Neural Networks 145 (2022) 271–287

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2021 Special Issue on AI and Brain Science

Reinforcement learning and its connections with neuroscience and psychology

Ajay Subramanian [a,c,*], Sharad Chitlangia [b,c], Veeky Baths [c,d,**]

[a] *Department of Psychology, New York University, New York, New York, 10003, USA*
[b] *Amazon*
[c] *Cognitive Neuroscience Lab, BITS Pilani K K Birla Goa Campus, NH-17B, Zuarinagar, Goa, 403726, India*
[d] *Department of Biological Sciences, BITS Pilani K K Birla Goa Campus, NH-17B, Zuarinagar, Goa, 403726, India*

ARTICLE INFO

Article history:
Available online 22 October 2021

Keywords:
Reinforcement learning
Neuroscience
Psychology

ABSTRACT

Reinforcement learning methods have recently been very successful at performing complex sequential tasks like playing Atari games, Go and Poker. These algorithms have outperformed humans in several tasks by learning from scratch, using only scalar rewards obtained through interaction with their environment. While there certainly has been considerable independent innovation to produce such results, many core ideas in reinforcement learning are inspired by phenomena in animal learning, psychology and neuroscience. In this paper, we comprehensively review a large number of findings in both neuroscience and psychology that evidence reinforcement learning as a promising candidate for modeling learning and decision making in the brain. In doing so, we construct a mapping between various classes of modern RL algorithms and specific findings in both neurophysiological and behavioral literature. We then discuss the implications of this observed relationship between RL, neuroscience and psychology and its role in advancing research in both AI and brain science.

1. Introduction

Reinforcement learning (RL) methods have been very successful on a variety of complex sequential tasks such as Atari (Mnih et al., 2015), Go (Silver et al., 2016), Poker (Heinrich & Silver, 2015) and Dota-2 (Berner et al., 2019), often far exceeding human-level performance. Though a large portion of these successes can be attributed to recent developments in deep reinforcement learning, many of the core ideas employed in these algorithms derive inspiration from findings in animal learning, psychology and neuroscience. There have been multiple works reviewing the correlates of reinforcement learning in neuroscience (Botvinick, Wang, Dabney, Miller, & Kurth-Nelson, 2020; Lee, Seo, & Jung, 2012; Niv, 2009). In 2012, Lee et al. (2012) reviewed several works reporting evidence of classical reinforcement learning ideas being implemented within the neural networks of the brain. Many commonly used building blocks of RL such as value functions, temporal difference learning and reward prediction errors (RPEs) have been validated by findings from neuroscience research, thus making reinforcement learning a promising candidate for computationally modeling human learning and decision making.

Since 2012 however, unprecedented advancement in RL research, accelerated by the arrival of deep learning has resulted in the emergence of several new ideas apart from the classical ideas for which neuroscience analogues had earlier been found. Relatively newer research areas like distributional RL (Bellemare, Dabney, & Munos, 2017), meta RL (Duan, Schulman, Chen, Bartlett, Sutskever, & Abbeel, 2016; Wang et al., 2018), and model-based RL (Sutton, 1990) have emerged, which has motivated work that seeks and in some cases finds, evidence for similar phenomena in neuroscience and psychology. In this review, we have incorporated these works, thus providing a well rounded and up-to-date review of the neural and behavioral correlates for modern reinforcement learning algorithms.

For this review, we employ the following structure. In Section 2, we provide a brief overview of classical reinforcement learning, its core, and the most popular ideas, in order to enable the uninformed reader to appreciate the findings and results discussed later on. Then, in Section 3 we discuss some of the building blocks of classical and modern RL: value functions, reward prediction error, eligibility traces and experience replay. While doing so, we discuss phenomena from neuroscience and

[*] Corresponding author at: Department of Psychology, New York University, New York, New York, 10003, USA.
[**] Corresponding author at: Cognitive Neuroscience Lab, BITS Pilani K K Birla Goa Campus, NH-17B, Zuarinagar, Goa, 403726, India.
E-mail addresses: as15003@nyu.edu (A. Subramanian), chitshar@amazon.com (S. Chitlangia), veeky@goa.bits-pilani.ac.in (V. Baths).

https://doi.org/10.1016/j.neunet.2021.10.003
0893-6080/© 2021 Elsevier Ltd. All rights reserved.

psychology that are analogous to these concepts and evidence that they are implemented in the brain. Following this, in Section 4 we discuss some modern RL algorithms and their neural and behavioral correlates: temporal difference learning, model-based RL, distributional RL, meta RL, causal RL and Hierarchical RL. Having explored all of these topics in considerable depth, we provide a mapping between specific reinforcement learning concepts and corresponding work validating their involvement in animal learning (Table 1). Finally in Section 5, we present a discussion on how research at the intersection of these fields can propel each of them forward. To do so, we discuss specific challenges in RL that brain science might hold key insight to, and vice versa.

The following two organizational choices govern our presentation of this review.

- **Two-way discussion**: We present an exhaustive review that simultaneously discusses important recent research, on how both neuroscience and psychology have influenced RL. Additionally, unlike previous work (Botvinick, Ritter, Wang, Kurth-Nelson, Blundell, & Hassabis, 2019; Botvinick et al., 2020; Lee et al., 2012), we also discuss how ideas from RL have recently impacted research in brain science.
- **Modularity**: From a computational perspective, decision making can be broken down conceptually into several capabilities. For example: planning, hierarchy, valuing choices, learning to learn etc. These happen to be distinct sub-fields of reinforcement learning (model-based RL, hierarchical RL, meta RL etc.). Therefore we divide our discussion into sections, each devoted to a specific module. This modular structure makes it easier to simultaneously describe, understand and compare ideas from RL, neuroscience and psychology.

2. Reinforcement learning: Background

The classical reinforcement learning framework describes an agent (human, robot etc.) interacting with its environment and learning to behave in a way that maximizes reward (Sutton & Barto, 1998). Fig. 1 illustrates this interaction. The agent is given a **state** S_t by the environment at a time t. The agent, using an internal **policy** $\pi(S_t)$ or strategy selects an **action** A_t. The action when applied to the environment moves the agent to a new state S_{t+1} and returns to it a scalar **reward** R_{t+1}. This sequence makes up a single **transition**. An agent's interaction with its environment comprises several such transitions. While considering these transitions, an explicit assumption is made that the future is independent of the past given the present. In other words, the next state is dependent only on the current state, action and environment properties, not on any states or actions previously taken. This is known as a Markov assumption and the process is therefore a **Markov Decision Process** (MDP).

In a reinforcement learning problem, the objective of the agent is to maximize the reward obtained over several transitions. In other words, the aim of the agent is to find a policy which when used to select actions, returns an optimal reward over a long duration. The most popular version of this maximization objective is to maximize discounted reward.

$$\pi_{\text{optimal}}(S_t) = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} R(S_\tau), \pi(S_\tau)\right] \tag{1}$$

where $\pi_{optimal}$ is the optimal policy and $\gamma \mid 0 < \gamma < 1$ is a factor to discount future rewards. It should be noted that maximizing the reward for each transition independently might not yield an optimal long-term reward. This aspect introduces several complexities in arriving at an optimal solution for the RL problem, and motivates features such as exploration and planning in RL algorithms.

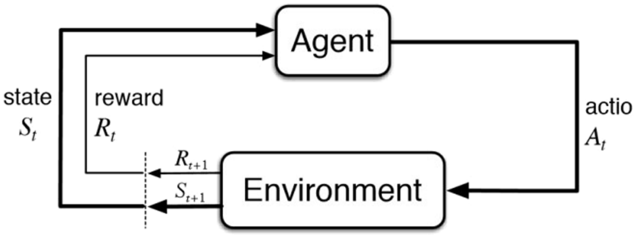

Fig. 1. The classical RL framework. The agent selects an action A_t at state S_t, in response to which it receives a corresponding reward R_{T+1}. The objective of the agent is to choose actions that maximize its reward over a long sequence of transitions.
Source: Sutton and Barto (1998).

Many reinforcement learning algorithms use a value function as a way to assign utility to states and actions. The value function of a state S_t is the expected reward that the agent is expected to receive if it starts at that state and executes a particular policy forever after.

$$V(S_t) = \mathbb{E}\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} R(S_\tau, \pi(S_\tau))\right] \tag{2}$$

where V is the state value function. A value function could also be assigned to a state–action pair in which case it represents the expected reward if a specific policy is executed *after* a given action is taken at the state.

$$Q_\pi(S_t, A_t) = \mathbb{E}\left[R_{t+1} + \sum_{\tau=t+1}^{\infty} \gamma^{\tau-t} R(S_\tau, \pi(S_\tau))\right] \tag{3}$$

where Q represents the value function for a state–action pair.

Value-function based RL algorithms often optimize value function estimates rather than directly optimizing policy. Once the optimal value function is learned, an optimal policy would then entail picking the highest value actions at each state. This procedure is called value iteration (Pashenkova, Rish, & Dechter, 1996) and finds application in various modern reinforcement learning algorithms. A common set of algorithms for optimizing the value function are the **dynamic programming (DP) methods**. These methods update value functions by bootstrapping value functions from other states (Bellman, 1954; Busoniu, Babuska, De Schutter, & Ernst, 2017). Examples of DP methods are Q-learning (Watkins & Dayan, 1992) and SARSA (Sutton & Barto, 1998). The optimization process involves updating the value function by ascending the gradient in the direction of the difference between **target** values and the currently estimated values, thus moving towards better estimates of rewards obtained during environment interaction. The target value is computed using DP bootstrapping. The difference between target and current value is termed as Reward Prediction Error (RPE). Dynamic programming methods that use value functions of states adjacent to the current state, to compute the target, are called temporal difference methods (Sutton, 1988) and are discussed in Section 4.1.

Now that we have given a brief background on some of the important core reinforcement learning concepts: MDP, policy, value functions and dynamic programming; we will explore the neural and behavioral correlates for some of the fundamental building blocks that make up classical and modern reinforcement learning algorithms.

Table 1
Table summarizing the mapping discussed between concepts in reinforcement learning (left column) and evidential phenomena reported in neuroscience and psychology research or specific areas of the brain responsible from them. Literature corresponding to them have also been referenced.

Reinforcement learning concept	Corresponding phenomena in neuroscience and psychology/Brain area responsible
state value function	reward expectancy in basal ganglia (Hikosaka, Nakamura, & Nakahara, 2006), prefrontal cortex (Wallis & Kennerley, 2010) and other areas (Schultz & Dickinson, 2000)
action value function	chosen value during decision making (Cai, Kim, & Lee, 2011; Lau & Glimcher, 2008; Padoa-Schioppa & Assad, 2006; Samejima, Ueda, Doya, & Kimura, 2005)
multi-task learning (Lazaric & Ghavamzadeh, 2010)	distributed reward signals (Cai et al., 2011; Kim, Hwang, & Lee, 2008; Murdoch, Chen, & Goldberg, 2018; Padoa-Schioppa & Assad, 2006; Pastor-Bernier & Cisek, 2011; Seo, Barraclough, & Lee, 2009; So & Stuphorn, 2010)
actor–critic baseline (Sutton, McAllester, Singh, & Mansour, 1999a)	relative values (Cai et al., 2011; Kim et al., 2008; Pastor-Bernier & Cisek, 2011; Seo et al., 2009; Seo & Lee, 2009)
reward prediction error/TD error	RPE signals in dopaminergic neurons (VTA) (Schultz, 2006), orbitofrontal cortex, lateral habenula, cingulate cortex, etc. (Hong & Hikosaka, 2008; Kim, Sul, Huh, Lee, & Jung, 2009; Matsumoto & Hikosaka, 2007; Matsumoto, Matsumoto, Abe, & Tanaka, 2007; Oyama, Hernádi, Iijima, & Tsutsui, 2010; Seo & Lee, 2007; Sul, Kim, Huh, Lee, & Jung, 2010)
credit assignment problem	orbitofrontal cortex (Fellows & Farah, 2003; Iversen & Mishkin, 1970; Murray, O'Doherty, & Schoenbaum, 2007; Schoenbaum, Nugent, Saddoris, & Setlow, 2002)
eligibility traces (Sutton & Barto, 1998)	prefrontal cortex, striatum, frontal cortex (Barraclough, Conroy, & Lee, 2004; Curtis & Lee, 2010; Kim, Huh, Lee, Baeg, Lee, & Jung, 2007; Kim et al., 2009; Seo et al., 2009; Seo & Lee, 2009; Sul, Jo, Lee, & Jung, 2011; Sul et al., 2010)
experience replay (Mnih et al., 2015; Schaul, Quan, Antonoglou, & Silver, 2016)	hippocampal place cells (Diba & Buzsáki, 2007; Foster & Wilson, 2006; Louie & Wilson, 2001; Moser, Kropff, & Moser, 2008; Skaggs & McNaughton, 1996), entorhinal cortices (Ólafsdóttir, Carpenter, & Barry, 2016), visual cortices (Ji & Wilson, 2007)
episodic memory (Lin, Zhao, Yang, & Zhang, 2018)	instance-based models of memory (Bornstein, Khaw, Shohamy, & Daw, 2017; Bornstein & Norman, 2017; Gershman & Daw, 2017; Lengyel & Dayan, 2007)
temporal difference (TD) learning (Samuel, 1959; Sutton, 1988)	reward prediction error hypothesis of dopamine neuron activity (Schultz, Dayan, & Montague, 1997)
TD(λ) (Sutton & Barto, 1998)	TD model of classical conditioning (Montague, Dayan, & Sejnowski, 1996)
model-based RL	Cognitive maps (Tolman, 1948); role of PFC (Gläscher, Daw, Dayan, & O'Doherty, 2010a), hippocampus (Benchenane et al., 2010; Hyman, Hasselmo, & Seamans, 2011; Sirota, Montgomery, Fujisawa, Isomura, Zugaro, & Buzsáki, 2008; Womelsdorf, Vinck, Stan Leung, & Everling, 2010)
successor representations (Dayan, 1993)	neural substrates (Gershman, 2018), SR as between model-free and model-based systems (Akam & Walton, 2021; Russek, Momennejad, Botvinick, Gershman, & Daw, 2017)
distributional TD learning (Bellemare et al., 2017)	distributional coding in non-RL domains (Dabney, Rowland, Bellemare, & Munos, 2018; Lammel, Lim, & Malenka, 2014; Pouget, Beck, Ma, & Latham, 2013), value coding in VTA of mice (Dabney et al., 2020)
meta reinforcement learning (Duan et al., 2016; Schaul & Schmidhuber, 2010; Wang et al., 2016)	learning to learn (Harlow, 1949), fast and slow learning (Botvinick et al., 2019), prefrontal cortex (Barraclough et al., 2004; Kim & Shadlen, 1999; Padoa-Schioppa & Assad, 2006; Rushworth & Behrens, 2008; Tsutsui, Grabenhorst, Kobayashi, & Schultz, 2016; Wang et al., 2018)
episodic meta RL (Ritter, Wang, Kurth-Nelson, & Botvinick, 2018a)	cerebral cortex (Ritter et al., 2018b; Santoro, Bartunov, Botvinick, Wierstra, & Lillicrap, 2016; Wayne et al., 2018), interaction between meta model-based control and episodic memory in human learning (Vikbladh, Shohamy, & Daw, 2017)
causality (Pearl, 2009)	children perform interventions (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; Sobel & Sommerville, 2010); counterfactuals (Palminteri, Lefebvre, Kilford, & Blakemore, 2017) in mPFC (Pischedda, Palminteri, & Coricelli, 2020), frontal cortex (Boorman, Behrens, & Rushworth, 2011), dopamine fluctuations (Kishida et al., 2016); causal induction (Gershman & Niv, 2010, 2012; Soto, Gershman, & Niv, 2014)
hierarchy (Barto & Mahadevan, 2003; Dietterich, 2000; Parr & Russell, 1998; Sutton, Precup, & Singh, 1999b)	goal-directed behavior (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004; Botvinick & Plaut, 2004; Lashley, 1951; Miller, Eugene, & Pribram, 2017; Newell, Shaw, & Simon, 1959; Schneider & Logan, 2006; Zacks, Speer, Swallow, Braver, & Reynolds, 2007), prefrontal cortex (Badre, 2008; Balaguer, Spiers, Hassabis, & Summerfield, 2016; Botvinick, 2008; Courtney, Roth, & Sala, 2007; Fuster, 1989; Koechlin, Ody, & Kouneiher, 2003; Ribas-Fernandes, Shahnazian, Holroyd, & Botvinick, 2019; Ribas-Fernandes et al., 2011), higher level representation of task context (Lashley, 1951), mental hierarchical organization (Newtson et al., 1976; Zacks & Tversky, 2001), gradual integration of skills (Bruner, 1973; Fischer, 1980; Greenfield, Nelson, & Saltzman, 1972), production-system based theories of cognition (ACT-R, Soar) (Anderson et al., 2004)
options (Sutton et al., 1999b)	task representation (Cohen, Dunbar, & McClelland, 1990; Cooper & Shallice, 2000; Monsell, 2003)
incompatibility between learning problem and temporally abstract actions (Botvinick, Niv, & Barto, 2009)	negative transfer problem (Luchins, 1942)
option discovery	integration of causal representations (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Gopnik & Schulz, 2004; Sommerville & Woodward, 2005a, 2005b), intrinsic rewards (Bunzeck & Düzel, 2006; Redgrave & Gurney, 2006; Schultz, Apicella, & Ljungberg, 1993), Bayesian models (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013; Solway et al., 2014; Tomov, Yagati, Kumar, Yang, & Gershman, 2020)

(*continued on next page*)

Table 1 (*continued*).

Reinforcement learning concept	Corresponding phenomena in neuroscience and psychology/Brain area responsible
graph partitioning to identify bottleneck states (Şimşek, Wolfe, & Barto, 2005; Mannor, Menache, Hoze, & Klein, 2004; Menache, Mannor, & Shimkin, 2002)	children integrate causal representations into a causal model (Gopnik et al., 2004; Gopnik & Schulz, 2004; Sommerville & Woodward, 2005a, 2005b)
intrinsic motivation (Barto, Singh, & Chentanez, 2004; Singh, Barto, & Chentanez, 2004)	intrinsic rewards in dopamine driven learning (Bunzeck & Düzel, 2006; Redgrave & Gurney, 2006; Schultz et al., 1993)

3. Building blocks of reinforcement learning: Neural and behavioral correlates

As briefly outlined in the previous section, the RL solution to an MDP can make use of several components or building-blocks. Among these, some of the most popular components are the value function, eligibility traces and reward prediction errors. More recently with the advent of deep RL, new components such as experience replay and episodic memory have emerged that are commonly incorporated within RL algorithms. In this section, we explore their neural and behavioral correlates. Many of these ideas have already been reviewed in much detail by Lee et al. (2012) and so our description of them will be concise relative to topics covered in future sections of the paper. For a more in-depth review of these, see Lee et al. (2012).

3.1. Value functions

As discussed in the previous section, a value function is a measure of reward expectation. This expectation is measured practically as the mean of discounted rewards over future states.

$$V(S_t) = \frac{1}{N} \sum_{\tau=t}^{t+N-1} \gamma^{\tau-t} R(S_\tau, \pi(S_\tau)) \tag{4}$$

where N is the number of sample states considered.

Neural signals containing information about reward expectancy have been found in many areas of the brain (Hikosaka et al., 2006; Schultz & Dickinson, 2000; Wallis & Kennerley, 2010). Resembling the two types of value functions prevalent in RL algorithms, evidence has been found for the brain too encoding both state values and action value. Action value functions (Samejima et al., 2005) are useful during motor responses when an action needs to be selected while state value functions might play an evaluative role. Transformations between the two types have been found to occur in the brain. For instance, during decision making, state value functions transform from a mean over all actions into value functions for the chosen action, which are often referred to as **chosen values** (Cai et al., 2011; Lau & Glimcher, 2008; Padoa-Schioppa & Assad, 2006).

Despite these similarities between neural value signals and value functions employed in reinforcement learning, they are different in some important ways. In RL, value functions for different decisions are all treated the same and represent the expected value of a single reward. But in the brain, activity for action value functions are observed in various areas for a single decision (Cai et al., 2011; Kim et al., 2008; Pastor-Bernier & Cisek, 2011; Seo et al., 2009; So & Stuphorn, 2010) and might apply to distinct reward signals as demonstrated by Murdoch et al. (2018) in the case of place preference (navigation) and song syllables (singing), in songbirds. They observed that strobe light flashes aversively conditioned place preference but not vocal learning, while noise bursts aversively conditioned vocal learning but not place preference. Another case of value signals in different brain regions encoding expectancy for different rewards for the same decision was seen in animals during a juice flavor decision making task (Padoa-Schioppa & Assad, 2006). Neurons encoding value in the supplemental motor area signaled desirable eye movements (to spatial location of targets) when considering the choice while those in the primate orbitofrontal cortex were associated with the juice flavors themselves (targets) (Tremblay & Schultz, 1999; Wallis & Miller, 2003). The function of distinct reward signals in the brain is analogous to a highly distributed version of backpropagation in artificial neural networks wherein distinct reward signals could selectively update distinct sets of parameters in the network.

Neural signals for chosen values are also distributed in multiple brain areas (Cai et al., 2011; Kim et al., 2009; Lau & Glimcher, 2008; Padoa-Schioppa & Assad, 2006; Sul et al., 2010). Also, some brain regions encode the difference of values between two alternative actions to determine likelihood of taking an action over another (Cai et al., 2011; Kim et al., 2008; Pastor-Bernier & Cisek, 2011; Seo et al., 2009; Seo & Lee, 2009). This is similar to the use of baseline values in **actor–critic** methods where value functions of each action are scaled by their mean to obtain values relative to other actions at the state (Sutton et al., 1999a). These baseline values are used to compute the 'advantage' of a given action relative to others available in the action space.

3.2. Reward prediction error

In order to make good decisions about states that are desirable to visit, we need a good estimate of the value of a state. Most reinforcement learning algorithms optimize the value function by minimizing a **reward prediction error** (RPE). If $V_\pi(S_t)$ is the value function (expected reward) at a state S_t and $G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots$ is the sum of rewards obtained after time t (also known as **return**), then a common formulation of the RPE at t is the difference between the two terms.

$$RPE(t) = G_t - V_\pi(S_t) \tag{5}$$

A policy that minimizes the reward prediction error for all states would return the best value estimates for all states in the state space. Thus, the RPE helps improve value estimates. Such RPE signals have been identified in midbrain dopaminergic neurons (Schultz, 2006) and many other areas such as the orbitofrontal cortex, lateral habenula and angular cingulate cortex (Hong & Hikosaka, 2008; Kim et al., 2009; Matsumoto & Hikosaka, 2007; Matsumoto et al., 2007; Oyama et al., 2010; Seo & Lee, 2007; Sul et al., 2010).

Action value functions in the brain are believed to be updated and stored at the synapses between cortical axons and striatal spiny dendrites (Hikosaka et al., 2006; Hong & Hikosaka, 2011; Lo & Wang, 2006; Reynolds, Hyland, & Wickens, 2001). RPEs are input to these synapses through terminals of dopaminergic neurons (Haber, Fudge, & McFarland, 2000; Haber & Knutson, 2010; Levey et al., 1993; Schultz, 2006) and value functions are updated.

3.3. Credit assignment and eligibility traces

In many cases and quite commonly in human behavior, rewards for a task are temporally delayed. In other words, decisions have to be made at several states before a reward feedback is obtained. An example of this is cooking, where a complete recipe has to be executed before actually tasting the dish and determining whether it is good or bad (reward). Now, say the reward was negative, we need to be able to effectively determine the step of the recipe where we went wrong so that it can be corrected on the next trial. This is known as the **credit assignment problem**. It is the challenge of finding the "responsibility" of each encountered state for an obtained reward. Behavioral experiments on reversal learning tasks have shown that credit assignment problems surface in animals with lesions in the orbitofrontal cortex which therefore suggests its involvement in assigning credit to states (Fellows & Farah, 2003; Iversen & Mishkin, 1970; Murray et al., 2007; Schoenbaum et al., 2002).

In RL literature, there are two prominent techniques that have been developed to solve the credit assignment problem. The first approach is to introduce intermediate states (Montague et al., 1996) to strengthen the connection between a state and its corresponding reward. This technique, however, does not correspond to any observation in neuroscience literature and is not consistent with profiles of dopamine neuron activity (Pan, Schmidt, Wickens, & Hyland, 2005). The second method is to use eligibility traces which are short term memory signals that assign state responsibility for a reward (Sutton & Barto, 1998). Eligibility traces are higher for states that are on average closer to the reward. Unlike intermediate states, eligibility traces have been observed in several animals and brain regions including the prefrontal cortex, striatum and frontal cortex (Barraclough et al., 2004; Curtis & Lee, 2010; Kim et al., 2007, 2009; Seo et al., 2009; Seo & Lee, 2009; Sul et al., 2011, 2010). The orbitofrontal cortex of the brain is believed to play an important role in credit assignment. Its involvement is evidenced by the observation that neurons in the orbitofrontal cortex show increased activity when a positive reward is obtained from a specific action (Abe & Lee, 2011; Barraclough et al., 2004; Kim et al., 2009; Roesch, Singh, Brown, Mullins, & Schoenbaum, 2009; Seo & Lee, 2009; Sul et al., 2010). Additionally, neurons in orbitofrontal cortex are believed to also encode relationships between actions and their corresponding outcomes (Barraclough et al., 2004; Kim et al., 2009; Seo et al., 2009; Sul et al., 2010). This observation could inspire future work in reinforcement learning research towards a solution to the credit assignment problem. Ongoing work is also exploring and testing the involvement of eligibility traces for credit assignment in synaptic value updates. A popular idea in this direction is neuromodulated STDP which attempts to model this by combining classical STDP with eligibility traces to add external reinforcement from reward signals (Gerfen & Surmeier, 2011; Shen, Flajolet, Greengard, & Surmeier, 2008).

3.4. Experience replay and episodic memory

Rodent research has led to the discovery that place cells and grid cells in the hippocampus encode a spatial map of the environment (Moser et al., 2008; O'Keefe & Dostrovsky, 1971; Sargolini et al., 2006). Along with mapping trajectories, these cells spontaneously recap previously experienced trajectories (Diba & Buzsáki, 2007; Foster & Wilson, 2006; Louie & Wilson, 2001; Skaggs & McNaughton, 1996). They can also explore new spatial trajectories which have not been experienced before (Gupta, van der Meer, Touretzky, & Redish, 2010; Ólafsdóttir, Barry, Saleem, Hassabis, & Spiers), a phenomenon which is known as **replay**. Replay's involvement in playing out trajectories that never happened suggest that it might be important in the brain's learning of world models (Foster, 2017; Pezzulo, van der Meer, Lansink, & Pennartz, 2014) which is used to generalize learned knowledge. Biological replay mechanisms have been recorded in the entorhinal cortices (Ólafsdóttir et al., 2016), prefrontal cortex (Peyrache, Khamassi, Benchenane, Wiener, & Battaglia, 2009) and visual cortices (Ji & Wilson, 2007). However, the information stored varies between these areas. Entorhinal cortical replay encodes spatial relationship between objects while visual cortical replay encodes sensory properties of events and objects. Replay mechanisms have been incorporated in modern deep reinforcement learning methods in the form of experience replay (Mnih et al., 2015; Schaul et al., 2016).

As discussed previously, deep RL methods that excel in performance in various tasks struggle with achieving sample efficiency similar to that of humans (Lake, Ullman, Tenenbaum, & Gershman, 2017; Tsividis, Pouncy, Xu, Tenenbaum, & Gershman, 2017). Experience replay is a popular component in reinforcement learning algorithms which enables them to learn tasks with fewer environment interactions by storing and reusing previously experienced transitions to update the policy. However, experience replay mechanisms in deep RL are still unable to mimic their biological counterparts. For instance, Liu, Dolan, Kurth-Nelson, and Behrens (2019) show that while experience replay in RL records experience in the same sequence as they occurred, hippocampal replay does not tend to follow this 'movie' sequence and rather employs an 'imagination' sequence in which experienced events are replayed in the order in which they are expected to occur according to learned internal models. Thus, integration of experience replay with model-based RL is an exciting avenue for future deep RL research.

Additionally, experience replay in deep RL involves using only previously played trajectories of the same task that the agent is learning and hence do not assist in learning new tasks. Another approach called episodic RL (Gershman, 2017; Lengyel & Dayan, 2007; Pritzel et al., 2017) uses such experience as an inductive bias to learn future tasks. An illustration of an episodic RL algorithm is shown in Fig. 2. These approaches employ a similarity network which reuses values for states that have already been learned, thus reducing the time to learn values for new states (Lin et al., 2018). This idea is similar to instance-based models of memory in where specific stored information from past experience is used for decision making in new situations (Bornstein et al., 2017; Bornstein & Norman, 2017; Gershman & Daw, 2017; Lengyel & Dayan, 2007).

4. Algorithms for reinforcement learning: Neural and behavioral correlates

Having covered the building blocks that most commonly make up RL algorithms, we now dive deep into various types of reinforcement learning algorithms along with work in both neuroscience and psychology suggesting that they might be promising models for certain aspects of animal learning and decision making.

4.1. Temporal difference learning

Temporal difference learning is one of the central ideas in reinforcement learning. The most common formulation of the reward prediction error discussed in the previous section, is the temporal difference (TD) error. The TD error δ_t is defined as:

$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(S_t) \tag{6}$$

The origins of TD learning can be traced back to the Rescorla–Wagner (RW) model for classical conditioning (Rescorla & Wagner, 1972), that learning occurs only when the animal is surprised. It proposed a trial-level associative learning rule which updates "associative strengths" of stimuli using a prediction error.

$$\delta_{RW} = \alpha(R_n - V(S_n)) \tag{7}$$

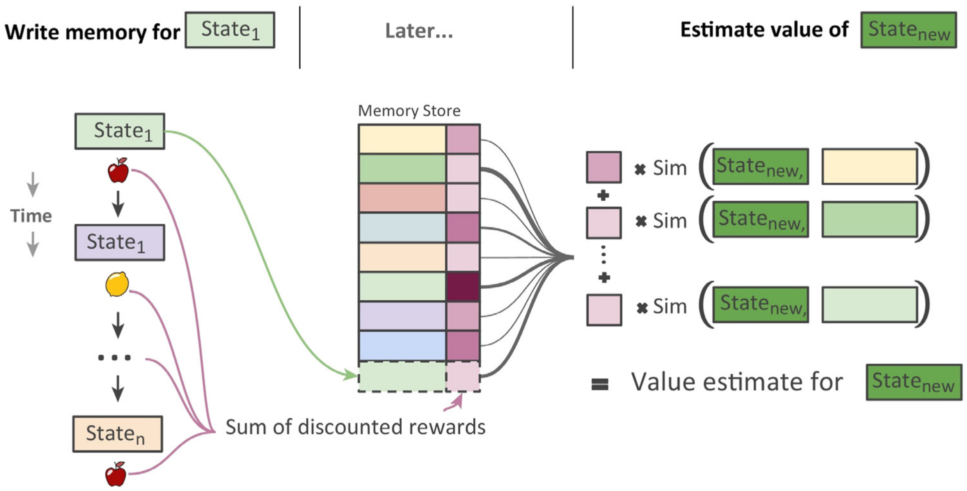

Fig. 2. Episodic reinforcement learning. Initially, the agent plays out a sequence of actions and stores the encountered states along with their learned value functions (expected sum of discounted rewards). When a new state is encountered in a new trajectory, its value can be estimated as linear combination of stored value functions weighted by the similarity of the stored states to the new state. This helps avoid learning value functions from scratch for every new state.
Source: Botvinick et al. (2019).

where α is the learning rate, R_n is the unconditional stimulus (US, reward) and $V(S_n)$ is the associative strength for conditional stimulus S which measures how well it predicts the US. While this model explains aspects of classical conditioning such as blocking (Kamin, 1967), it is trial-based and so does not explain temporal dependencies of learning and consequently, higher-order conditioning. The TD model fills these gaps.

The earliest known use of temporal difference learning in artificial intelligence research dates back to 1959 when Samuel (1959) demonstrated its usage for a checkers-playing program. Sutton (1988) developed the first theoretical formulation of TD learning and showed that it was more efficient and more accurate than conventional supervised learning methods.

Following this, work in computational neuroscience suggested that the firing of dopamine neurons signaled a reward prediction error (Montague, Dayan, Nowlan, Pouget, & Sejnowski, 1993). Later work Sejnowski, Dayan, and Montague (1995) showed that the TD model allows a formulation of expectations through value functions to influence synaptic changes, via a Hebbian learning framework. Later, from the extensive experiments conducted by Schultz et al. (1997), a major breakthrough in relating TD methods to actual biological phenomena was made by Montague et al. (1996) when they related fluctuation levels in dopamine delivery, from the VTA/SNc to cortical and subcortical target neuronal structures, to TD reward prediction errors.

The TD error formulation in RL is a very specific case of the more general TD(λ) proposed by Sutton and Barto (1998) which accounts for eligibility traces. The concept of eligibility traces was inspired from ideas like trace conditioning (Pavlov & Anrep, 1927) and Hull's learning theory (Hull, 1943). Two new terms are introduced to account for it, the weight vector w_t and the eligibility trace z_t, modeled as:

$$\delta_t = R_{t+1} + \gamma V(S_t, w_t) - V(S_t, w_t) \tag{8}$$

$$w_t = w_{t-1} + \alpha \delta_{t-1} z_{t-1} \tag{9}$$

$$z_t = \gamma \lambda z_{t-1} + \gamma V(S_{t-1}, w_{t-1}) \tag{10}$$

When $\lambda = 1$, this formulation perfectly mimics the behavior of **Monte Carlo algorithms** and the credit given to previous steps decreases by a factor of γ. On the other hand, when $\lambda = 0$, it transforms into the TD formulation discussed earlier, where only the previous state is given credit. Sutton and Barto (1998) showed that this TD(λ) formulation is the same formulation as TD model of classical conditioning as used in the framework proposed by Montague et al. (1996) to verify the results of TD learning. Thus, the TD(λ) formulation combined RPEs and eligibility traces into a single framework.

The TD model accounts for many limitations of the RW model. Bootstrapping of value functions (using a value estimate in the target expression) allows it to explain higher-order conditioning. It also provides flexibility in terms of real-time stimulus representation; with existing work having shown that one in particular, the microstimulus representation corresponds well with several empirical phenomena (Ludvig, Sutton, & Kehoe, 2012). Additionally, given its temporal nature, it has been able to make predictions about aspects of animal learning some of which were confirmed later on. For instance, some of its prominent predictions (Sutton & Barto, 1981, 1990) were: (a) conditioning requires a positive inter-stimulus interval (ISI), (b) remote associations facilitate conditioning, (c) blocking reverses if the new CS temporally precedes pretrained CS (later verified experimentally in rabbits (Kehoe, Schreurs, & Graham, 1987)). The TD model has also been successful in modeling observations from neuroscience. Perhaps the most commonly related neuroscience phenomenon to the TD reward prediction error was given by Schultz et al. (1997) as the reward prediction error hypothesis of dopamine neuron activity.

The TD model is, however, limited in a few ways. Firstly, it says nothing about how responses are learned, given the learned ability to predict stimuli. Secondly, the model in its simplest form cannot express uncertainty because it uses the point-average as an estimate for future reward. Recent approaches which model the value function as a distribution, account for uncertainty (Dabney et al., 2020; Gershman, 2015). These are discussed further in Section 4.3.

4.2. Model-based reinforcement learning

The classical reinforcement learning framework accounts only for learning that occurs through interaction. However, a large portion of learning in humans and animals involves imagined scenarios, planning out consequences of actions, replay (as discussed in Section 3.4) and so on. Model-based reinforcement learning seeks to mimic these capabilities and is a promising area both in RL (Silver et al., 2016) (Fig. 3) and as a computational model for biological learning (Doll, Simon, & Daw, 2012).

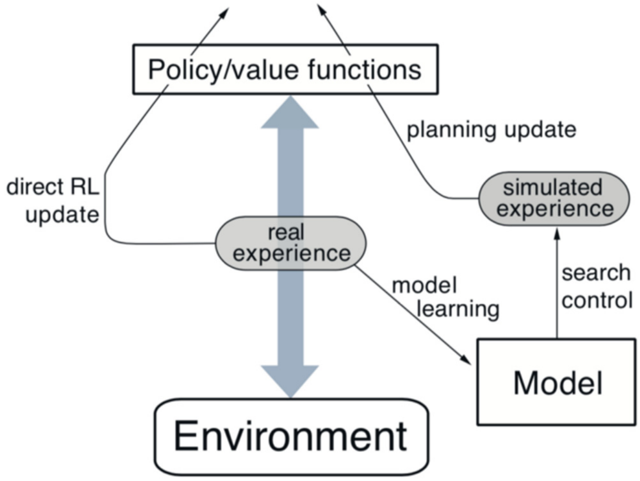

Fig. 3. Schematic of the Dyna model-based reinforcement learning framework (Sutton, 1990). Involves simultaneous model-free and model-based procedures. The model-free procedure involves using interactions with the real environment ('real experience') to directly learn the policy and/or value functions. The model-based procedure uses the real experience to learn a dynamics model $P(S_{t+1}, R_{t+1}|S_t, A_t)$ which can be used to generate artificial/simulated experience as another way to update policy and value functions.

The animal learning community in the early 20th century saw a divide between Thorndike's Law of Effect (Thorndike, 1933) and Tolman's Cognitive Maps (Tolman, 1948). Thorndike posited that humans associate rewards to actions and our future choices are driven by the type of reward we receive. On the other hand, Tolman stated that learning can still happen in the case that a reward is not immediately received, strengthening the argument for a type of **latent learning**, requiring goal-directed planning and reasoning. Thorndike's Law of Effect and Tolman's Cognitive Maps have served as foundational behavioral evidence for the two major types of learning systems concerned with action valuation in our brain, model-free and model-based learning.

Model-based learning systems involve building mental models through experience. There is evidence that model-based algorithms are implemented in biological systems. For instance, Gläscher et al. (2010a) observed increased activity in the lateral prefrontal cortex when previously unknown state transitions were observed. This evidence showed that the brain integrates unknown transitions into its transition model. Additionally, the hippocampus might play a role in integrating information about the current task and behavioral context. This integration might rely on synchronous activity in the theta band of frequencies (Benchenane et al., 2010; Hyman et al., 2011; Sirota et al., 2008; Womelsdorf et al., 2010). Langdon, Sharpe, Schoenbaum, and Niv (2018) reviewed recent findings of the association of dopaminergic prediction errors with model based learning and hypothesized that the underlying system might be multiplexing model-free scalar RPEs with model-based multi-dimensional RPEs.

Although there have been numerous advancements in finding neural correlates for model-free reinforcement learning (Delgado, Nystrom, Fissell, Noll, & Fiez, 2000; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; Knutson & Gibbs, 2006), the last two decades have witnessed research that bolsters evidence for the existence of a model-based system especially in a combined setting with the model-free learning system (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Gläscher, Daw, Dayan, & O'Doherty, 2010b; Ito & Doya, 2011; Kool, Cushman, & Gershman, 2018; Seo et al., 2009). Human neural systems are known to use information from both model-free and model-based sources (Daw, Niv, & Dayan, 2005; Gläscher et al., 2010a; Pan, Sawa, Tsuda, Tsukada, & Sakagami, 2008). Typical experiments involve a multi-staged decision making task while simultaneously recording BOLD (blood-oxygen-level-dependent) signals through fMRI. Results from these experiments suggest strongly coupled decision making systems (existence of reward prediction probability signals of both types of information (Lau & Glimcher, 2008)) in the ventromedial prefrontal cortex (Daw et al., 2011; Tsutsui et al., 2016), ventral striatum (Daw et al., 2011) and a model based behavior in the lateral pre-frontal cortex (Daw et al., 2005; Dolan & Dayan, 2013; Gläscher et al., 2010b) and anterior cingulate cortex (Akam & Walton, 2021). Combined social information and reward history can also be traced to the different regions of the anterior cingulate cortex (Apps, Rushworth, & Chang, 2016).

Recently, the idea of **successor representations** (Dayan, 1993) has been revived, especially in the context of how humans build cognitive maps of the environment. The underlying idea behind successor representations is to build a 'predictive map' of states of the environment in terms of future state *occupancies*. Recent modeling work Russek et al. (2017) has re-confirmed previous work on how successor representations might lie between model based and model free learning when compared over the spectrum of flexibility and efficiency. Gershman (2018) explains and summarizes recent literature on successor representations and its neural correlates. Akam and Walton (2021) recently proposed that successor representations are a key module in how model-based reinforcement learning systems are coupled with model-free reinforcement learning systems. Specifically, they proposed that future behaviors and RPEs are refined by imagined (or offline) planning and successor type representations to be incorporated as and when necessary (during online interaction).

Humans are known to develop *habits*: fast decisions or action sequences over time (Dolan & Dayan, 2013). Habits have been associated with model-free reinforcement learning, in particular, as responses to stimuli i.e., forming an association between an action and antecedent stimuli (Balleine & O'Doherty, 2010; Dezfouli & Balleine, 2012; Miller, Shenhav, & Ludvig, 2019). Additionally it was also recently observed that in the context of habit development for a set of individuals there was no involvement of model-based learning (Gillan, Otto, Phelps, & Daw, 2015).

The idea of building models is much more central and has a lot of implications to other aspects of human intelligence including but not limited to structure learning and social intelligence. Lake et al. (2017) suggest that human learning systems arbitrate to tradeoff between flexibility and speed.

4.3. Distributional reinforcement learning

In classical temporal difference learning, as discussed earlier, the value of a state is the expectation of cumulative future reward. As shown by Eq. (4), this expectation is expressed as the **mean** over future rewards. This measure is limited since it does not account for variance. Also, because the mean is a point estimator, it cannot capture multimodal return distributions. Much recent work has been done on developing a distributional framework for RL that maintains a return **distribution** rather than a single average value over future rewards (Bellemare et al., 2017). The formulation of this idea, termed as **distributional reinforcement learning**, replaces the value function $V(s_t)$ in the temporal difference update (Eq. (6)) with Z, the probability distribution of returns for state s_t.

$$\delta_t = R_{t+1} + \gamma Z(s_{t+1}) - Z(s_t) \tag{11}$$

$Z(s_t)$ stores the probability of occurrence for each value of return possible at s_t. Since it stores the complete distribution of

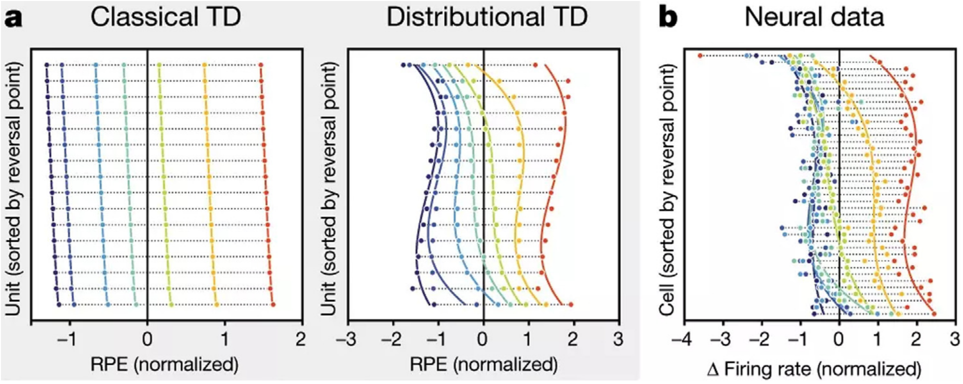

Fig. 4. Comparison of distributional TD and classical TD RPEs. On each trial, the animal receives one of seven possible reward values, chosen at random. **a.** RPEs produced by classical and distributional TD simulations. Each horizontal line is one simulated neuron. Each color corresponds to a particular reward magnitude. The x axis is the cells response when reward is received. In classical TD, all cells expected approximately the same RPE for a given reward signal. In contrast, in distributional TD, the simulated neurons showed significantly different degrees of reward expectation. **b.** Responses recorded from light-identified dopamine neurons in mice on the same task. A large variation in RPEs was observed between cells, which resembles distributional TD more than the classical TD simulation.
Source: Dabney et al. (2020).

return, the expected return (value function) is simply its expectation.

$$V(s_t) = \mathbb{E}[Z(s_t)] \tag{12}$$

Additionally, learning the return distribution helps the agent account for variance and capture multimodality. Its benefit can be immediately seen when we consider the simple example of a one step task. Suppose there are three states A, B and C. The agent starts at A and can choose to either go to B or C. Unknown to the agent: reward obtained at B is either +1 or −1 with equal probability, while reward at C is always 0.5. Value functions learned using classical TD would learn that B and C are equivalent since expected reward is equal. On the other hand, value functions learned using distributional TD would know that B is in fact much more "risky" than C.

Practically, it is difficult to assign probabilities to every possible return value in cases where the distribution is continuous. To solve this issue, algorithms usually discretize (bin) the return distribution with the number of bins tuned as a hyperparameter. For example, the C51 model uses 51 bins to represent the return distribution for Atari games (Bellemare et al., 2017).

Past work had provided evidence for distributional coding in the brain for non-RL domains (Pouget et al., 2013). Moreover, distributional reinforcement learning had earlier been shown to be biologically plausible (Dabney et al., 2018; Lammel et al., 2014). Recently, Dabney et al. (2020) carried out single-unit recordings of the ventral tegmental area (VTA) in mice and showed that for a given dopamine-based reward, different cells show different Reward Prediction Errors (RPEs). These RPEs can be either positive or negative which indicates that some cells are optimistic i.e., expect a larger reward than what is obtained, while others are pessimistic and expect a lower reward. Each cell RPE here is analogous to a bin used in practical implementations of distributional RL, since the cell's spiking activity represents expectation of a specific reward prediction error. Eq. (13) describes how the cell RPEs together form the Z distribution seen in the RL formulation.

$$\begin{aligned} \delta_{t1} &= R_{t+1} + \gamma V_1(s_{t+1}) - V_1(s_t) \\ \delta_{t2} &= R_{t+1} + \gamma V_2(s_{t+1}) - V_2(s_t) \\ &\vdots \\ \delta_{tN} &= R_{t+1} + \gamma V_N(s_{t+1}) - V_N(s_t) \\ Z(s_t) &= \{\delta_{t1}, \delta_{t2}, \dots \delta_{tN}\} \end{aligned} \tag{13}$$

where $n = 1 \dots N$ represent N neurons and δ_n is the RPE of the nth neuron. All of these RPEs taken together form the Z distribution we saw during the discussion on distributional RL.

Through extensive experiments, they compared the distributional coding with other models that attempt to explain RL in neural circuits, and showed that distributional RL most accurately predicts RPE reversal points and future rewards in the brain. Fig. 4 shows plots comparing distributional TD and classical TD on points obtained via single cell recording.

Unlike many of the RL algorithms we have seen so far, distributional RL is one of the algorithms whose involvement in neural circuits was identified after the idea was first independently proposed in AI literature. Hence, it offers evidence that better and more efficient computational models can potentially result in advances in brain research.

4.4. Meta reinforcement learning

While modern deep reinforcement learning methods have been able to achieve superhuman performance on a variety of tasks, they are many orders less sample efficient than the average human and possess weak inductive biases that deter transfer of learned knowledge (Lake et al., 2017; Marcus, 2018).

One way to increase sample efficiency is to avoid learning tasks from scratch each time and instead use previous learning experiences to guide the current learning process. In machine learning literature, this leveraging of past experience to speed up learning of new tasks is called **meta-learning** (Schaul & Schmidhuber, 2010). The original idea of "learning to learn" is often attributed to Harlow's 1949 work (Harlow, 1949) wherein a monkey was presented with two unseen objects, only one of which contained a reward. The monkey was then made to pick one of the objects after which the reward was revealed and the positions of the objects possibly reversed. All of this constituted a single trial. A given set of objects were used for a set of 6 trials before switching them for different objects, observing 6 trials and so on. The reward when tracked across several such rounds yielded an interesting observation. As the monkey was exposed to more sets of objects, the number of steps it needed to solve the problem for a new object set decreased. Thus, the monkey demonstrated capabilities of transferring knowledge between similar tasks. A recent idea that has helped in modeling such behavior is the hypothesis that underlying the fast learning problem of each object set, there was a slow learning process that figured out the problem dynamics and helped the monkey improve its sample efficiency on related problems (Botvinick et al., 2019).

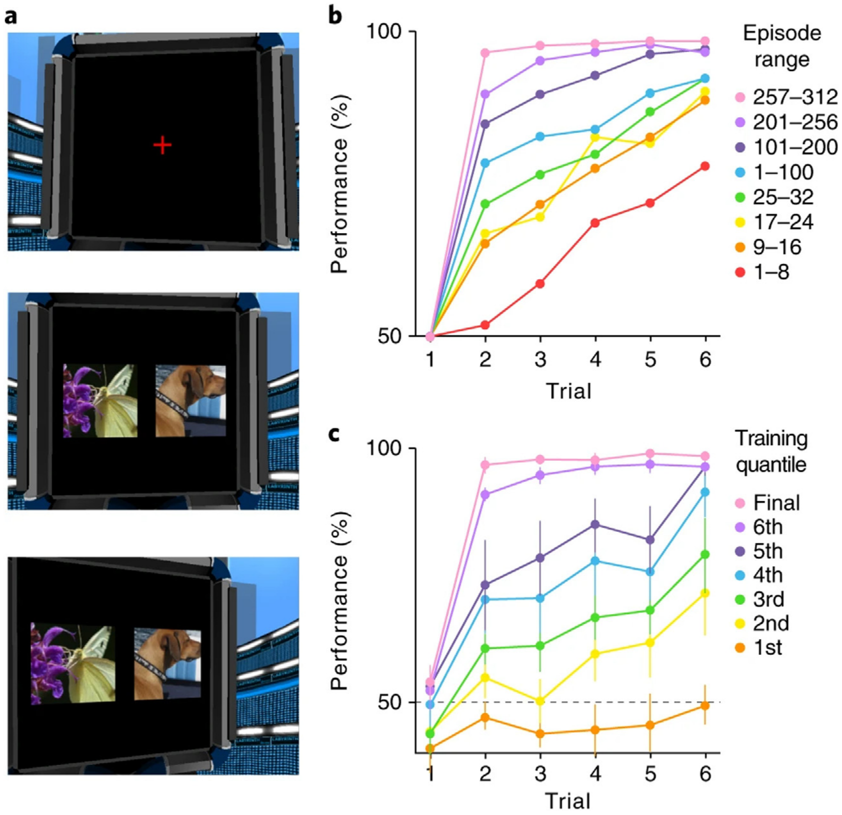

Fig. 5. Comparison between meta RL's (recurrent network) and Harlow's experimental results. **a.** Inputs for simulation experiments showing fixation cross (top), initial stimulus (middle), and outcome of saccade (bottom). **b.** Performance reward gathered during each trial of Harlow's monkey experiment (Harlow, 1949). **c.** Simulation performance at different phases of training. Performance improvement trend over time resembles Harlow's results.
Source: Wang et al. (2018).

Very recently, this core idea has been applied to reinforcement learning as meta RL to accelerate the learning process. Meta RL can be formulated as a modified version of the classical RL problem where the policy $\pi(s_t)$ now additionally depends on the previous reward r_t and action a_{t-1}, thus becoming $\pi(s_t, a_{t-1}, r_t)$. These additional dependencies allow the policy to internalize dynamics of MDP. Recurrent networks when used as part of a reinforcement learning algorithm for tasks similar to Harlow's yielded similar reward curves. This suggests that over a long period of exposure to related tasks, RNNs are able to capture the underlying activity dynamics due to their ability to memorize sequential information (Duan et al., 2016; Wang et al., 2016) (Fig. 5). Wang et al. (2018) also noticed that such meta-learning methods formed a part of dopaminergic reward-based learning in the prefrontal cortex of the brain. Much recent work on the prefrontal cortex (PFC) suggests that humans do more than just learning representations of expected reward. The PFC encodes latent representations on recent rewards and choices, and some sectors also encode the expected values of actions, objects, and states (Barraclough et al., 2004; Kim & Shadlen, 1999; Padoa-Schioppa & Assad, 2006; Rushworth & Behrens, 2008; Tsutsui et al., 2016).

As an extension of meta RL, a set of recent computational approaches combines episodic memory with meta RL which results in stronger inductive biases and higher sample efficiency. Inspired from the observation that episodic memory circuits reinstate patterns of activity in the cerebral cortex (Ritter et al., 2018b), Ritter et al. (2018a) developed a framework for how such episodic memory functions can strategically reuse information about previously learned tasks and thereby improve learning efficiency (Santoro et al., 2016; Wayne et al., 2018). This work also evidences recent interactions between meta model-based control and episodic memory in human learning (Vikbladh et al., 2017).

4.5. Causal reinforcement learning

The ability of humans to reason about cause and effect relationships is not unknown. The field of causal inference and reasoning looks at studying this paradigm in great detail. Pearl in 2019 (Pearl, 2009) introduced a 3-level causal hierarchy which he called the "causal ladder". The causal ladder consists of associations, interventions and counterfactuals at the different levels of the hierarchy. Or more simply: seeing, doing and imagining respectively.

Much of supervised learning is at the first level i.e. learning associations from observed data. Reinforcement learning, in contrast, simultaneously has an agent learning associations and performing interventions in a world. Most model-free RL algorithms learn observation action associations by performing interventions in the environment. It is important to note here that to learn from interventions, an agent does not necessarily have to maintain a model of the world as long as it has access to the world or a good model of it. Humans are known to perform interventions to infer basic cause–effect relationships of the world from a very young age. It has been shown that children can learn to identify causal relationships in basic 3-variable models by performing interventions (McCormack et al., 2016; Sobel & Sommerville, 2010). What this means is that, in their environments, children explore and perform interventions freely to explore and understand the world.

The third level of the hierarchy, which consists of counterfactuals (or imagining), usually requires an agent maintaining an internal model of the world, using which it tries to imagine counterfactual 'what if?' scenarios. Being able to construct intuitive world models is a hallmark of human intelligence because it helps us in planning and reasoning among other aspects of our daily interactions. In particular, models encoding causal structure serve as strong priors for planning and reasoning about tasks (Lake et al., 2017). These causal models are especially helpful in contemplating scenarios for better planning i.e., reasoning through counterfactual actions and not necessarily taking the immediately rewarding action. A counterfactual scenario can be thought of as a combination of observation and imagined interventions. In formal terms of causal inference, with a structural causal model (SCM) of the environment, counterfactual scenarios can be simulated. Recent works Buesing et al. (2018) and Oberst and Sontag (2019) focus on this by modeling the environment as a structural causal model. By building an SCM, it can be intervened to take an action that was not originally taken and simulate counterfactual experience. Compared to Vanilla model-based policy search, a counterfactual policy search has been shown to perform better (Buesing et al., 2018).

In an experiment involving Parkinson's disease diagnosed patients, sub-second temporal resolution dopamine levels were monitored through blood-oxygen-level-dependent (BOLD) imaging to study the action of specific neurotransmitters (Kishida et al., 2016). It was observed that the subsecond dopamine fluctuations encoded reward and *counterfactual prediction errors* in superposition, paving the way for neural evidence of counterfactual outcomes. Much work on factual learning has suggested that humans have a valence-induced bias towards positive prediction errors over negative prediction errors (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017; Ouden et al., 2013). This means humans prefer outcomes that lead to higher reward than expected than outcomes that lead to lower rewards than expected. But as previously mentioned, an important sect of human intelligence involves learning from foregone outcomes. Recently, it was shown that humans have a bias towards negative prediction errors rather than positive prediction errors during counterfactual learning (Palminteri et al., 2017). More generally, this means that humans have a "confirmation bias" towards their own choices that guides learning than either positive or negative prediction errors. Recent evidence (Pischedda et al., 2020) showed that counterfactual outcomes are encoded negatively in the ventral medial prefrontal cortex and positively in the dorsal medial prefrontal cortex. Consistent with the previously described findings, factual learning is encoded in the opposite pattern in both these regions (Klein, Ullsperger, & Jocham, 2017; Li & Daw, 2011). More generally, experimental findings suggested that human learning behavior is significantly increased when complete information is presented (i.e., counterfactual outcomes are presented). It has also been shown that neurons in the lateral frontal polar cortex (lFPC), dorsomedial frontal cortex (DMFC), and posteromedial cortex (PMC) encode reward-based evidence favoring the best counterfactual option at future decisions (Boorman et al., 2011).

Building internal models of the world allows humans to deal with partial observability on a daily basis (Gershman & Daw, 2017). These internal models are grounded in concepts observed through partitioning observations in a well-organized manner i.e. structure learning from observations, which allows them to make decisions under incomplete information. In the context of reinforcement learning, algorithms depend on a representation of the environment and hence directly affect the algorithm's learning quality (efficiency and efficacy) (Gershman, Norman, & Niv, 2015). An important component of the structure learnt is causality. In other words, the learnt structure should be able to capture the appropriate discrete causal structure underlying the continuous world (Gershman & Niv, 2010). The ability of humans to build a structure consisting of *latent causes* (i.e. hidden causes) from observations is remarkable yet not entirely understood and allows them to perform inference over these latent causes to reason (Gershman & Niv, 2012). Recent work in computational modeling supports this by building a Bayesian non-parametric prior over latent causes (Soto et al., 2014). The idea behind structure learning is central to human intelligence, even beyond the context of reinforcement learning (Braun, Mehring, & Wolpert, 2010).

A standard modeling framework for building causal knowledge of the world (artificial or real) requires first identifying potential qualitative relationships (i.e., learning the structure of variables) followed by estimating their strength (i.e., quantitative cause–effect estimation). Analogously, humans tend to extract relevant information to build high-level causal knowledge and then improve the knowledge by estimating each cause's quantitative effect. As previously noted Lagnado, Waldmann, Hagmayer, and Sloman (2007), much past research on causal learning and reasoning has focused on quantitative estimation of cause–effect relationships from pre-fixed variables (usually available through a dataset). However, building qualitative causal relationships is also important and foundational to recent work on "intuitive theories" (Gerstenberg & Tenenbaum, 2017). As an example of building qualitative causal relationships, there are cases where we may know that two variables have a particular relationship but we may not know the extent to which they are related and which variable is *causing* the other. Building causal knowledge of relevant high level variables bottom-up (e.g. from rich inputs such as pixels) is a fairly challenging task. CausalWorld (Ahmed et al., 2020), a recently proposed environment suite, involves learning the causal structure of high level variables and should enable research in this direction in the AI community. Artificial environments also already allow the ability to intervene, which is foundational to *evaluating* which causal structure might be a better model of the world (Hagmayer, Sloman, Lagnado, & Waldmann, 2007; Lagnado & Sloman, 2004).

There have been various theories behind the psychology of causal induction, with the two most prominent being the causal power theory (Cheng, 1997) and the ΔP model (Lober & Shanks, 2000). These models address the question of how humans learn the association between causes and effects. Although these models have been quite widely argued and debated over, they fail to account for formal definitions in terms of graphical models, an idea central to modern ideas in causality (Pearl, 2009). Recently, Tenenbaum and Griffiths (2001) postulated that performing inferences over learned causal structures is a central human tendency and in an attempt to bridge the psychology, computer science and philosophy literatures proposed Bayesian Causal Support and the χ^2 model. Both these models extend the original causal power theory model and the ΔP model by incorporating Bayesian Inference which works on graphical models rather than simple parameter estimation.

4.6. Hierarchical reinforcement learning

General RL algorithms scale poorly with the size of state space due to difficulty in exploration and effects of catastrophic forgetting that arise in larger task domains. In order to solve this problem also known as the **scaling problem**, a popular computational framework known as temporal abstraction (Barto & Mahadevan, 2003; Dietterich, 2000; Parr & Russell, 1998; Sutton et al., 1999b) was developed, which suggested learning temporally extended actions that were composed of primitive actions. A common way in which these temporally extended actions are implemented is through the **options** (Sutton et al., 1999b) framework. Options are actions that span more than a single state and consist of multiple primitive actions. For example, the option "walk towards the door" would be composed of several primitive actions including motor movements and maintaining balance. Mathematically, the simplest form of options is formulated as a triplet $(\pi, \beta, \mathbb{I})$ where π is the policy, $\beta(t) \in [0, 1]$ is the termination condition i.e. the probability that the option terminates at the current state, and $\mathbb{I}$ is the initiation set of states, which determines if the option can be chosen at the current state.

Hierarchical reinforcement learning combines temporally extended actions to maximize reward on goal-directed tasks. In psychology, hierarchy has played a significant role in explaining goal-directed behavior (Anderson et al., 2004; Botvinick & Plaut, 2004; Lashley, 1951; Miller et al., 2017; Newell et al., 1959; Schneider & Logan, 2006; Zacks et al., 2007). Even in neuroscience, existing literature accounts for the prefrontal cortex being largely responsible for hierarchical behavior (Badre, 2008; Botvinick, 2008; Courtney et al., 2007; Fuster, 1989; Koechlin et al., 2003). Thus, even though HRL was not developed to answer questions in psychology and neuroscience, it addresses an issue with standard RL methods which might also be prevalent in the brain.

Early work in psychology had also postulated the presence of hierarchy in human behavior. That determining the sequence of primitive actions requires higher-level representations of task context, was first formalized by Lashley in 1951 (Lashley, 1951). The concept of task representation (Cohen et al., 1990; Cooper & Shallice, 2000; Monsell, 2003) is very similar to the option construct (discussed earlier) that was developed in reinforcement learning literature. Empirical evidence that human mental representations are organized hierarchically was also found (Newtson et al., 1976; Zacks & Tversky, 2001). Hierarchy has also been observed in the behavior of children through their childhood. Children learn elementary skills which are gradually integrated into more complex skills and knowledge as they grow (Bruner, 1973; Fischer, 1980; Greenfield et al., 1972).

The strongest resemblance to HRL is found in the production-system based theories of cognition, especially ACT-R (Anderson et al., 2004) and Soar (Lehman, Laird, & Rosenbloom, 1996). These frameworks propose that the solution to a problem can make use of shorter action sequences called "chunks". Given a problem, high-level decisions can be used to trigger these chunks. Though these frameworks are similar to HRL in many regards, they differ in the aspect of not being based around a single reward maximization objective.

Thus, HRL shares attributes with multiple theories in behavioral psychology. However, ideas in psychology go even beyond the positive transfer problem that we have until now discussed i.e., sequencing temporally abstracted actions to develop goal-directed policies; and discuss downsides of hierarchical learning in humans. Luchins in 1942 (Luchins, 1942) introduced the negative transfer problem; that pre-existing knowledge with context differing from the current problem can hinder problem-solving in human subjects. Surprisingly, HRL aligns with behavioral theories even in these downsides. A direct analog to the negative transfer problem has been observed in HRL (Botvinick et al., 2009).

As a natural extension to the above-discussed work that provides strong evidence for the hierarchical nature of human behavior, recent neuroscience research has delved deeper into neural correlates for hierarchy. Ribas-Fernandes et al. (2011) showed for the first time that the medial PFC processes subgoal-related RPEs. This was followed by further study on the nature of these RPEs, which showed that subgoal-related RPEs are unsigned (Ribas-Fernandes et al., 2019). These results reinforce recent evidence that RPEs for different task levels are induced as separable signals in the basal ganglia (Diuk, Tsai, Wallis, Botvinick, & Niv, 2013). Balaguer et al. (2016) extend these studies to hierarchical planning, by reporting that neural signals in the dmPFC encode the cost of representing hierarchical plans.

A major challenge in hierarchical RL has been option discovery, that is, how to form chunks of 'reusable' actions from primitive ones. One approach to option discovery is to keep a record of states that occur frequently on paths to goals and label them as subgoals or bottleneck states that a good solution must pass through (Mcgovern, 2002; Pickett & Barto, 2002; Schwartz & Thrun, 1995). This bottleneck theory is also consistent with work that shows that humans are sensitive to repeating sequences of events. Another approach to option discovery from HRL literature is to construct a graph of states and all the transitions possible. Then, graph partitioning can be used to identify bottleneck states which can be used as subgoals during the learning process (Şimşek et al., 2005; Mannor et al., 2004; Menache et al., 2002). Existing work in psychology provides empirical evidence that children identify causal representations that they then integrate into a large causal model (Gopnik et al., 2004; Gopnik & Schulz, 2004; Sommerville & Woodward, 2005a, 2005b). More recent work in HRL uses task agnostic approaches to discover options, by using intrinsic rewards for exploration (Barto et al., 2004; Singh et al., 2004). Existing neuroscience literature also provides evidence for something similar to this notion of intrinsic reward driven learning. It has been found that the same dopaminergic neurons that code reward prediction errors also respond to novel stimuli (Bunzeck & Düzel, 2006; Redgrave & Gurney, 2006; Schultz et al., 1993). In psychology, intrinsic rewards are strongly tied to ideas of motivation which like their RL counterpart, depend on the animal's action as well as current state. For example, a rabbit in a state of hunger is more motivated to obtain a food reward. Despite the similarity, RL algorithms still do not capture the correlation between strength/vigor of actions and motivation (which is seen in animals) (Skinner, 1988), since simple environments commonly used in RL do not allow the agent to control the rate at which actions are performed.

A parallel line of research tries to approach the problem of hierarchy discovery, by building computational models of how the brain might do so. Solway et al. (2014) developed a Bayesian model selection approach to identify optimal hierarchies. Structure discovered using this approach explains various behavioral effects like bottleneck states, transitions and hierarchical planning. Tomov et al. (2020) extended this approach by proposing a Bayesian model that additionally captures uncertainty-based learning and reward generalization, both of which were unexplained by previous models (Schapiro et al., 2013; Solway et al., 2014).

5. Discussion

Reinforcement learning's emergence as a state-of-the-art machine learning framework and concurrently, its promising ability to model several aspects of biological learning and decision making, have enabled research at the intersection of reinforcement learning, neuroscience and psychology. Through this review, we have attempted to comprehensively illustrate the various classes of RL methods and validation of these methods in brain science literature. Table 1 summarizes all the findings discussed in the paper and provides a mapping between RL concepts and evidence corresponding to them in neuroscience and psychology.

As detailed in earlier sections, findings in brain science have played an important role in inspiring new reinforcement learning ideas such as value functions, eligibility traces, TD learning, meta RL and hierarchical RL. Given this success, we now discuss some ways in which existing work in brain science could potentially further influence reinforcement learning research.

- **Distinct rewards:** In the brain, distinct rewards condition behavior for a single decision, and each reward might be encoded by different brain regions (Cai et al., 2011; Kim

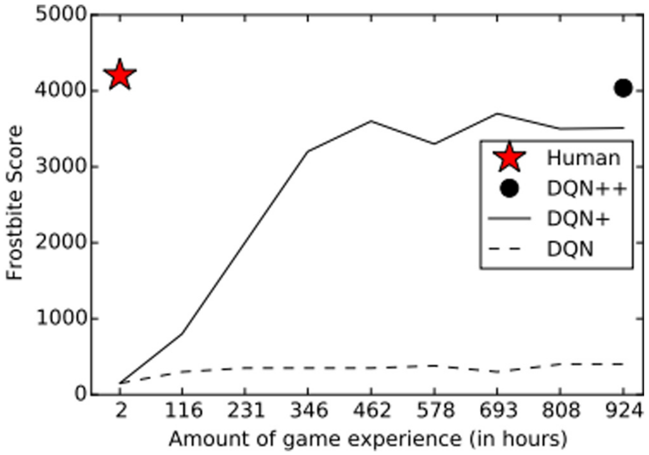

Fig. 6. Comparison of humans with state of the art deep reinforcement learning methods on the Atari game 'Frostbite'. The best Deep RL agent, DQN++ takes significantly more time (∼400 h) than humans to achieve similar performance. Other agents (DQN and DQN+) fail to match performance even after a large amount of experience.
Source: Lake et al. (2017).

et al., 2008; Murdoch et al., 2018; Padoa-Schioppa & Assad, 2006; Pastor-Bernier & Cisek, 2011; Seo et al., 2009; So & Stuphorn, 2010; Tremblay & Schultz, 1999; Wallis & Miller, 2003). Rewards in modern deep RL algorithms always update all parameters in the network, which is inefficient and could potentially impact learning in cases where objectives desired by two or more reward signals conflict.

- **Action eligibility traces:** Neurons in the orbitofrontal cortex are believed to encode relationships between actions and their corresponding outcomes (Barraclough et al., 2004; Kim et al., 2009; Seo et al., 2009; Sul et al., 2010). RL algorithms predominantly use only state eligibility traces, which is sufficient when the same actions are available at every state. However, in real-world cases where different states allow different sets of actions, a trace for actions might be useful.
- **'Imagination' replay:** Experience replay in deep RL models tend to replay past-experienced sequences of states in the order in which they occurred (Mnih et al., 2015; Schaul et al., 2016). Recent observations in neuroscience suggest that hippocampal replay might replay states based on the expected sequence according to an internal world model (Liu et al., 2019). This event-based replay is currently unexplored in RL literature.
- **Action vigor:** Work on the role of motivation in animal behavior observes a strong correlation between motivation and action vigor (Skinner, 1988). Though modern RL algorithms do model motivation using intrinsic motivation and curiosity (Barto et al., 2004; Singh et al., 2004), most RL environments do not allow agents to adjust the rate at which actions are applied. This currently prevents potentially interesting analysis on vigor, caution and motivation in RL agent behavior.
- **Grounded language learning**: Language plays an important role in many aspects of human learning such as exploration and forming internal representations. Recent work in computational linguistics also emphasizes the role of pragmatic communication in human representation learning (Cohn-Gordon, Goodman, & Potts, 2019). Recent work in RL has embraced language and has made strides towards proposing algorithms and challenges that present language as an important tool for learning (Colas et al., 2020; Narasimhan, Barzilay, & Jaakkola, 2018).
- **Social learning**: A key aspect of human learning is our ability to learn via social interaction. Theory of Mind (TOM) (Premack & Woodruff, 1978) involves the ability to understand others by attributing mental beliefs, intents, desires, emotions, and knowledge to them. Recent studies show that neural signals carry social information, as reviewed comprehensively by Insel and Fernald (2004). Social situations are much more complex which may not be completely expressible from a single reward value.
- **Modularity**: The human brain is known to be very modular in nature. Specific regions specialize in specific roles such as the occipital lobe deals primarily with vision signals, the temporal lobe deals primarily with auditory signals, etc. Modularity makes learning systems flexible, adaptable and allow quick transfer of decision making knowledge allowing humans to be versatile at learning. Although work exists on modular reinforcement Learning (Simpkins & Isbell, 2019; Sprague & Ballard, 2003), most current RL algorithms and computational models lack modularity and are unable to adapt to new tasks, domains, etc quickly, as there is a lack of formalism guiding the definition and use of modularity. It is a core component for both our internal world models as well as in sequential decision making (Fodor, 1983). In the former case, modularity leads to building parts that may specialize in a domain, for instance, processing a particular type of sensory signal. In the latter case, modularity results in efficient adaptation to new tasks, domains, etc. Recently, Chang, Kaushik, Griffiths, and Levine (2021) built principles for modular decision making in the model-free setting by establishing principles to identify credit assignment algorithms which allow independent modification of components.

Leveraging and implementing ideas from neuroscience and psychology for RL could also potentially help address the important issues of sample efficiency and exploration in modern-day algorithms. Humans can learn new tasks with very little data (Fig. 6). Moreover, they can perform variations of a learned task (with different goals, handicaps etc.) without having to re-learn from scratch (Lake et al., 2017; Tsividis et al., 2017). Unlike deep RL models, humans form rich, generalizable representations which are transferable across tasks. Principles such as compositionality, causality, intuitive physics and intuitive psychology have been observed in human learning behavior, which if replicated or modeled in RL, could produce large gains in sample efficiency and robustness (Lake et al., 2017).

Studies in neuroscience and psychology have also been motivated by ideas originally developed in RL literature. For instance, the temporal difference model made several important predictions about classical conditioning in animals (Sutton & Barto, 1981, 1990), some of which were verified experimentally only later (Kehoe et al., 1987). Studies into the possibility of distributional coding in the brain (Dabney et al., 2020) were motivated by the distributional TD model developed a few years before (Bellemare et al., 2017). Another example is meta reinforcement learning which according to Wang et al. (2018) was first observed as an emergent phenomenon in recurrent networks trained using an RL algorithm, which inspired research into the possibility of the prefrontal cortex encoding similar behavior. All of these examples suggest that reinforcement learning is a promising model for learning in the brain and therefore that experimenting with RL models could yield predictions that motivate future research in neuroscience and psychology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was done while Ajay and Sharad were undergraduate students at BITS Pilani. Ajay and Sharad would like to thank the Cognitive Neuroscience Lab at BITS Pilani K K Birla Goa Campus for supporting and enabling this review.

References

Abe, H., & Lee, D. (2011). Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. *Neuron*, *70*(4), 731–741. http://dx.doi.org/10.1016/j.neuron.2011.03.026.

Ahmed, O., Träuble, F., Goyal, A., Neitz, A., Bengio, Y., Schölkopf, B., et al. (2020). Causalworld: A robotic manipulation benchmark for causal structure and transfer learning. arXiv preprint arXiv:2010.04296.

Akam, T., & Walton, M. (2021). What is dopamine doing in model-based reinforcement learning? *Current Opinion in Behavioral Sciences*, *38*, 74–82.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036.

Apps, M., Rushworth, M., & Chang, S. W. C. (2016). The anterior cingulate gyrus and social cognition: Tracking the motivation of others. *Neuron*, *90*, 692–707.

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

Balaguer, J., Spiers, H., Hassabis, D., & Summerfield, C. (2016). Neural mechanisms of hierarchical planning in a virtual subway network. *Neuron*, *90*(4), 893–903.

Balleine, B., & O'Doherty, J. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*, 48–69.

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, http://dx.doi.org/10.1038/nn1209.

Barto, A. G., & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, *13*(1), 41–77.

Barto, A. G., Singh, S., & Chentanez, N. (2004). Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd international conference on development and learning* (pp. 112–119). Piscataway, NJ.

Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. In *34th international conference on machine learning, ICML*.

Bellman, R. (1954). The theory of dynamic programming. *American Mathematical Society. Bulletin*, http://dx.doi.org/10.1090/S0002-9904-1954-09848-8.

Benchenane, K., Peyrache, A., Khamassi, M., Tierney, P. L., Gioanni, Y., Battaglia, F. P., et al. (2010). Coherent theta oscillations and reorganization of spike timing in the hippocampal- prefrontal network upon learning. *Neuron*, http://dx.doi.org/10.1016/j.neuron.2010.05.013.

OpenAI Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. ArXiv.

Boorman, E., Behrens, T. E. J., & Rushworth, M. (2011). Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. *PLoS Biology*, *9*.

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*(1), 15958. http://dx.doi.org/10.1038/ncomms15958, Number: 1 Publisher: Nature Publishing Group.

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*(7), 997–1003. http://dx.doi.org/10.1038/nn.4573, Number: 7 Publisher: Nature Publishing Group.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in Cognitive Sciences*, *12*(5), 201–208.

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. http://dx.doi.org/10.1016/j.cognition.2008.08.011.

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, http://dx.doi.org/10.1037/0033-295X.111.2.395.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. http://dx.doi.org/10.1016/j.tics.2019.02.006, Publisher: Elsevier.

Botvinick, M., Wang, J. X., Dabney, W., Miller, K. J., & Kurth-Nelson, Z. (2020). Deep reinforcement learning and its neuroscientific implications. *Neuron*.

Braun, D., Mehring, C., & Wolpert, D. (2010). Structure learning in action. *Behavioural Brain Research*, *206*, 157–165.

Bruner, J. S. (1973). Organization of early skilled action. *Child Development*, http://dx.doi.org/10.1111/j.1467-8624.1973.tb02105.x.

Buesing, L., Weber, T., Zwols, Y., Racaniere, S., Guez, A., Lespiau, J.-B., et al. (2018). Woulda, coulda, shoulda: Counterfactually-guided policy search. arXiv preprint arXiv:1811.06272.

Bunzeck, N., & Düzel, E. (2006). Absolute coding of stimulus novelty in the human substantia nigra/VTA. *Neuron*, http://dx.doi.org/10.1016/j.neuron.2006.06.021.

Busoniu, L., Babuska, R., De Schutter, B., & Ernst, D. (2017). *Reinforcement learning and dynamic programming using function approximators*. CRC Press.

Cai, X., Kim, S., & Lee, D. (2011). Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. *Neuron*, *69*(1), 170–182. http://dx.doi.org/10.1016/j.neuron.2010.11.041.

Chang, M., Kaushik, S., Griffiths, T. L., & Levine, S. (2021). Modularity in reinforcement learning via algorithmic independence in credit assignment. In *Learning to learn-workshop at ICLR 2021*.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367–405.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the stroop effect. *Psychological Review*, *97*(3), 332.

Cohn-Gordon, R., Goodman, N., & Potts, C. (2019). An incremental iterated response model of pragmatics. In *Proceedings of the society for computation in linguistics (SCiL) 2019* (pp. 81–90).

Colas, C., Karch, T., Lair, N., Dussoux, J.-M., Moulin-Frier, C., Dominey, P., et al. (2020). Language as a cognitive tool to imagine goals in curiosity driven exploration. *Advances in Neural Information Processing Systems*, *33*.

Cooper, R., & Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, http://dx.doi.org/10.1080/026432900380427.

Courtney, S. M., Roth, J. K., & Sala, J. B. (2007). A hierarchical biased-competition model of domain-dependent working memory maintenance and executive control. *Working Memory: Behavioural and Neural Correlates*, 369–383.

Şimşek, O., Wolfe, A. P., & Barto, A. G. (2005). Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the 22nd international conference on machine learning - ICML '05* (pp. 816–823). Bonn, Germany: ACM Press, http://dx.doi.org/10.1145/1102351.1102454.

Curtis, C. E., & Lee, D. (2010). Beyond working memory: the role of persistent activity in decision making. *Trends in Cognitive Sciences*, *14*(5), 216–222. http://dx.doi.org/10.1016/j.tics.2010.03.006.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., et al. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671–675. http://dx.doi.org/10.1038/s41586-019-1924-6.

Dabney, W., Rowland, M., Bellemare, M. G., & Munos, R. (2018). Distributional reinforcement learning with quantile regression. In *32nd AAAI conference on artificial intelligence, AAAI 2018* (pp. 2892–2901).

Daw, N., Gershman, S., Seymour, B., Dayan, P., & Dolan, R. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, http://dx.doi.org/10.1038/nn1560.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*, 613–624.

Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., & Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, *84*(6), 3072–3077. http://dx.doi.org/10.1152/jn.2000.84.6.3072.

Dezfouli, A., & Balleine, B. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*.

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, *10*(10), 1241–1242. http://dx.doi.org/10.1038/nn1961.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, http://dx.doi.org/10.1613/jair.639.

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*(13), 5797–5805.

Dolan, R., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*, 312–325.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL^2: Fast reinforcement learning via slow reinforcement learning. (pp. 1–14). ArXiv.

Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain: A Journal of Neurology*, *126*(Pt 8), 1830–1837. http://dx.doi.org/10.1093/brain/awg180.

Fischer, K. W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, http://dx.doi.org/10.1037/0033-295X.87.6.477.

Fodor, J. A. (1983). *The modularity of mind*. MIT Press.

Foster, D. J. (2017). Replay comes of age. *Annual Review of Neuroscience*, *40*(1), 581–602. http://dx.doi.org/10.1146/annurev-neuro-072116-031538.

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, *440*(7084), 680–683. http://dx.doi.org/10.1038/nature04587.

Frank, M. J., Moustafa, A., Haughey, H., Curran, T., & Hutchison, K. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences*, *104*, 16311–16316.

Fuster, J. M. (1989). A theory of prefrontal functions: The prefrontal cortex and the temporal organization of behavior. *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe*, 157–192.

Gerfen, C. R., & Surmeier, D. J. (2011). Modulation of striatal projection systems by dopamine. *Annual Review of Neuroscience*, *34*, 441–466. http://dx.doi.org/10.1146/annurev-neuro-061010-113641.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*(11), Article e1004567.

Gershman, S. J. (2017). Reinforcement learning and causal models. *The Oxford Handbook of Causal Reasoning*, 295.

Gershman, S. (2018). The successor representation: Its computational logic and neural substrates. *The Journal of Neuroscience*, *38*, 7193–7200.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*(1), 101–128. http://dx.doi.org/10.1146/annurev-psych-122414-033625, Publisher: Annual Reviews.

Gershman, S., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, *20*, 251–256.

Gershman, S., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior*, *40*, 255–268.

Gershman, S., Norman, K., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. *Oxford Handbook of Causal Reasoning*, 515–548.

Gillan, C., Otto, A. R., Phelps, E., & Daw, N. (2015). Model-based learning protects against forming habits. *Cognitive, Affective & Behavioral Neuroscience*, *15*, 523–536.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010a). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, http://dx.doi.org/10.1016/j.neuron.2010.04.016.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. (2010b). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*, 585–595.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological Review*, *111*(1), 3.

Gopnik, A., & Schulz, L. (2004). *Vol. 8, Mechanisms of theory formation in young children* (pp. 371–377). Elsevier,

Greenfield, P. M., Nelson, K., & Saltzman, E. (1972). The development of rulebound strategies for manipulating seriated cups: A parallel between action and grammar. *Cognitive Psychology*, *3*(2), 291–310.

Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron*, *65*(5), 695–705. http://dx.doi.org/10.1016/j.neuron.2010.01.034.

Haber, S. N., Fudge, J. L., & McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *20*(6), 2369–2382.

Haber, S. N., & Knutson, B. (2010). The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology*, *35*(1), 4–26. http://dx.doi.org/10.1038/npp.2009.129.

Hagmayer, Y., Sloman, S. A., Lagnado, D. A., & Waldmann, M. R. (2007). Causal reasoning through intervention. *Causal Learning: Psychology, Philosophy, and Computation*, 86–100.

Hare, T., O'Doherty, J., Camerer, C., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience*, *28*, 5623–5630.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, http://dx.doi.org/10.1037/h0062474.

Heinrich, J., & Silver, D. (2015). Smooth UCT search in computer poker. In *IJCAI international joint conference on artificial intelligence*.

Hikosaka, O., Nakamura, K., & Nakahara, H. (2006). Basal ganglia orient eyes to reward. *Journal of Neurophysiology*, *95*(2), 567–584. http://dx.doi.org/10.1152/jn.00458.2005.

Hong, S., & Hikosaka, O. (2008). The globus pallidus sends reward-related signals to the lateral habenula. *Neuron*, *60*(4), 720–729. http://dx.doi.org/10.1016/j.neuron.2008.09.035.

Hong, S., & Hikosaka, O. (2011). Dopamine-mediated learning and switching in cortico-striatal circuit explain behavioral changes in reinforcement learning. *Frontiers in Behavioral Neuroscience*, *5*, 15. http://dx.doi.org/10.3389/fnbeh.2011.00015.

Hull, C. L. (1943). *Vol. 422, Principles of behavior*. Appleton-century-crofts New York.

Hyman, J. M., Hasselmo, M. E., & Seamans, J. K. (2011). What is the functional relevance of prefrontal cortex entrainment to hippocampal theta rhythms? *Frontiers in Neuroscience*, *5*, 24.

Insel, T. R., & Fernald, R. D. (2004). How the brain processes social information: searching for the social brain. *Annual Review of Neuroscience*, *27*, 697–722.

Ito, M., & Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, *21*, 368–373.

Iversen, S. D., & Mishkin, M. (1970). Perseverative interference in monkeys following selective lesions of the inferior prefrontal convexity. *Experimental Brain Research*, *11*(4), 376–386. http://dx.doi.org/10.1007/BF00237911.

Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, *10*(1), 100–107. http://dx.doi.org/10.1038/nn1825.

Kamin, L. J. (1967). Predictability, surprise, attention, and conditioning. *Punishment Aversive Behavior*.

Kehoe, E. J., Schreurs, B. G., & Graham, P. (1987). Temporal primacy overrides prior training in serial compound conditioning of the rabbit's nictitating membrane response. *Animal Learning & Behavior*, *15*(4), 455–464.

Kim, Y. B., Huh, N., Lee, H., Baeg, E. H., Lee, D., & Jung, M. W. (2007). Encoding of action history in the rat ventral striatum. *Journal of Neurophysiology*, *98*(6), 3548–3556. http://dx.doi.org/10.1152/jn.00310.2007.

Kim, S., Hwang, J., & Lee, D. (2008). Prefrontal coding of temporally discounted values during intertemporal choice. *Neuron*, *59*(1), 161–172. http://dx.doi.org/10.1016/j.neuron.2008.05.010.

Kim, J. N., & Shadlen, M. N. (1999). Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience*, http://dx.doi.org/10.1038/5739.

Kim, H., Sul, J. H., Huh, N., Lee, D., & Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(47), 14701–14712. http://dx.doi.org/10.1523/JNEUROSCI.2728-09.2009.

Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., et al. (2016). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences*, *113*(1), 200–205.

Klein, T. A., Ullsperger, M., & Jocham, G. (2017). Learning relative values in the striatum induces violations of normative decision making. *Nature Communications*, *8*.

Knutson, B., & Gibbs, S. E. (2006). Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacology*, *191*, 813–822.

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science*, http://dx.doi.org/10.1126/science.1088545.

Kool, W., Cushman, F. A., & Gershman, S. J. (2018). Competition and cooperation between multiple reinforcement learning systems. *Goal-Directed Decision Making*, 153–178.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation. *Causal Learning: Psychology, Philosophy, and Computation*, 154–172.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lammel, S., Lim, B. K., & Malenka, R. C. (2014). Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology*, *76*(PART B), 351–359. http://dx.doi.org/10.1016/j.neuropharm.2013.03.019, Publisher: Elsevier Ltd.

Langdon, A. J., Sharpe, M., Schoenbaum, G., & Niv, Y. (2018). Model-based predictions for dopamine. *Current Opinion in Neurobiology*, *49*, 1–7.

Lashley, K. S. (1951). *Vol. 21, The problem of serial order in behavior*. Bobbs-Merrill Oxford, United Kingdom.

Lau, B., & Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, *58*(3), 451–463.

Lazaric, A., & Ghavamzadeh, M. (2010). Bayesian multi-task reinforcement learning. In *ICML-27th international conference on machine learning* (pp. 599–606). Omnipress.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, http://dx.doi.org/10.1146/annurev-neuro-062111-150512.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*.

Lehman, J., Laird, J., & Rosenbloom, P. (1996). A gentle introduction to soar, an architecture for human cognition. *Invitation To Cognitive Science*.

Lengyel, M., & Dayan, P. (2007). Hippocampal contributions to control: the third way. *Advances in Neural Information Processing Systems*, *20*, 889–896.

Levey, A. I., Hersch, S. M., Rye, D. B., Sunahara, R. K., Niznik, H. B., Kitt, C. A., et al. (1993). Localization of D1 and D2 dopamine receptors in brain with subtype-specific antibodies. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(19), 8861–8865. http://dx.doi.org/10.1073/pnas.90.19.8861.

Li, J., & Daw, N. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *The Journal of Neuroscience*, *31*, 5504–5511.

Lin, Z., Zhao, T., Yang, G., & Zhang, L. (2018). Episodic memory deep Q-networks. arXiv:1805.07603.

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. (2019). Human replay spontaneously reorganizes experience. *Cell*, *178*(3), 640–652.e14. http://dx.doi.org/10.1016/j.cell.2019.06.012.

Lo, C.-C., & Wang, X.-J. (2006). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience*, *9*(7), 956–963. http://dx.doi.org/10.1038/nn1722.

Lober, K., & Shanks, D. (2000). Is causal induction based on causal power? Critique of cheng (1997). *Psychological Review*, *107 1*, 195–212.

Louie, K., & Wilson, M. A. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, *29*(1), 145–156. http://dx.doi.org/10.1016/s0896-6273(01)00186-6.

Luchins, A. S. (1942). Mechanization in problem solving: The effect of einstellung. *Psychological Monographs*, http://dx.doi.org/10.1037/h0093502.

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2012). Evaluating the TD model of classical conditioning. *Learning & Behavior*, *40*(3), 305–319.

Mannor, S., Menache, I., Hoze, A., & Klein, U. (2004). Dynamic abstraction in reinforcement learning via clustering. In *Proceedings of the twenty-first international conference on machine learning* (p. 71).

Marcus, G. (2018). Deep learning: A critical appraisal. arXiv preprint arXiv:1801.00631.

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, *447*(7148), 1111–1115. http://dx.doi.org/10.1038/nature05860.

Matsumoto, M., Matsumoto, K., Abe, H., & Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature Neuroscience*, *10*(5), 647–656. http://dx.doi.org/10.1038/nn1890.

McCormack, T., Bramley, N. R., Frosch, C. A., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure.. *Journal of Experimental Child Psychology*, *141*, 1–22.

Mcgovern, E. A. (2002). Autonomous discovery of temporal abstractions from interaction with an environment. *Power*.

Menache, I., Mannor, S., & Shimkin, N. (2002). Q-cut—dynamic discovery of subgoals in reinforcement learning. In *European conference on machine learning* (pp. 295–306). Springer.

Miller, G. A., Eugene, G., & Pribram, K. H. (2017). *Plans and the structure of behaviour*. Routledge.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological Review*, *126*, 292–311.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140.

Montague, P. R., Dayan, P., Nowlan, S. J., Pouget, A., & Sejnowski, T. (1993). Using aperiodic reinforcement for directed self-organization during development. In *Advances in neural information processing systems* (pp. 969–976).

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*(5), 1936–1947. http://dx.doi.org/10.1523/JNEUROSCI.16-05-01936.1996, Publisher: Society for Neuroscience.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, *31*(1), 69–89. http://dx.doi.org/10.1146/annurev.neuro.31.061307.090723.

Murdoch, D., Chen, R., & Goldberg, J. H. (2018). Place preference and vocal learning rely on distinct reinforcers in songbirds. *Scientific Reports*, *8*(1), 1–9.

Murray, E. A., O'Doherty, J. P., & Schoenbaum, G. (2007). What we know and do not know about the functions of the orbitofrontal cortex after 20 years of cross-species studies. *Journal of Neuroscience*, *27*(31), 8166–8169. http://dx.doi.org/10.1523/JNEUROSCI.1556-07.2007.

Narasimhan, K., Barzilay, R., & Jaakkola, T. (2018). Grounding language for transfer in deep reinforcement learning. *Journal of Artificial Intelligence Research*, *63*, 849–874.

Newell, A., Shaw, J. C., & Simon, H. A. (1959). Report on a general problem solving program. *Vol. 256*, In *IFIP congress* (p. 64). Pittsburgh, PA.

Newtson, D., et al. (1976). Foundations of attribution: The perception of ongoing behavior. *New Directions in Attribution Research*, *1*, 223–247.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154.

Oberst, M., & Sontag, D. (2019). Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International conference on machine learning* (pp. 4881–4890). PMLR.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*.

Ólafsdóttir, H. F., Barry, C., Saleem, A. B., Hassabis, D., & Spiers, H. J. Hippocampal place cells construct reward related sequences through unexplored space, eLife 4. http://dx.doi.org/10.7554/eLife.06063.

Ólafsdóttir, H. F., Carpenter, F., & Barry, C. (2016). Coordinated grid and place cell replay during rest. *Nature Neuroscience*, *19*(6), 792–794. http://dx.doi.org/10.1038/nn.4291.

Ouden, H. D., Daw, N., Fernández, G., Elshout, J., Rijpkema, M., Hoogman, M., et al. (2013). Dissociable effects of dopamine and serotonin on reversal learning. *Neuron*, *80*, 1090–1100.

Oyama, K., Hernádi, I., Iijima, T., & Tsutsui, K.-I. (2010). Reward prediction error coding in dorsal striatal neurons. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(34), 11447–11457. http://dx.doi.org/10.1523/JNEUROSCI.1719-10.2010.

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223–226.

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, *13*.

Pan, X., Sawa, K., Tsuda, I., Tsukada, M., & Sakagami, M. (2008). Reward prediction based on stimulus categorization in primate lateral prefrontal cortex. *Nature Neuroscience*, *11*(6), 703–712.

Pan, W.-X., Schmidt, R., Wickens, J. R., & Hyland, B. I. (2005). Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *25*(26), 6235–6242. http://dx.doi.org/10.1523/JNEUROSCI.1478-05.2005.

Parr, R., & Russell, S. (1998). Reinforcement learning with hierarchies of machines. *Advances in Neural Information Processing Systems*, 1043–1049.

Pashenkova, E., Rish, I., & Dechter, R. (1996). Value iteration and policy iteration algorithms for Markov decision problem. In *AAAI: workshop on structural issues in planning and temporal reasoning*.

Pastor-Bernier, A., & Cisek, P. (2011). Neural correlates of biased competition in premotor cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(19), 7083–7088. http://dx.doi.org/10.1523/JNEUROSCI.5681-10.2011.

Pavlov, I. P., & Anrep, G. V. (1927). *Vol. 142*, *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. london: oxford University Press.

Pearl, J. (2009). *Causality: models, reasoning and inference* (2nd ed.). USA: Cambridge University Press.

Peyrache, A., Khamassi, M., Benchenane, K., Wiener, S. I., & Battaglia, F. P. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. *Nature Neuroscience*, *12*(7), 919–926.

Pezzulo, G., van der Meer, M. A. A., Lansink, C. S., & Pennartz, C. M. A. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Sciences*, *18*(12), 647–657. http://dx.doi.org/10.1016/j.tics.2014.06.011.

Pickett, M., & Barto, A. G. (2002). Policyblocks: An algorithm for creating useful macro-actions in reinforcement learning. *Vol. 19*, In *ICML* (pp. 506–513).

Pischedda, D., Palminteri, S., & Coricelli, G. (2020). The effect of counterfactual information on outcome value coding in medial prefrontal and cingulate cortex: From an absolute to a relative neural code. *The Journal of Neuroscience*, *40*, 3268–3277.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature Neuroscience*, *16*(9), 1170–1178.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Pritzel, A., Uria, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., et al. (2017). Neural episodic control. In *International conference on machine learning* (pp. 2827–2836). PMLR.

Redgrave, P., & Gurney, K. (2006). *Vol. 7*, *The short-latency dopamine signal: A role in discovering novel actions?* (pp. 967–975). Nature Publishing Group,

Rescorla, R. A., & Wagner, A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current Research and Theory*, 64–99.

Reynolds, J. N., Hyland, B. I., & Wickens, J. R. (2001). A cellular mechanism of reward-related learning. *Nature*, *413*(6851), 67–70. http://dx.doi.org/10.1038/35092560.

Ribas-Fernandes, J. J., 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*(1), 8–23.

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*(2), 370–379.

Ritter, S., Wang, J., Kurth-Nelson, Z., & Botvinick, M. (2018a). Episodic control as meta-reinforcement learning. http://dx.doi.org/10.1101/360537, BioRxiv.

Ritter, S., Wang, J., Kurth-Nelson, Z., Jayakumar, S., Blundell, C., Pascanu, R., et al. (2018b). Been there, done that: Meta-learning with episodic recall. In *International conference on machine learning* (pp. 4354–4363). PMLR.

Roesch, M. R., Singh, T., Brown, P. L., Mullins, S. E., & Schoenbaum, G. (2009). Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(42), 13365–13376. http://dx.doi.org/10.1523/JNEUROSCI.2572-09.2009.

Rushworth, M. F., & Behrens, T. E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*(4), 389–397.

Russek, E., Momennejad, I., Botvinick, M., Gershman, S., & Daw, N. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, *13*.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337–1340.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *International conference on machine learning* (pp. 1842–1850). PMLR.

Sargolini, F., Fyhn, M., Hafting, T., McNaughton, B. L., Witter, M. P., Moser, M.-B., et al. (2006). Conjunctive representation of position, direction, and velocity in entorhinal cortex. *Science*, *312*(5774), 758–762.

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.

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2016). Prioritized experience replay. arXiv:1511.05952.

Schaul, T., & Schmidhuber, J. (2010). Metalearning. *Scholarpedia*, http://dx.doi.org/10.4249/scholarpedia.4650.

Schneider, D. W., & Logan, G. D. (2006). Hierarchical control of cognitive processes: switching tasks in sequences. *Journal of Experimental Psychology: General*, *135*(4), 623.

Schoenbaum, G., Nugent, S. L., Saddoris, M. P., & Setlow, B. (2002). Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. *Neuroreport*, *13*(6), 885–890. http://dx.doi.org/10.1097/00001756-200205070-00030.

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, *57*, 87–115. http://dx.doi.org/10.1146/annurev.psych.56.091103.070229.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, http://dx.doi.org/10.1523/jneurosci.13-03-00900.1993.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience*, *23*(1), 473–500. http://dx.doi.org/10.1146/annurev.neuro.23.1.473.

Schwartz, A., & Thrun, S. (1995). Finding structure in reinforcement learning. *Advances in Neural Information Processing Systems*, *7*, 385–392.

Sejnowski, T. J., Dayan, P., & Montague, P. R. (1995). Predictive hebbian learning. In *COLT '95, Proceedings of the eighth annual conference on computational learning theory* (pp. 15–18). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/225298.225300.

Seo, H., Barraclough, D. J., & Lee, D. (2009). Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(22), 7278–7289. http://dx.doi.org/10.1523/JNEUROSCI.1479-09.2009.

Seo, H., & Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(31), 8366–8377. http://dx.doi.org/10.1523/JNEUROSCI.2369-07.2007.

Seo, H., & Lee, D. (2009). Behavioral and neural changes after gains and losses of conditioned reinforcers. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(11), 3627–3641. http://dx.doi.org/10.1523/JNEUROSCI.4726-08.2009.

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, *321*(5890), 848–851. http://dx.doi.org/10.1126/science.1160575.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*.

Simpkins, C., & Isbell, C. (2019). Composable modular reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence, Vol. 33* (pp. 4975–4982).

Singh, S. P., Barto, A. G., & Chentanez, N. (2004). Intrinsically motivated reinforcement learning. *NIPS*.

Sirota, A., Montgomery, S., Fujisawa, S., Isomura, Y., Zugaro, M., & Buzsáki, G. (2008). Entrainment of neocortical neurons and Gamma oscillations by the hippocampal theta rhythm. *Neuron*, http://dx.doi.org/10.1016/j.neuron.2008.09.014.

Skaggs, W. E., & McNaughton, B. L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science*, *271*(5257), 1870–1873. http://dx.doi.org/10.1126/science.271.5257.1870.

Skinner, B. (1988). The operant side of behavior therapy. *Journal of Behavior Therapy and Experimental Psychiatry*, *19*(3), 171–179.

So, N.-Y., & Stuphorn, V. (2010). Supplementary eye field encodes option and action value for saccades with variable reward. *Journal of Neurophysiology*, *104*(5), 2634–2653. http://dx.doi.org/10.1152/jn.00430.2010.

Sobel, D. M., & Sommerville, J. (2010). The importance of discovery in children's causal learning from interventions. *Frontiers in Psychology*, *1*.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., et al. (2014). Optimal behavioral hierarchy. *PLoS Computational Biology*, *10*(8), Article e1003779.

Sommerville, J. A., & Woodward, A. L. (2005a). *Vol. 8, Infants' sensitivity to the causal features of means-end support sequences in action and perception* (pp. 119–145). Wiley Online Library,

Sommerville, J. A., & Woodward, A. L. (2005b). Pulling out the intentional structure of action: The relation between action processing and action production in infancy. *Cognition*, http://dx.doi.org/10.1016/j.cognition.2003.12.004.

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

Sprague, N., & Ballard, D. (2003). Multiple-goal reinforcement learning with modular sarsa(o). In *IJCAI'03, Proceedings of the 18th international joint conference on artificial intelligence* (pp. 1445–1447). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Sul, J. H., Jo, S., Lee, D., & Jung, M. W. (2011). Role of rodent secondary motor cortex in value-based action selection. *Nature Neuroscience*, *14*(9), 1202–1208. http://dx.doi.org/10.1038/nn.2881.

Sul, J. H., Kim, H., Huh, N., Lee, D., & Jung, M. W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron*, *66*(3), 449–460. http://dx.doi.org/10.1016/j.neuron.2010.03.033.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, http://dx.doi.org/10.1023/A:1022633531479.

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*. http://dx.doi.org/10.1016/b978-1-55860-141-3.50030-4.

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*(2), 135.

Sutton, R. S., & Barto, A. G. (1990). *Time-derivative models of pavlovian reinforcement*. The MIT Press.

Sutton, R., & Barto, A. (1998). Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks*, http://dx.doi.org/10.1109/tnn.1998.712192.

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999a). Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, *12*, 1057–1063.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. *Advances in Neural Information Processing Systems*, 59–65.

Thorndike, E. L. (1933). A proof of the law of effect. *Science*, *77*(1989), 173–175. http://dx.doi.org/10.1126/science.77.1989.173-a.

Tolman, E. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189—208. http://dx.doi.org/10.1037/h0061626.

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2020). Discovery of hierarchical representations for efficient planning. *PLoS Computational Biology*, *16*(4), Article e1007594.

Tremblay, L., & Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, *398*(6729), 704–708.

Tsividis, P. A., Pouncy, T., Xu, J. L., Tenenbaum, J. B., & Gershman, S. J. (2017). Human learning in Atari. In *2017 AAAI spring symposium series*.

Tsutsui, K. I., Grabenhorst, F., Kobayashi, S., & Schultz, W. (2016). A dynamic code for economic object valuation in prefrontal cortex neurons. *Nature Communications*, http://dx.doi.org/10.1038/ncomms12554.

Vikbladh, O., Shohamy, D., & Daw, N. D. (2017). Episodic contributions to model - based reinforcement learning. In *Cognitive computational neuroscience conference*.

Wallis, J. D., & Kennerley, S. W. (2010). Heterogeneous reward signals in prefrontal cortex. *Current Opinion in Neurobiology*, *20*(2), 191–198. http://dx.doi.org/10.1016/j.conb.2010.02.009.

Wallis, J. D., & Miller, E. K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*, *18*(7), 2069–2081.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, http://dx.doi.org/10.1038/s41593-018-0147-8.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., et al. (2016). Learning to reinforcement learn. arXiv preprint arXiv:1611.05763.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, http://dx.doi.org/10.1007/bf00992698.

Wayne, G., Hung, C.-C., Amos, D., Mirza, M., Ahuja, A., Grabska-Barwinska, A., et al. (2018). Unsupervised predictive memory in a goal-directed agent. arXiv preprint arXiv:1803.10760.

Womelsdorf, T., Vinck, M., Stan Leung, L., & Everling, S. (2010). Selective theta-synchronization of choice-relevant information subserves goal-directed behavior. *Frontiers in Human Neuroscience*, http://dx.doi.org/10.3389/fnhum.2010.00210.

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

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, http://dx.doi.org/10.1037/0033-2909.127.1.3.