Reinforcement Learning and its Connections with Neuroscience and Psychology

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

[a]*Cognitive Neuroscience Lab, BITS Pilani K K Birla Goa Campus, NH-17B, Zuarinagar, Goa, India 403726*
[b]*Dept. of Electrical and Electronics Engineering, BITS Pilani K K Birla Goa Campus, NH-17B, Zuarinagar, Goa, India 403726*
[c]*Dept. of Biological Sciences, BITS Pilani K K Birla Goa Campus, NH-17B, Zuarinagar, Goa, India 403726*

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 in the area to produce such results, many core ideas in reinforcement learning are inspired by phenomena of animal learning, psychology and neuroscience. In this paper, we comprehensively review a number of findings in neuroscience and psychology that provide evidence for the plausibility of reinforcement learning being a promising model for phenomena in human learning, decision making and behavior. We do so by a) exploring neuroscientific evidence for various classes of RL algorithms along with their building blocks, and b) mapping specific RL ideas to findings in neuroscience and psychology. Finally, we discuss the implications of these findings and their role in advancing research in both AI and brain science.

Keywords: reinforcement learning, neuroscience, psychology

---

*Corresponding authors
  Email addresses: f20170371@goa.bits-pilani.ac.in (Ajay Subramanian),
f20170472@goa.bits-pilani.ac.in (Sharad Chitlangia), veeky@goa.bits-pilani.ac.in
(Veeky Baths)

1. Introduction

Reinforcement learning methods have been very successful at exceeding human-level performance on a variety of complex sequential tasks such as Atari [1], Go [2], Poker [3] and Dota-2 [4]. 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. Though there have been multiple works reviewing the implications of reinforcement learning in neuroscience [5, 6, 7]. In 2012, Lee et al. [5] reviewed several works reporting evidence of core reinforcement learning ideas being implemented within the neural networks in the brain. According to them, commonly used building blocks of reinforcement learning such as value functions, temporal difference learning and reward prediction errors (RPEs) have been validated by findings in neuroscience research, thus making reinforcement learning a promising candidate for computationally modelling phenomena in 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 areas of research like distributional RL [8], meta RL [9, 10], and model-based RL [11] have emerged, which has motivated work that researches evidence for similar phenomena in neuroscience and psychology. In this review, we have incorporated these works too, in order to provide a well rounded and up-to-date review of the neural and psychological correlates for modern reinforcement learning algorithms.

In presenting a comprehensive review of the same, we employ the following structure. First, 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, we dis-

cuss 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 psychology that are analogous to these concepts and evidence that they are implemented in the brain. Following this, we discuss some modern RL algorithms and their neural and psychological 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 concise yet detailed table mapping specific reinforcement learning concepts and corresponding work validating their involvement in animal learning and decision making. Finally, we present an opinionated 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 research might hold key insight to, and vice versa.

Unlike previous works that approach reviewing this intersection from a neuroscience angle, we present a fresh perspective keeping two principles in mind.

- **RL-first approach:** Topics are categorized based on concepts in reinforcement learning. While discussing them, we provide the RL concept first followed by its validation in brain science. This is unlike much previous work which presents discussion from the perspective of neuroscientific phenomena.

- **Modularity:** Decision making can be broken down conceptually into several capabilities. For example: planning, hierarchy, valuing options, learning to learn etc. The distinct sub-fields of reinforcement learning research (model-based RL, hierarchical RL, meta RL etc.) capture this modularity very well and so we discuss them each separately.

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

3

that maximizes a reward [12]. Figure 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 agents interaction with its environment comprises several such transitions. While considering these transitions, we make an explicit assumption that the future is independent of the past given the present. In other words, the next state in our process 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.

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

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.

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.
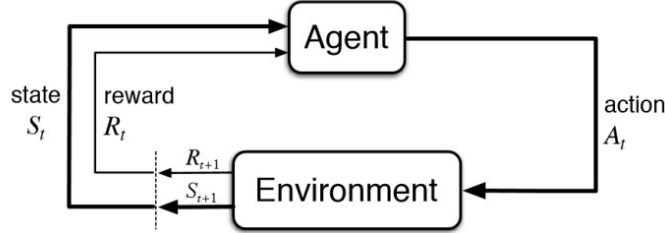
4

Figure 1: The classical RL framework [12]. The Agent selects actions A_t to update its state S_t so as to receive maximum reward R_{t+1} from the environment over the long term. [Source: [12]]

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

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 we execute a specific policy *after* a specific action is taken at the state.

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

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 [13] and finds application in various modern reinforcement learning algorithms. A common set of algorithms for optimizing the value function are the **dynamic programming methods**. These methods update value functions by bootstrapping value functions from other states [14, 15]. Examples of DP methods are Q-learning [16] and SARSA [12]. The optimization process involves update their value function by ascending the gradient in the direction of the difference between a **target** value and a currently obtained reward, 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 reward 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 [17] and are covered in a future section.

Now that we have given a brief background on some of the important core reinforcement learning concepts, we will move to exploring the neural and psychological correlates for some of these fundamental building blocks that make up classical and modern reinforcement learning algorithms.

3. Building Blocks of Reinforcement Learning and their Neural 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 the neural inspiration and evidence for these ideas. Most of these ideas have already been reviewed in much detail by Lee et al. [5] 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, refer [5].

3.1. Value Functions

As discussed in the previous section, a value function is a measure of reward expectation. Neural signals containing information about reward expectancy have been shown to exist in many areas of the brain [18, 19, 20]. Resembling the types of value function types prevalent in RL algorithms, evidence has been found for the brain too using both state values and action value encoding. Action value functions 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 value**. [21, 22].

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 [23, 24, 25, 22, 26] and might apply to distinct reward signals as for instance, demonstrated by Murdoch et al. [27] in the case of songbirds. 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 [21]. Neurons encoding value in the supplemental motor area signaled desirable eye movements (spatial location of targets) while considering the choice while those in the primate orbitofrontal cortex were associated with the juice flavors themselves (targets) [28, 29]. The function of distinct reward signals in the brain is analogous to a highly distributed version of credit assignment via backpropagation in artificial neural networks.

Neural signals for chosen values are are also distributed in multiple brain areas [21, 30, 31, 32, 22]. Also, some brain regions encode the difference of values between two alternative actions to determine likelihood of taking an action over another [33, 24, 25, 23, 22]. 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 [34].

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 $G_t = R_{t+1} + R_{t+2} + R_{t+3} + \ldots$ 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.

$$RPE(t) = G_t - V_\pi(S_t) \tag{4}$$

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 to improve value estimates. Such RPE signals have been identified in midbrain dopaminergic neurons [35] and many other areas such as the orbitofrontal cortex, lateral habenula and angular cingulate cortex [36, 37, 38, 39, 32, 31, 40].

Action value functions in the brain are believed to be updated and stored at the synapses between cortical axons and striatal spiny dendrites [41, 19, 42, 43]. RPEs are input to these synapses through terminals of dopaminergic neurons [44, 35, 45, 46] and value functions are updated using spike-time dependent plasticity. Ongoing work is exploring and testing the involvement of eligibility traces for credit assignment in these value updates. A popular idea in this direction is neuromodulated STDP which combines classical STDP with eligibility traces to add external reinforcement from reward signals [47, 48].

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 [49, 50, 51, 52].

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 [53] to strengthen the connection between a state and its corresponding reward. This technique however, doesn't correspond to any observation in neuroscience literature and is not consistent with profiles of dopamine neuron activity [54]. The second method is to use eligibility traces which are short term memory signals that assign state responsibility for a reward [12]. 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 [55, 33, 23, 56, 31, 32, 57, 58]. 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. [55, 33, 31, 59, 32, 60]. Additionally, neurons in orbitofrontal cortex are believed to encode relationships between actions and their corresponding outcomes [55, 23, 31, 32]. This observation could inspire future work in reinforcement learning research towards a solution to the credit assignment problem.

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 [62]. Along with mapping
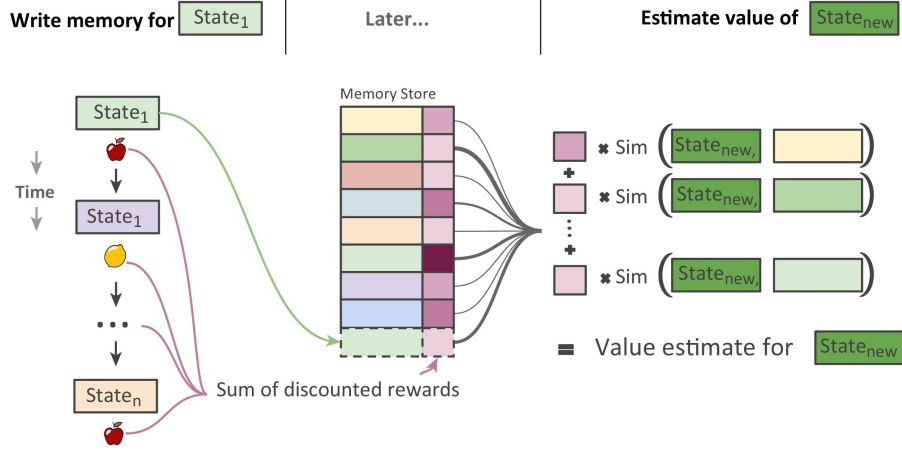
Figure 2: Episodic reinforcement learning [61]. Memories of previously experienced states and corresponding rewards are stored, and value of new, unseen states is computed based on degree of similarity to new states. This resembles instance-based models of memory. [Source: [61]]

trajectories, these cells spontaneously recap previously experienced trajectories [63, 64, 65, 66]. They can also explore new spatial trajectories which haven't been experienced before [67, 68]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 [69, 70] which is used to generalize learned knowledge. Biological replay mechanisms have been recorded in both the entorhinal cortices [71] and visual cortices [72]. 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 [1, 73].

As discussed previously, deep RL methods that excel in performance in various tasks struggle with achieving sample efficiency similar to that of humans [74, 75]. Experience replay is a popular component in reinforcement learning

algorithms which enables them learn tasks with fewer environment interactions by storing and reusing previously experienced transitions to update the policy. However, experience replay mechanisms employed in deep RL are still unable to mimic their biological counterparts. For instance, Liu et al. [76] 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 [77, 78, 79] uses such experience as an inductive bias to learn future tasks. One approach to do so, is to use a similarity network that reuses values for states that have already been learned, thus reducing the time to learn values [80]. This is similar to instance-based models of memory in where specific stored information from past experience is used for decision making in new situations. [81, 79, 82, 83].

4. Psychological and Neural Evidence for RL Algorithms

Having covered the building blocks that most commonly make up RL algorithms, we dive deep into various types of reinforcement learning algorithms along with work suggesting that they might be promising models for 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:

$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(S_t) \tag{5}$$

The earliest known use of temporal difference learning in Artificial Intelligence research dates back to 1959 when Samuel [84] demonstrated its usage for a checkers-playing program. Sutton [17] developed the first theoretical formulation of TD learning and showed that they were more efficient and yet more accurate than conventional supervised learning methods. Following this, work in computational neuroscience suggested that the firing of dopamine neurons signalled a reward prediction error [85]. Later work [86] showed that the TD model allows a formulation of expectations through value functions to influence synaptic changes, via a Hebbian learning framework. Perhaps the most commonly related neuroscience phenomenon to the TD reward prediction error was given by Schultz [87] as the reward prediction error hypothesis of dopamine neuron activity.

From the extensive experiments conducted by [87], a major breakthrough in relating TD methods to actual biological phenomena was made by Montague and colleagues [53] when they related fluctuation levels in dopamine delivery, from the VTA to cortical and subcortical target neuronal structures, to TD reward prediction errors.

The TD error formulation is a very specific case of the more general TD(λ) later proposed by Sutton [12] which accounts for eligibility traces. Two new terms are introduced to account for it, the weight vector w_t and the eligibility trace z_t, modeled as:

12

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

$$w_t = w_{t-1} + \alpha \delta_{t-1} z_{t-1} \tag{7}$$

$$z_t = \gamma \lambda z_{t-1} + \gamma V(S_{t-1}, w_{t-1}) \tag{8}$$

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 . On the other hand, when $\lambda = 0$, it transforms into the TD formulation discussed earlier, where only the previous state is given credit. [12] showed that this TD(λ) formulation is the same formulation as TD model of classical conditioning as used in the framework proposed by Montague and others [53] to verify the results of TD learning. Thus, the TD(λ) formulation combined RPEs and eligibility traces into a single framework.

As we shall discuss in a future section, recent experiments have suggested that overlaying ideas from distributional reinforcement learning onto the current TD-RPE framework might be better at explaining recordings of reward prediction errors in the brain [88].

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 and so on. Model-based reinforcement learning seeks to mimic these capabilities and is a promising area both in RL [2] (Figure 3) and as a computational model for biological learning [89].

The animal learning psychology community in the early 20th century saw a divide between Thorndike's Law of Effect [90] and Tolman's Cognitive Maps [91]. 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,
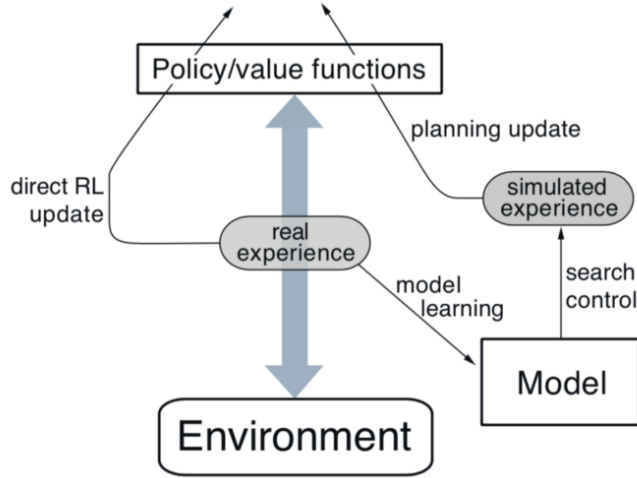
Figure 3: Schematic of the Dyna model-based reinforcement learning framework. Unlike model-free methods, these use both real and simulated experience to update their policy and value functions. In order to generate realistic simulated trajectories, a model learning process occurs simultaneously. [Source: [11]]

Tolman stated that learning can still happen in the case that a reward is not immediately received strengthening the argument on 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 psychological evidences for the two major types of learning systems concerned with action valuation in our brain, model free and model based learning.

Although we've seen numerous advancements in finding neural correlates for model free reinforcement learning [92, 93, 94], the last 2 decades have seen proving the existence of a model based system especially in a combined setting with the model free learning system [23, 95, 96, 97]. Typical experiments involve a multi-staged decision making task while simultaneous recording of 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 [30]) in the ventro-

14

medial prefrontal cortex [98, 95], ventral striatum [95] and a model based RL behavior in the lateral pre-frontal cortex [96, 99, 100]. Combined social information and reward history can also be traced to the different regions of the anterior cingulate cortex [101].

Model-based learning systems involve building mental models through experience, followed by search of desired action sequences using this model. Humans are known to develop *habits* i.e. fast decisions over time [99]. Few works [74] suggest that human learning systems arbitrate to tradeoff between flexibility and speed. By simulating possibilities offline through the use of the model built by the model-based learning system, model based choices can be made more efficient.

Human neural systems are known to use information from both model-free and model-based sources [102, 103, 104]. There is extant evidence that model-based algorithms are implemented in biological systems. For instance, Glascher et al. [104] 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 [105, 106, 107, 108].

4.3. Distributional Reinforcement Learning

In classical temporal difference learning, as discussed earlier, the value of a state is represented as an expectation of future rewards starting from that state, and this value is updated using values of future states as shown in Equation 5. Since the reward prediction error uses a single expectation value to anticipate a future reward for a given state, it squeezes the entire reward distribution into a single number and so, might not capture multimodal reward distributions. Much
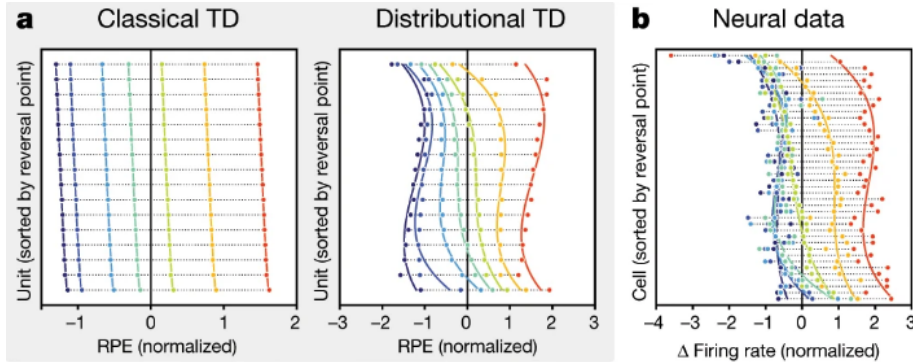
15

Figure 4: Comparison of distributional TD and classical TD RPEs. **a**. horizontal bars denote individual neurons. x-axis measures RPE value. For classical TD, neurons differ very little in their expected rewards while in distributional, there is a lot of variance; some neurons are optimistic while others are pessimistic. **b**. RPE signals obtained from dopamine neurons of the mouse brain show high variance for which distributional TD was a better fit than classical TD. [Source: [88]]

recent work has been done on developing a distributional framework for RL that maintains an expected reward **distribution** rather than a single average value over future rewards [8], which enables identification of bimodal and multimodal properties in the reward distribution. The formulation of this idea, termed as **distributional reinforcement learning**, involves updating a random variable $Z(s, a)$ whose expectation is the value $Q(s, a)$.

$$Z(s_t, a_t) = R_{t+1} + \gamma Z(s_{t+1}, a_{t+1}) \tag{9}$$

Past work has provided evidence for distributional coding in the brain for non-RL domains [109]. Moreover, distributional reinforcement learning has been shown to be biologically plausible [110, 111]. Recently, Dabney et al. [88] carried out single-unit recordings of the ventral tegmental area 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 for a single reward which goes to show that cells display varying degrees of optimism and pessimism for achieving a particular goal. Through extensive experiments,

16

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. Figure 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.
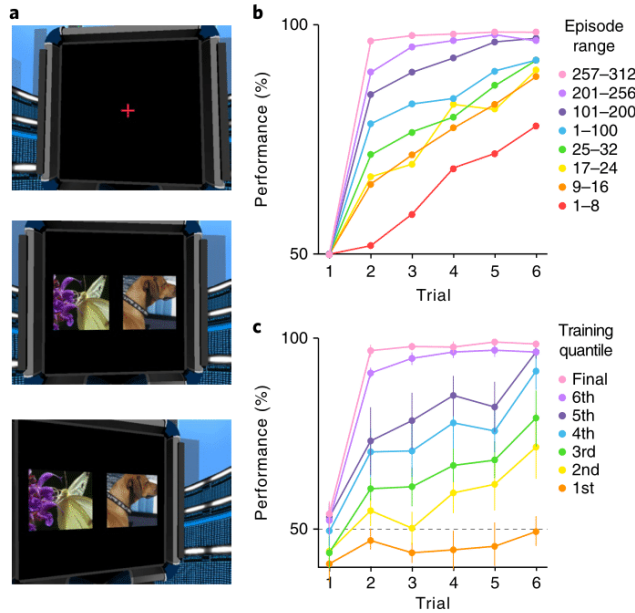

Figure 5: Comparison between meta RL's and Harlow's experimental results. **a**. Inputs for simulation experiments showing fixation cross (top), initial stimulus (middle), and outcome of saccade (bottom). **b**. Performance reward at each trial of Harlow's monkey experiment [112]. **c**. Simulation performance at different phases of training. Improvement trend over time resembles that from Harlow. [Source: [9]]

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 [74, 113].

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 new task learning is called meta-learning [114]. The original idea of "learning to learn" is often attributed to Harlow's 1949 work [112] 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 modelling 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 [61].

Very recently, this core idea has been applied to reinforcement learning to accelerate the learning process. 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 [115, 10] (Figure 5). Wang and colleagues [9] also noticed that such meta-learning methods formed a part

18

of dopaminergic reward-based learning in the prefrontal cortex of the brain. Much recent work on the prefrontal cortex (PFC) [98, 116, 21, 117, 55] suggests that humans do more than just learn an abstraction of only tweaking synaptic weights for learning. The PFC encodes latent representations on recent rewards and choices, and some sectors also encode the expected values of actions, objects, and states.

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 [118], Ritter et al [119] developed a framework for how such episodic memory functions can strategically reuse information about previously learned tasks and thereby improve learning efficiency [120, 121].This work also evidences recent interactions between meta model-based control and episodic memory in human learning [122].

4.5. Causal Reinforcement Learning

The ability of humans to reason about cause and effect relationships in not unknown. The field of Causal Inference and Reasoning looks at studying this paradigm in great detail. In the context of Causal Inference, a reinforcement learning agent taking actions in an environment can be considered as performing interventions in the world. But since in the model-free setting, we are merely operating by observing a world and not by maintaining a model of the world, there is a limited possibility of discernment of causal relationships. In particular, a model-free agent might only be able to determine the causal effects of its action but not actually reason, for instance through counterfactual actions, questions concerning, "What could've been?".

Being able to construct intuitive models of the world is a hallmark of human intelligence because it helps us in planning and reasoning among other aspects

19

of our daily interaction with the world. In particular, models encoding the causal structure of the world serve as strong priors for planning and reasoning about the worldly tasks [74]. These causal models are especially helpful in contemplating scenarios for better planning i.e., reasoning through counterfactual actions. In formal terms of Causal Inference, with a structural causal model (SCM) of the environment, counterfactual trajectories can be simulated. Recent works [123, 124] focus on this by modelling 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 [124] as well.

The concepts of Causality are not only applicable to planning but also in hidden state inference which is used in scenarios where there is partial observability. Gershman and Daw [81] expands on this in their Latent Cause Theory. Humans deal with partial observability on a daily basis, but what makes it easy for them is to reason through the world by building an internal model. This internal model is grounded in concepts observed through partitioning observations in a well-organized manner i.e., structure learning. Humans are then able to perform inference over these latent causes to reason if two different signals are coming from the same latent cause [125]. Recent work supports this by building a bayesian non-parametric prior over latent causes [126].

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 that was developed was temporal abstraction [127, 128, 129, 130] which suggested learning temporally extended actions that were composed of primitive actions. These temporally extended actions are commonly referred to as **options** [130]. For example, the option "walk to-

wards the door" would be composed of several primitive actions including motor movements and maintaining balance.

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 [131, 132, 133, 134, 135, 136, 137]. Even in neuroscience, existing literature accounts for the prefrontal cortex being largely responsible for hierarchical behavior [138, 139, 140, 141, 142]. 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 that was first formalized by Lashley in 1951 [131]. The concept of task representation [143, 144, 145] 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 [146, 147]. 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 [148, 149, 150].

The strongest resemblance to HRL is found in the production-system based theories of cognition, specially ACT-R [134] and Soar [151]. 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 [152] 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 psychological theories even in these downsides. A direct analog to the negative transfer problem has been observed in HRL [153].

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 [154, 155, 156]. 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 [157, 158, 159]. Existing work in psychology provides empirical evidence that children identify causal representations that they then integrate into a large causal model [160, 161, 162, 163]. More recent work in HRL uses task agnostic approaches to discover options, by using intrinsic rewards for exploration [164, 165]. 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 [166, 167, 168].
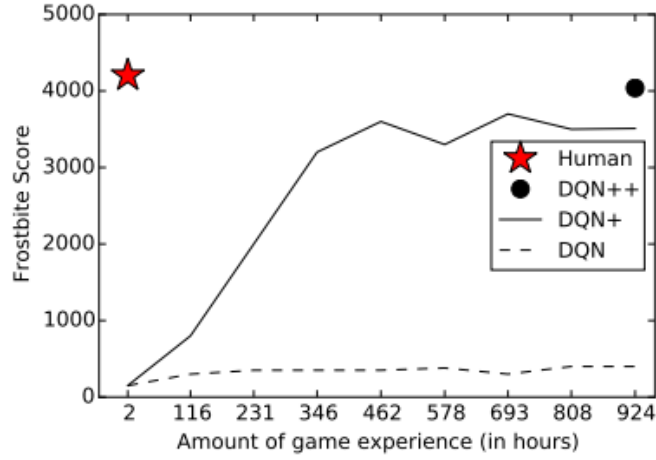
Figure 6: Comparison of humans with state of the art deep reinforcement learning methods on the Atari game 'Frostbite'. [Source: [74]]

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 [19], prefrontal cortex [20] and other areas [18]
action value function	chosen value during decision making [21, 22]
multi-task learning [169]	distributed reward signals [23, 24, 25, 26, 22, 27, 21]
actor-critic baseline [34]	relative values [33, 24, 25, 23, 22]

23

Continuation of Table 1	
Reinforcement learning concept	**Corresponding phenomena in neuroscience and psychology / Brain area responsible**
reward prediction error / TD error	RPE signals in dopaminergic neurons (VTA) [35], orbitofrontal cortex, lateral habenula, cingulate cortex, etc. [36, 37, 38, 39, 32, 31, 40]
credit assignment problem	orbitofrontal cortex [49, 50, 51, 52]
eligibility traces [12]	prefrontal cortex, striatum, frontal cortex [55, 33, 23, 56, 31, 32, 57, 58]
experience replay [1, 73]	hippocampal place cells [62, 63, 64, 65, 66], entorhinal cortices [71], visual cortices [72]
episodic memory [80]	instance-based models of memory [81, 79, 82, 83]
temporal difference (TD) learning [84, 17]	reward prediction error hypothesis of dopamine neuron activity [87]
TD(λ) [12]	TD model of classical conditioning [53]
distributional TD learning [8]	distributional coding in non-RL domains [109, 110, 111], value coding in VTA of mice [88]
meta reinforcement learning [114, 115, 10]	learning to learn [112], fast and slow learning [61], prefrontal cortex [9, 98, 116, 21, 117, 55]
episodic meta RL [119]	cerebral cortex [118, 120, 121], interaction between meta model-based control and episodic memory in human learning [122]
hierarchy [127, 128, 129, 130]	goal-directed behavior [131, 132, 133, 134, 135, 136, 137], prefrontal cortex [138, 139, 140, 141, 142], higher level representation of task context [131], mental hierarchical organization [146, 147], gradual integration of skills [148, 149, 150], production-system based theories of cognition (ACT-R, Soar) [134]

Continuation of Table 1	
Reinforcement learning concept	**Corresponding phenomena in neuroscience and psychology / Brain area responsible**
options [130]	task representation [143, 144, 145]
incompatibility between learning problem and temporally abstract actions [153]	negative transfer problem [152]
option discovery	humans are sensitive to repeating sequences of events
graph partitioning to identify bottleneck states [157, 158, 159]	children integrate causal representations into a causal model [160, 161, 162, 163]
intrinsic motivation [164, 165]	intrinsic rewards in dopamine driven learning [166, 167, 168]
End of Table	

5. Discussion and Conclusions

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 via a mapping between RL algorithms and evidence corresponding to them in neuroscience and psychology. By doing so, we hope to inspire future research that bridges the gap between the fields of RL, more generally machine learning, and brain science.

Even though several similarities exist between the components used in RL algorithms and those we believe to make up human learning ability, humans

capabilities are still a far reach from those of modern RL. Humans can learning new tasks with very little data (Figure 6). Moreover, they can perform variations of a learned task (with different goals, handicaps etc.) without having to re-learn from scratch [74]. 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. Neuroscience research has also yielded concepts such as memory neurons and distributed rewards which share no counterpart in RL algorithms. Moreover, studying the nature of representations of stimuli and concepts in the brain can inspire ideas for representations in deep RL and more generally deep learning. Additionally, humans are capable of performing tasks which are very different from those used as benchmark tasks in RL research. Exploring experiments conducted in psychology could open doors for new, challenging RL tasks. A simulated version of Harlow's monkey experiments being used as a task for meta RL algorithms is a fine example. Another aspect of human intelligence that evades present-day RL is our ability to effectively make use of multiple modalities (vision, language, interaction, etc.) during learning. Equipping RL agents with similar perception abilities might be a promising avenue for future work.

Though we have covered a number of important RL algorithms and concepts along with their biological correlates, there are a number of relevant non-RL ideas that might also be useful for RL research. Some of them are discussed below.

- **Inductive bias**: Humans come into a learning task with a lot of previous experience and conceptual knowledge. This might be a contributing factor to our amazing data efficiency. On the other hand, RL algorithms almost always learn from scratch. This disparity encourages future research on transfer learning, meta learning and representation learning in RL.

- **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 [170]. 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. [171, 172].

- **Social learning**: A key aspect of human learning is our ability to learn via social interaction. For instance, communicating with other humans forces us to develop clearer representations. Moreover, social learning offers huge advantages for tackling exploration problems and also transfer learning from other people. Ideas from social psychology could therefore offer assistance in solving the prevalent credit assignment and exploration in multi-agent RL research.

6. Acknowledgements

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing Atari with Deep Reinforcement Learning, arXiv:1312.5602 [cs]ArXiv: 1312.5602.
URL http://arxiv.org/abs/1312.5602

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Naturedoi:10.1038/nature16961.

[3] J. Heinrich, D. Silver, Smooth UCT search in computer poker, in: IJ-CAI International Joint Conference on Artificial Intelligence, 2015, iSSN: 10450823.

[4] OpenAI, :, C. Berner, G. Brockman, B. Chan, V. Cheung, P. D\c ebiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Jzefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. O. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, S. Zhang, Dota 2 with Large Scale Deep Reinforcement Learning_eprint: 1912.06680.
URL http://arxiv.org/abs/1912.06680

[5] D. Lee, H. Seo, M. W. Jung, Neural Basis of Reinforcement Learning and Decision Making, Annual Review of Neurosciencedoi:10.1146/annurev-neuro-062111-150512.

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

[7] M. Botvinick, J. X. Wang, W. Dabney, K. J. Miller, Z. Kurth-Nelson, Deep reinforcement learning and its neuroscientific implications, Neuron.

[8] M. G. Bellemare, W. Dabney, R. Munos, A distributional perspective on reinforcement learning, in: 34th International Conference on Machine Learning, ICML 2017, 2017, _eprint: 1707.06887.

[9] J. X. Wang, Z. Kurth-Nelson, D. Kumaran, D. Tirumala, H. Soyer, J. Z. Leibo, D. Hassabis, M. Botvinick, Prefrontal cortex as a meta-reinforcement learning system, Nature Neurosciencedoi:10.1038/s41593-018-0147-8.

[10] Y. Duan, J. Schulman, X. Chen, P. L. Bartlett, I. Sutskever, P. Abbeel, RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning (2016) 1–14_eprint: 1611.02779.
URL http://arxiv.org/abs/1611.02779

[11] R. S. Sutton, Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming, in: Machine Learning Proceedings 1990, 1990. `doi:10.1016/b978-1-55860-141-3.50030-4`.

[12] R. Sutton, A. Barto, Reinforcement Learning: An Introduction, IEEE Transactions on Neural Networks`doi:10.1109/tnn.1998.712192`.

[13] E. Pashenkova, I. Rish, R. Dechter, Value iteration and Policy iteration algorithms for Markov decision problem, Relation.

[14] R. Bellman, The Theory of Dynamic Programming, Bulletin of the American Mathematical Society`doi:10.1090/S0002-9904-1954-09848-8`.

[15] L. Buoniu, R. Babuka, B. De Schutter, D. Ernst, Reinforcement learning and dynamic programming using function approximators, 2010, publication Title: Reinforcement Learning and Dynamic Programming Using Function Approximators. `doi:10.1201/9781439821091`.

[16] C. J. C. H. Watkins, P. Dayan, Q-learning, Machine Learning`doi:10.1007/bf00992698`.

[17] R. S. Sutton, Learning to Predict by the Methods of Temporal Differences, Machine Learning`doi:10.1023/A:1022633531479`.

[18] W. Schultz, A. Dickinson, Neuronal Coding of Prediction Errors, Annual Review of Neuroscience 23 (1) (2000) 473–500, _eprint: https://doi.org/10.1146/annurev.neuro.23.1.473. `doi:10.1146/annurev.neuro.23.1.473`.
URL `https://doi.org/10.1146/annurev.neuro.23.1.473`

[19] O. Hikosaka, K. Nakamura, H. Nakahara, Basal ganglia orient eyes to reward, Journal of Neurophysiology 95 (2) (2006) 567–584. `doi:10.1152/jn.00458.2005`.

[20] J. D. Wallis, S. W. Kennerley, Heterogeneous reward signals in pre-frontal cortex, Current Opinion in Neurobiology 20 (2) (2010) 191–198. `doi:10.1016/j.conb.2010.02.009`.
URL `http://www.sciencedirect.com/science/article/pii/S0959438810000310`

[21] C. Padoa-Schioppa, J. A. Assad, Neurons in the orbitofrontal cortex en-code economic value, Nature`doi:10.1038/nature04676`.

[22] X. Cai, S. Kim, D. Lee, Heterogeneous Coding of Temporally Discounted Values in the Dorsal and Ventral Striatum during Intertemporal Choice, Neuron 69 (1) (2011) 170–182. `doi:10.1016/j.neuron.2010.11.041`.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3034314/`

[23] H. Seo, D. J. Barraclough, D. Lee, Lateral intraparietal cortex and rein-forcement learning during a mixed-strategy game, The Journal of Neu-roscience: The Official Journal of the Society for Neuroscience 29 (22) (2009) 7278–7289. `doi:10.1523/JNEUROSCI.1479-09.2009`.

[24] S. Kim, J. Hwang, D. Lee, Prefrontal coding of temporally discounted values during intertemporal choice, Neuron 59 (1) (2008) 161–172. `doi:10.1016/j.neuron.2008.05.010`.

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

[26] N.-Y. So, V. Stuphorn, Supplementary eye field encodes option and ac-tion value for saccades with variable reward, Journal of Neurophysiology 104 (5) (2010) 2634–2653. `doi:10.1152/jn.00430.2010`.

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

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

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

[30] B. Lau, P. W. Glimcher, DYNAMIC RESPONSE-BY-RESPONSE MODELS OF MATCHING BEHAVIOR IN RHESUS MONKEYS, Journal of the Experimental Analysis of Behavior`doi:10.1901/jeab.2005.110-04`.

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

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

[33] H. Seo, D. Lee, 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) (2009) 3627–3641. `doi:10.1523/JNEUROSCI.4726-08.2009`.

[34] R. S. Sutton, D. McAllester, S. Singh, Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, Advances in neural information processing systems 12 (1999) 1057–1063.

[35] W. Schultz, Behavioral theories and the neurophysiology of reward, Annual Review of Psychology 57 (2006) 87–115. `doi:10.1146/annurev.psych.56.091103.070229`.

[36] M. Matsumoto, O. Hikosaka, Lateral habenula as a source of negative reward signals in dopamine neurons, Nature 447 (7148) (2007) 1111–1115. `doi:10.1038/nature05860`.

[37] S. Hong, O. Hikosaka, The globus pallidus sends reward-related signals to the lateral habenula, Neuron 60 (4) (2008) 720–729. `doi:10.1016/j.neuron.2008.09.035`.

[38] M. Matsumoto, K. Matsumoto, H. Abe, K. Tanaka, Medial prefrontal cell activity signaling prediction errors of action values, Nature Neuroscience 10 (5) (2007) 647–656. `doi:10.1038/nn1890`.

[39] H. Seo, D. Lee, 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) (2007) 8366–8377. `doi:10.1523/JNEUROSCI.2369-07.2007`.

[40] K. Oyama, I. Herndi, T. Iijima, K.-I. Tsutsui, Reward prediction error coding in dorsal striatal neurons, The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 30 (34) (2010) 11447–11457. `doi:10.1523/JNEUROSCI.1719-10.2010`.

[41] J. N. Reynolds, B. I. Hyland, J. R. Wickens, A cellular mechanism of reward-related learning, Nature 413 (6851) (2001) 67–70. `doi:10.1038/35092560`.

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

[43] S. Hong, O. Hikosaka, Dopamine-mediated learning and switching in cortico-striatal circuit explain behavioral changes in reinforcement learning, Frontiers in Behavioral Neuroscience 5 (2011) 15. `doi:10.3389/fnbeh.2011.00015`.

[44] A. I. Levey, S. M. Hersch, D. B. Rye, R. K. Sunahara, H. B. Niznik, C. A. Kitt, D. L. Price, R. Maggio, M. R. Brann, B. J. Ciliax, 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) (1993) 8861–8865. doi:10.1073/pnas.90.19.8861.

[45] S. N. Haber, J. L. Fudge, N. R. McFarland, 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) (2000) 2369–2382.

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

[47] W. Shen, M. Flajolet, P. Greengard, D. J. Surmeier, Dichotomous dopaminergic control of striatal synaptic plasticity, Science (New York, N.Y.) 321 (5890) (2008) 848–851. doi:10.1126/science.1160575.

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

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

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

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

[52] E. A. Murray, J. P. O'Doherty, G. Schoenbaum, 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) (2007) 8166–8169. `doi:10.1523/JNEUROSCI.1556-07.2007`.
URL `http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.1556-07.2007`

[53] P. R. Montague, P. Dayan, T. J. Sejnowski, A framework for mesencephalic dopamine systems based on predictive Hebbian learning, Journal of Neuroscience 16 (5) (1996) 1936–1947, publisher: Society for Neuroscience. `doi:10.1523/JNEUROSCI.16-05-01936.1996`.
URL `https://www.jneurosci.org/content/16/5/1936`

[54] W.-X. Pan, R. Schmidt, J. R. Wickens, B. I. Hyland, 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) (2005) 6235–6242. `doi:10.1523/JNEUROSCI.1478-05.2005`.

[55] D. J. Barraclough, M. L. Conroy, D. Lee, Prefrontal cortex and decision making in a mixed-strategy game, Nature Neuroscience`doi:10.1038/nn1209`.

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

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

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

[59] M. R. Roesch, T. Singh, P. L. Brown, S. E. Mullins, G. Schoenbaum, 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) (2009) 13365–13376. doi:10.1523/JNEUROSCI.2572-09.2009.

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

[61] M. Botvinick, S. Ritter, J. X. Wang, Z. Kurth-Nelson, C. Blundell, D. Hassabis, Reinforcement Learning, Fast and Slow, Trends in Cognitive Sciences 23 (5) (2019) 408–422, publisher: Elsevier. doi:10.1016/j.tics.2019.02.006.
URL https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(19)30061-0

[62] E. I. Moser, E. Kropff, M.-B. Moser, Place Cells, Grid Cells, and the Brain's Spatial Representation System, Annual Review of Neuroscience 31 (1) (2008) 69–89, _eprint: https://doi.org/10.1146/annurev.neuro.31.061307.090723. doi:10.1146/annurev.neuro.31.061307.090723.
URL https://doi.org/10.1146/annurev.neuro.31.061307.090723

[63] K. Diba, G. Buzski, Forward and reverse hippocampal place-cell sequences during ripples, Nature neuroscience 10 (10) (2007) 1241–1242. doi:10.1038/nn1961.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2039924/

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

[65] K. Louie, M. A. Wilson, Temporally structured replay of awake hippocam-

pal ensemble activity during rapid eye movement sleep, Neuron 29 (1) (2001) 145–156. `doi:10.1016/s0896-6273(01)00186-6`.

[66] W. E. Skaggs, B. L. McNaughton, Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience, Science (New York, N.Y.) 271 (5257) (1996) 1870–1873. `doi:10.1126/science.271.5257.1870`.

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

[68] H. F. Iafsdttir, C. Barry, A. B. Saleem, D. Hassabis, H. J. Spiers, Hippocampal place cells construct reward related sequences through unexplored space, eLife 4. `doi:10.7554/eLife.06063`.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4479790/`

[69] D. J. Foster, Replay Comes of Age, Annual Review of Neuroscience 40 (1) (2017) 581–602, _eprint: https://doi.org/10.1146/annurev-neuro-072116-031538. `doi:10.1146/annurev-neuro-072116-031538`.
URL `https://doi.org/10.1146/annurev-neuro-072116-031538`

[70] G. Pezzulo, M. A. A. van der Meer, C. S. Lansink, C. M. A. Pennartz, Internally generated sequences in learning and executing goal-directed behavior, Trends in Cognitive Sciences 18 (12) (2014) 647–657. `doi:10.1016/j.tics.2014.06.011`.
URL `http://www.sciencedirect.com/science/article/pii/S1364661314001570`

[71] H. F. Iafsdttir, F. Carpenter, C. Barry, Coordinated grid and place cell replay during rest, Nature Neuroscience 19 (6) (2016) 792–794. `doi:10.1038/nn.4291`.

[72] D. Ji, M. A. Wilson, Coordinated memory replay in the visual cortex and

hippocampus during sleep, Nature Neuroscience 10 (1) (2007) 100–107. `doi:10.1038/nn1825`.

[73] T. Schaul, J. Quan, I. Antonoglou, D. Silver, Prioritized Experience Replay, arXiv:1511.05952 [cs]ArXiv: 1511.05952.
URL `http://arxiv.org/abs/1511.05952`

[74] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, Building Machines That Learn and Think Like People, arXiv:1604.00289 [cs, stat]ArXiv: 1604.00289.
URL `http://arxiv.org/abs/1604.00289`

[75] P. A. Tsividis, T. Pouncy, J. L. Xu, J. B. Tenenbaum, S. J. Gershman, Human Learning in Atari.
URL `http://gershmanlab.webfactional.com/pubs/Tsividis17.pdf`

[76] Y. Liu, R. J. Dolan, Z. Kurth-Nelson, T. E. Behrens, Human Replay Spontaneously Reorganizes Experience, Cell 178 (3) (2019) 640–652.e14. `doi:10.1016/j.cell.2019.06.012`.
URL `https://linkinghub.elsevier.com/retrieve/pii/S0092867419306403`

[77] A. Pritzel, B. Uria, S. Srinivasan, A. Puigdomnech, O. Vinyals, D. Hassabis, D. Wierstra, C. Blundell, Neural Episodic Control, arXiv:1703.01988 [cs, stat]ArXiv: 1703.01988.
URL `http://arxiv.org/abs/1703.01988`

[78] S. J. Gershman, Reinforcement learning and causal models, Oxford Handbook of Causal Reasoning (2015) 1–32.

[79] M. Lengyel, P. Dayan, Hippocampal Contributions to Control: The Third Way 8.

[80] Z. Lin, T. Zhao, G. Yang, L. Zhang, Episodic Memory Deep Q-Networks, arXiv:1805.07603 [cs, stat]ArXiv: 1805.07603.
URL `http://arxiv.org/abs/1805.07603`

[81] S. J. Gershman, N. D. Daw, Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework, Annual Review of Psychology 68 (1) (2017) 101–128, publisher: Annual Reviews. `doi:10.1146/annurev-psych-122414-033625`.
URL `https://www.annualreviews.org/doi/10.1146/annurev-psych-122414-033625`

[82] A. M. Bornstein, K. A. Norman, Reinstated episodic context guides sampling-based decisions for reward, Nature Neuroscience 20 (7) (2017) 997–1003, number: 7 Publisher: Nature Publishing Group. `doi:10.1038/nn.4573`.
URL `https://www.nature.com/articles/nn.4573`

[83] A. M. Bornstein, M. W. Khaw, D. Shohamy, N. D. Daw, Reminders of past choices bias decisions for reward in humans, Nature Communications 8 (1) (2017) 15958, number: 1 Publisher: Nature Publishing Group. `doi:10.1038/ncomms15958`.
URL `https://www.nature.com/articles/ncomms15958`

[84] A. L. Samuel, Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development 3 (3) (1959) 210–229.

[85] P. R. Montague, P. Dayan, S. J. Nowlan, A. Pouget, T. J. Sejnowski, Using Aperiodic Reinforcement for Directed Self-Organization During Development, in: S. J. Hanson, J. D. Cowan, C. L. Giles (Eds.), Advances in Neural Information Processing Systems 5, Morgan-Kaufmann, 1993, pp. 969–976.
URL `http://papers.nips.cc/paper/606-using-aperiodic-reinforcement-for-directed-self-or pdf`

[86] T. J. Sejnowski, P. Dayan, P. R. Montague, Predictive Hebbian Learning, in: Proceedings of the Eighth Annual Conference on Computational Learning Theory, COLT '95, Association for Computing Machinery, New

York, NY, USA, 1995, pp. 15–18. `doi:10.1145/225298.225300`.
URL `https://doi.org/10.1145/225298.225300`

[87] W. Schultz, Predictive Reward Signal of Dopamine Neurons, Journal of Neurophysiology 80 (1) (1998) 1–27. `doi:10.1152/jn.1998.80.1.1`.
URL `https://doi.org/10.1152/jn.1998.80.1.1`

[88] W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, M. Botvinick, A distributional code for value in dopamine-based reinforcement learning, Nature 577 (7792) (2020) 671–675. `doi: 10.1038/s41586-019-1924-6`.

[89] B. B. Doll, D. A. Simon, N. D. Daw, The ubiquity of model-based reinforcement learning, Current opinion in neurobiology 22 (6) (2012) 1075–1081.

[90] E. L. Thorndike, A proof of the law of effect, Science 77 (1989) (1933) 173–175. `arXiv:https://science.sciencemag.org/content/77/1989/173.2.full.pdf`, `doi:10.1126/science.77.1989.173-a`.
URL `https://science.sciencemag.org/content/77/1989/173.2`

[91] E. TOLMAN, Cognitive maps in rats and men, Psychological review 55 (4) (1948) 189208. `doi:10.1037/h0061626`.
URL `https://doi.org/10.1037/h0061626`

[92] M. R. Delgado, L. E. Nystrom, C. Fissell, D. C. Noll, J. A. Fiez, Tracking the hemodynamic responses to reward and punishment in the striatum, Journal of Neurophysiology 84 (6) (2000) 3072–3077, pMID: 11110834. `arXiv:https://doi.org/10.1152/jn.2000.84.6.3072`, `doi: 10.1152/jn.2000.84.6.3072`.
URL `https://doi.org/10.1152/jn.2000.84.6.3072`

[93] T. Hare, J. O'Doherty, C. Camerer, W. Schultz, A. Rangel, Dissociating the role of the orbitofrontal cortex and the striatum in the computation of

goal values and prediction errors, The Journal of Neuroscience 28 (2008) 5623 – 5630.

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

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

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

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

[98] K. I. Tsutsui, F. Grabenhorst, S. Kobayashi, W. Schultz, A dynamic code for economic object valuation in prefrontal cortex neurons, Nature Communications doi:10.1038/ncomms12554.

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

[100] N. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control, Nature Neuroscience 8 (2005) 1704–1711.

[101] E. S. Bromberg-Martin, M. Matsumoto, O. Hikosaka, Dopamine in Motivational Control: Rewarding, Aversive, and Alerting, 2010, iSSN: 08966273 Publication Title: Neuron. doi:10.1016/j.neuron.2010.11.022.

[102] N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control, Nature Neuroscience doi:10.1038/nn1560.

[103] X. Pan, K. Sawa, I. Tsuda, M. Tsukada, M. Sakagami, Reward prediction based on stimulus categorization in primate lateral prefrontal cortex, Nature Neuroscience`doi:10.1038/nn.2128`.

[104] J. Glscher, N. Daw, P. Dayan, J. P. O'Doherty, States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning, Neuron`doi:10.1016/j.neuron.2010.04.016`.

[105] A. Sirota, S. Montgomery, S. Fujisawa, Y. Isomura, M. Zugaro, G. Buzski, Entrainment of Neocortical Neurons and Gamma Oscillations by the Hippocampal Theta Rhythm, Neuron`doi:10.1016/j.neuron.2008.09.014`.

[106] K. Benchenane, A. Peyrache, M. Khamassi, P. L. Tierney, Y. Gioanni, F. P. Battaglia, S. I. Wiener, Coherent Theta Oscillations and Reorganization of Spike Timing in the Hippocampal- Prefrontal Network upon Learning, Neuron`doi:10.1016/j.neuron.2010.05.013`.

[107] J. M. Hyman, M. E. Hasselmo, J. K. Seamans, What is the functional relevance of prefrontal cortex entrainment to hippocampal theta rhythms?, 2011, iSSN: 16624548 Publication Title: Frontiers in Neuroscience. `doi:10.3389/fnins.2011.00024`.

[108] T. Womelsdorf, M. Vinck, L. Stan Leung, S. Everling, Selective theta-synchronization of choice-relevant information subserves goal-directed behavior, Frontiers in Human Neuroscience`doi:10.3389/fnhum.2010.00210`.

[109] A. Pouget, J. M. Beck, W. J. Ma, P. E. Latham, Probabilistic brains: Knowns and unknowns, Nature Neuroscience 16 (9) (2013) 1170–1178, publisher: Nature Publishing Group. `doi:10.1038/nn.3495`.
URL `http://dx.doi.org/10.1038/nn.3495`

[110] S. Lammel, B. K. Lim, R. C. Malenka, Reward and aversion in a heterogeneous midbrain dopamine system, Neuropharmacology 76 (PART B)

(2014) 351–359, publisher: Elsevier Ltd. `doi:10.1016/j.neuropharm.`
`2013.03.019.`
URL `http://dx.doi.org/10.1016/j.neuropharm.2013.03.019`

[111] W. Dabney, M. Rowland, M. G. Bellemare, R. Munos, Distributional reinforcement learning with quantile regression, 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 (2018) 2892–2901ISBN: 9781577358008 _eprint: 1710.10044.

[112] H. F. Harlow, The formation of learning sets, Psychological Review`doi:` `10.1037/h0062474.`

[113] G. Marcus, Deep Learning: A Critical Appraisal (2018) 1–27_eprint: 1801.00631.
URL `http://arxiv.org/abs/1801.00631`

[114] T. Schaul, J. Schmidhuber, Metalearning, Scholarpedia`doi:10.4249/` `scholarpedia.4650.`

[115] J. X. Wang, Z. Kurth-Nelson, D. Tirumala, H. Soyer, J. Z. Leibo, R. Munos, C. Blundell, D. Kumaran, M. Botvinick, Learning to reinforcement learn (2016) 1–17_eprint: 1611.05763.
URL `http://arxiv.org/abs/1611.05763`

[116] M. F. Rushworth, T. E. Behrens, Choice, uncertainty and value in prefrontal and cingulate cortex, 2008, iSSN: 10976256 Publication Title: Nature Neuroscience. `doi:10.1038/nn2066.`

[117] J. N. Kim, M. N. Shadlen, Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque, Nature Neuroscience`doi:` `10.1038/5739.`

[118] S. Ritter, J. X. Wang, Z. Kurth-Nelson, S. M. Jayakumar, C. Blundell, R. Pascanu, M. Botvinick, Been there, done that: Meta-learning with episodic recall, in: 35th International Conference on Machine Learning, ICML 2018, 2018.

[119] S. Ritter, J. Wang, Z. Kurth-Nelson, M. Botvinick, Episodic Control as Meta-Reinforcement Learning, bioRxivdoi:10.1101/360537.

[120] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, T. Lillicrap, Meta-Learning with Memory-Augmented Neural Networks, in: 33rd International Conference on Machine Learning, ICML 2016, 2016.

[121] G. Wayne, C.-C. Hung, D. Amos, M. Mirza, A. Ahuja, A. Grabska-Barwinska, J. Rae, P. Mirowski, J. Z. Leibo, A. Santoro, M. Gemici, M. Reynolds, T. Harley, J. Abramson, S. Mohamed, D. Rezende, D. Saxton, A. Cain, C. Hillier, D. Silver, K. Kavukcuoglu, M. Botvinick, D. Hassabis, T. Lillicrap, Unsupervised Predictive Memory in a Goal-Directed Agent_eprint: 1803.10760.
URL http://arxiv.org/abs/1803.10760

[122] O. Vikbladh, D. Shohamy, N. D. Daw, Episodic Contributions to Model - Based Reinforcement Learning, Cognitive Computational Neuroscience Conference.

[123] M. Oberst, D. Sontag, Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models, ArXiv abs/1905.05824.

[124] L. Buesing, T. Weber, Y. Zwols, S. Racanire, A. Guez, J.-B. Lespiau, N. Heess, Woulda, Coulda, Shoulda: Counterfactually-Guided Policy Search, CoRR abs/1811.06272, _eprint: 1811.06272.
URL http://arxiv.org/abs/1811.06272

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

[126] F. A. Soto, S. Gershman, Y. Niv, Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization., Psychological review 121 3 (2014) 526–58.

[127] A. G. Barto, S. Mahadevan, Recent Advances in Hierarchical Reinforcement Learning, 2003, iSSN: 09246703 Publication Title: Discrete

Event Dynamic Systems: Theory and Applications. `doi:10.1023/A:1022140919877`.

[128] T. G. Dietterich, Hierarchical Reinforcement Learning with the MAXQ Value Function Decomposition, Journal of Artificial Intelligence Research_eprint: 9905014. `doi:10.1613/jair.639`.

[129] R. Parr, S. Russell, Reinforcement learning with hierarchies of machines, in: Advances in Neural Information Processing Systems, 1998, iSSN: 10495258.

[130] R. S. Sutton, D. Precup, S. Singh, Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning, Artificial Intelligence`doi:10.1016/S0004-3702(99)00052-1`.

[131] K. Lashley, The problem of serial order in behavior, in: Cerebral mechanisms in behavior, 1951.

[132] G. A. Miller, E. Galanter, K. H. Pribram, Plans and the structure of behaviour, in: Systems Research for Behavioral Science: A Sourcebook, 2017, iSSN: 00978507. `doi:10.2307/411065`.

[133] a. Newell, J. C. Shaw, H. a. Simon, Report on a general problem-solving program, 1959, iSSN: 00280836 Publication Title: IFIP Congress. `doi:10.1038/224923a0`.

[134] J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, Y. Qin, An integrated theory of the mind, 2004, iSSN: 0033295X Publication Title: Psychological Review. `doi:10.1037/0033-295X.111.4.1036`.

[135] M. Botvinick, D. C. Plaut, Doing Without Schema Hierarchies: A Recurrent Connectionist Approach to Normal and Impaired Routine Sequential Action, Psychological Review`doi:10.1037/0033-295X.111.2.395`.

[136] D. W. Schneider, G. D. Logan, Hierarchical control of cognitive processes: Switching tasks in sequences, Journal of Experimental Psychology: General`doi:10.1037/0096-3445.135.4.623`.

[137] J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, J. R. Reynolds, Event perception: A mind-brain perspective, 2007, iSSN: 00332909 Publication Title: Psychological Bulletin. `doi:10.1037/0033-2909.133.2.273`.

[138] D. Badre, Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes, 2008, iSSN: 13646613 Publication Title: Trends in Cognitive Sciences. `doi:10.1016/j.tics.2008.02.004`.

[139] M. M. Botvinick, Hierarchical models of behavior and prefrontal function, 2008, iSSN: 13646613 Publication Title: Trends in Cognitive Sciences. `doi:10.1016/j.tics.2008.02.009`.

[140] S. M. Courtney, J. K. Roth, J. B. Sala, A hierarchical biased-competition model of domain-dependent working memory maintenance and executive control, in: The Cognitive Neuroscience of Working Memory, 2012. `doi:10.1093/acprof:oso/9780198570394.003.0021`.

[141] J. M. Fuster, A theory of prefrontal functions: The prefrontal cortex and the temporal organization of behavior, in: The prefrontal cortex: Anatomy, physiology and neuropsychology of the frontal lobe, 1989.

[142] E. Koechlin, C. Ody, F. Kouneiher, The Architecture of Cognitive Control in the Human Prefrontal Cortex, Science`doi:10.1126/science.1088545`.

[143] J. D. Cohen, K. Dunbar, J. L. McClelland, On the control of automatic processes: A parallel distributed processing account of the stroop effect, Psychological Review`doi:10.1037/0033-295X.97.3.332`.

[144] R. Cooper, T. Shallice, Contention scheduling and the control of routine activities, Cognitive Neuropsychology`doi:10.1080/026432900380427`.

[145] S. Monsell, Task switching, 2003, iSSN: 13646613 Publication Title: Trends in Cognitive Sciences. `doi:10.1016/S1364-6613(03)00028-7`.

[146] D. Newtson, Foundations of attribution: The perceoption of ongoing behavior, in: New Directions in Attribution Research, 1976.

[147] J. M. Zacks, B. Tversky, Event structure in perception and conception, Psychological Bulletin`doi:10.1037/0033-2909.127.1.3`.

[148] J. S. Bruner, Organization of early skilled action., Child development`doi:10.1111/j.1467-8624.1973.tb02105.x`.

[149] K. W. Fischer, A theory of cognitive development: The control and construction of hierarchies of skills, Psychological Review`doi:10.1037/0033-295X.87.6.477`.

[150] P. M. Greenfield, K. Nelson, E. Saltzman, The development of rulebound strategies for manipulating seriated cups: A parallel between action and grammar, Cognitive Psychology`doi:10.1016/0010-0285(72)90009-6`.

[151] J. Lehman, J. Laird, P. Rosenbloom, A gentle introduction to Soar, an architecture for human cognition, Invitation to cognitive science.

[152] A. S. Luchins, Mechanization in problem solving: The effect of Einstellung., Psychological Monographs`doi:10.1037/h0093502`.

[153] M. M. Botvinick, Y. Niv, A. C. Barto, Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective, Cognition 113 (3) (2009) 262–280. `doi:10.1016/j.cognition.2008.08.011`.

[154] E. A. Mcgovern, Autonomous discovery of temporal abstractions from interaction with an environment, PowerISBN: 0-493-71665-3.

[155] M. Pickett, A. G. Barto, PolicyBlocks: An Algorithm for Creating Useful Macro-Actions in Reinforcement Learning, Proceedings of the 19th International Conference on Machine Learning (ICML)ISBN: 1-55860-873-7 _eprint: arXiv:1011.1669v3. `doi:10.1017/CBO9781107415324.004`.

[156] S. Thrun, A. Schwartz, Finding structure in reinforcement learning, Advances in Neural Information Processing Systems (NIPS) 7`doi:10.1.1.44.1621`.

[157] S. Mannor, I. Menache, A. Hoze, U. Klein, Dynamic abstraction in reinforcement learning via clustering, in: Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004, 2004. `doi: 10.1145/1015330.1015355`.

[158] I. Menache, S. Mannor, N. Shimkin, Q-cut - Dynamic discovery of subgoals in reinforcement learning, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2002, iSSN: 16113349.

[159] . imek, A. P. Wolfe, A. G. Barto, Identifying useful subgoals in reinforcement learning by local graph partitioning, in: Proceedings of the 22nd international conference on Machine learning - ICML '05, ACM Press, Bonn, Germany, 2005, pp. 816–823. `doi:10.1145/1102351.1102454`. URL `http://portal.acm.org/citation.cfm?doid=1102351.1102454`

[160] A. Gopnik, L. Schulz, Mechanisms of theory formation in young children, 2004, iSSN: 13646613 Publication Title: Trends in Cognitive Sciences. `doi:10.1016/j.tics.2004.06.005`.

[161] A. Gopnik, D. M. Sobel, D. Danks, C. Glymour, L. E. Schulz, T. Kushnir, A Theory of Causal Learning in Children: Causal Maps and Bayes Nets, Psychological Review`doi:10.1037/0033-295X.111.1.3`.

[162] J. A. Sommerville, A. L. Woodward, Infant's sensitivity to the casual features of means-end support sequences in action and perception, 2005, iSSN: 15250008 Publication Title: Infancy. `doi:10.1207/s15327078in0802_2`.

[163] J. A. Sommerville, A. L. Woodward, Pulling out the intentional structure

of action: The relation between action processing and action production in infancy, Cognitiondoi:10.1016/j.cognition.2003.12.004.

[164] A. G. Barto, S. Singh, N. Chentanez, Intrinsically motivated learning of hierarchical collections of skills, in: Proceedings of the 3rd International Conference on Development and Learning, 2004.

[165] S. Singh, A. G. Barto, N. Chentanez, Intrinsically motivated reinforcement learning, in: Advances in Neural Information Processing Systems, 2005, iSSN: 10495258.

[166] N. Bunzeck, E. Dzel, Absolute Coding of Stimulus Novelty in the Human Substantia Nigra/VTA, Neurondoi:10.1016/j.neuron.2006.06.021.

[167] P. Redgrave, K. Gurney, The short-latency dopamine signal: A role in discovering novel actions?, 2006, iSSN: 1471003X Publication Title: Nature Reviews Neuroscience. doi:10.1038/nrn2022.

[168] W. Schultz, P. Apicella, T. Ljungberg, Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task, Journal of Neurosciencedoi:10.1523/jneurosci.13-03-00900.1993.

[169] A. Lazaric, M. Ghavamzadeh, Bayesian multi-task reinforcement learning, 2010.

[170] R. Cohn-Gordon, N. D. Goodman, C. Potts, An incremental iterated response model of pragmatics, arXiv preprint arXiv:1810.00367.

[171] C. Colas, T. Karch, N. Lair, J.-M. Dussoux, C. Moulin-Frier, P. F. Dominey, P.-Y. Oudeyer, Language as a cognitive tool to imagine goals in curiosity-driven exploration, arXiv preprint arXiv:2002.09253.

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