2011.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience, 14*(10), 1338–1344. http://doi.org/10.1038/nn.2921.

Alexander, W. H., & Brown, J. W. (2014). A general role for medial prefrontal cortex in event prediction. *Frontiers in Computational Neuroscience, 8*(69). http://doi.org/10.3389/fncom.2014.00069.

Alexander, W. H., & Brown, J. W. (2015). Hierarchical error representation: A computational model of anterior cingulate and dorsolateral prefrontal cortex. *Neural Computation, 27*(11), 2354–2410. http://doi.org/10.1162/NECO_a_00779.

Alexander, G., & DeLong, M. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. *Trends in Cognitive Sciences, 12*(5), 193–200. http://doi.org/10.1016/j.tics.2008.02.004.

Badre, D., & Frank, M. J. (2011). Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: Evidence from fMRI. *Cerebral Cortex (New York, N.Y.: 1991)*, 1–10. http://doi.org/10.1093/cercor/bhr117.

Badre, D., Kayser, A. S., & Esposito, M. D. (2010). Article frontal Cortex and the Discovery of abstract action rules. *Neuron, 66*(2), 315–326. http://doi.org/10.1016/j.neuron.2010.03.025.

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). What's past is present: Reminders of past choices bias decisions for reward in humans. *bioRxiv*.

Bornstein, A. M., & Norman, K. A. (2017). Putting value in context: A role for context memory in decisions for reward. *bioRxiv*.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement-learning perspective. *Cognition, 113*(3), 262–280.

[**AU6**]

Bromberg-Martin, E. S., & Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron, 63*(1), 119–126. http://doi.org/10.1016/j.neuron.2009.06.009.

Cavanagh, J. F., Masters, S. E., Bath, K., & Frank, M. J. (2014). Conflict acts as an implicit cost in reinforcement learning. *Nature Communications, 5*(5394). http://doi.org/10.1038/ncomms6394.

Collins, A. G. E., Cavanagh, J. F., & Frank, M. J. (2014). Human EEG uncovers latent generalizable rule structure during learning. *The Journal of Neuroscience, 34*(13), 4677–4685. http://doi.org/10.1523/JNEUROSCI.3900-13.2014.

Collins, A. G. E., Ciullo, B., Frank, M. J., & Badre, D. (2017). Working memory load strengthens reward prediction errors. *The Journal of Neuroscience, 37*(16), 2700–2716. http://doi.org/10.1523/JNEUROSCI.2700-16.2017.

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *The European Journal of Neuroscience, 35*(7), 1024–1035. http://doi.org/10.1111/j.1460-9568.2011.07980.x.

Collins, A. G. E., & Frank, M. J. M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review, 120*(1), 190–229. http://doi.org/10.1037/a0030852.

Collins, A. G. E., & Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review, 121*(3), 337–366. http://doi.org/10.1037/a0037015.

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

[**AU7**]

Collins, A. G. E., & Frank, M. J. (2016b). Motor demands constrain cognitive rule structures. *PLoS Computational Biology, 12*(3), e1004785. http://doi.org/10.1371/journal.pcbi.1004785.

Collins, A. G. E., & Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *Plos Biology, 10*(3), e1001293. http://doi.org/10.1371/journal.pbio.1001293.

Davidow, J. Y., Foerde, K., Galvan, A., & Shohamy, D. (2016). An upside to reward Sensitivity: The Hippocampus supports enhanced reinforcement learning in adolescence. *Neuron, 92*(1), 93–99. http://doi.org/10.1016/j.neuron.2016.08.031.

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*(6), 1204–1215. http://doi.org/10.1016/j.neuron.2011.02.027.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness. In *Proceedings of the 15th International Academic MindTrek Conference on Envisioning future Media environments - MindTrek '11 (p. 9)*. New York, New York, USA: ACM Press. http://doi.org/10.1145/2181037.2181040.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M., & Niv, Y. (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33*(13), 5797–5805. http://doi.org/10.1523/JNEUROSCI.5445-12.2013.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, (February), 1–9. http://doi.org/10.1038/nn.3981.

Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science, 344*(6191), 1481–1486. http://doi.org/10.1126/science.1252254.

Dunovan, K., & Verstynen, T. (2016). Believer-skeptic meets actor-critic: Rethinking the role of basal ganglia pathways during decision-making and reinforcement learning. *Frontiers in Neuroscience, 10*(106). http://doi.org/10.3389/fnins.2016.00106.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. *Journal of Cognitive Neuroscience, 17*(1), 51–72. http://doi.org/10.1162/0898929052880093.

Frank, M. J., & Badre, D. (2011). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex (New York, N.Y.: 1991), 2010*, 1–18. http://doi.org/10.1093/cercor/bhr114.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America, 104*(41), 16311–16316. http://doi.org/10.1073/pnas.0706111104.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science (New York, N.Y.), 306*(5703), 1940–1943. http://doi.org/10.1126/science.1102941.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117*(1), 197–209. http://doi.org/10.1037/a0017808.

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. [**AU8**] *Current Opinion in Behavioral Sciences, 5*, 43–50. http://doi.org/10.1016/j.cobeha.2015.07.007.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification Work? – a literature review of empirical studies on gamification. In *2014 47th Hawaii International Conference on system Sciences* (pp. 3025–3034). IEEE. http://doi.org/10.1109/HICSS.2014.377.

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., … Berke, J. D. (2015). Mesolimbic dopamine signals the value of work. *Nature Neuroscience, 19*(1), 117–126. http://doi.org/10.1038/nn.4173.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *The Journal of Neuroscience: the Official Journal of the Society for Neuroscience, 26*(32), 8360–8367. http://doi.org/10.1523/JNEUROSCI.1010-06.2006.

Hart, A. S., Rutledge, R. B., Glimcher, P. W., & Phillips, P. E. M. (2014). Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. *Journal of Neuroscience, 34*(3), 698–704. http://doi.org/10.1523/JNEUROSCI.2489-13.2014.

Holroyd, C. B., & McClure, S. S. M. (2015). Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. *Psychological Review, 122*(1), 54–83. http://doi.org/10.1037/a0038339.

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences, 16*(2), 122–128. http://doi.org/10.1016/j.tics.2011.12.008.

Huys, Q. J. M., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., … Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences of the United States of America, 112*(10), 3098–3103. http://doi.org/10.1073/pnas.1414219112.

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks, 15*(4), 549–559. http://doi.org/10.1016/S0893-6080(02)00048-5.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science (New York, N.Y.),, 302*(5648), 1181–1185. http://doi.org/10.1126/science.1088545.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences, 11*(6), 229–235. http://doi.org/10.1016/j.tics.2007.04.005.

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General, 143*(1), 131–141. http://doi.org/10.1037/a0031048.

Kravitz, A. V., Tye, L. D., & Kreitzer, A. C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neuroscience, 15*(6), 816–818. http://doi.org/10.1038/nn.3100.

Leong, Y. C., Radulescu, A., Daniel, R., Dewoskin, V., Niv, Y., & Partners, T. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron, 93*(2), 451–463. http://doi.org/10.1016/j.neuron.2016.12.040.

Lieder, F., Griffiths, T.L. (n.d.). Helping people make better decisions using optimal gamification.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience: the Official Journal of the Society for Neuroscience, 16*(5), 1936–1947.

Nee, D. E., & D'Esposito, M. (2016). The hierarchical organization of the lateral prefrontal cortex. *eLife, 5*(March 2016), 1–26. http://doi.org/10.7554/eLife.12112.

Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 35*(21), 8145–8157. http://doi.org/10.1523/JNEUROSCI.2978-14.2015.

Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron, 71*(2), 370–379. http://doi.org/10.1016/j.neuron.2011.05.042.

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*(4), 486–492. http://doi.org/10.1038/nn.3331.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2016). Complementary learning systems within the hippocampus: A neural network modeling approach to reconciling episodic memory with statistical learning. *bioRxiv, 51870*. http://doi.org/10.1101/051870.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (2014). Optimal behavioral hierarchy. *PLoS Computational Biology, 10*(8). http://doi.org/10.1371/journal.pcbi.1003779.

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review, 121*(3), 526–558. http://doi.org/10.1037/a0037018.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning* (Vol. 9). MIT Press.

Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A., & Wilbrecht, L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neuroscience, 15*(9), 1281–1289. http://doi.org/10.1038/nn.3188.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience, 15*(2), 395–415. http://doi.org/10.3758/s13415-015-0334-y.

Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in Human Neuroscience, 5*(January), 1–14. http://doi.org/10.3389/fnhum.2011.00189.

Non-Print Items

Abstract
How the brain uses reinforcement feedback to make simple choices that lead to reward is well understood. However, this ability is often considered insufficient to account for the flexibility and efficiency of human decision-making. In this chapter, we show that the computations of model-free reinforcement learning (RL) can in fact account for complex human learning abilities, such as generalization, transfer, and fast learning in high-dimensional, dynamic environments. Specifically, we show that humans structure their current information and choices into useful state and action spaces and that applying simple RL computations to these spaces—sometimes hierarchically—enables rich decision-making. Thus, RL computations enable humans to learn to represent the information they acquire in structured ways. Such structured RL simplifies complex problems (through representation learning), affords transfer of information (by building abstract rules and relating them to relevant contexts), and enables efficient exploration (by grouping together subsequences or identifying subpolicies).

Keywords:
Exploration; Generalization; Hierarchical reinforcement learning; Reinforcement learning; Representation learning; Rule learning; Structure learning.