Reinforcement Learning Inspired by Psychology and Neuroscience

Hongkun Wu[1, *]

[1]Department of Mathematics, University College London, London, United Kingdom

*Corresponding author: ucahhwu@ucl.ac.uk

Abstract. Decision-making is a crucial intelligence shared by animals and humans and it helps them to act in an intricate environment to seek rewards and avoid punishments. Psychologists first became interested in this ability and studied the conditional behavioral problems associated with it, then these studies have further led to the need for unified quantitative explanation models, among which Reinforcement learning is still the most convincing and data-backed model today. The model itself, in turn, facilitates research in neuroscience. In this paper, the researcher first introduces the original framework of reinforcement learning and the potential neural correlates to it. Then the paper reviews new developments in reinforcement learning algorithms that address the limitations of the original model as well as variants further inspired by neuroscience. Finally, the study highlights some new directions for future research. This study focuses on the evolution of reinforcement learning algorithms inspired by neuroscience, shows the relationship of mutual promotion and common development between reinforcement Learning and neuroscience, and clarifies some concerns for future exploration.

Keywords: Reinforcement learning; cognition; reward; artificial intelligence.

1. Introduction

One of the core abilities of animals and humans is to make decisions in complex environments, on a more general level, it is worth studying how to choose different behaviors for long-term rewards based on whether the environment gives rewards or punishment. Behavioral psychologists were the first to study this kind of problem in Pavlovian conditioning. Later, due to the increasing need for mathematical psychology and the increasing research in behavioral psychology, the model of reinforcement learning (RL) was proposed by Sutton and Barto to serve as a normative framework to understand the decision-making process [1]. With the development of brain science, electrophysiology and brain imaging technology, researchers have begun to explore the relationship between the brain response during decision-making and the reinforcement learning model. For example, A lot of evidence suggests that reinforcement learning models may be related to how the dopaminergic neurons in the brain work. On the other hand, cognitive and behavioral psychology also inspired the improvement of the reinforcement learning model. For example, inspired by the fact that human behavior and cognition are hierarchically organized, Botvinick, Niv and Barto proposed hierarchical reinforcement learning [2].

On a computational level, behavioral experiments about decision-making can be categorized into two types. Represented by Pavlovian conditioning, the aim of prediction learning is to learn the external environment state and event correlation [3]. While the other type of experiment requires the subject to learn to take a series of actions in a specific environment to maximize the payoff, which is called instrumental conditioning. In order to give a computational framework to instrumental conditioning, temporal difference learning is suggested by Sutton and Barto [4]. This model will generate the temporal difference reward prediction error when its predictions don't match the environment. This reinforcement learning signal is later detected in the brain by using functional imaging, and an increasing number of studies have shown that RL models are associated with dopaminergic neurons. Therefore, it is reasonable to believe that RL research and neuroscience research will mutually enhance each other. For instance, dopamine has been implicated in disorders such as Parkinson's disease and addictive behaviors [5]. RL provides a quantitative modeling framework that allows researchers to better analyze brain data about the dopamine mechanism.

Meanwhile, research into behavioral and neuroscience has also posed new challenges to the RL model, revealing some of its limitations. For instance, questions such as how learning from one task influences learning other related tasks or how to learn to do a complex task with a hierarchical structure need to be answered. The need to solve these problems has given rise to several algorithms such as deep reinforcement learning, hierarchical reinforcement learning and the combination of model-based and model-free algorithms. These algorithms have stronger explanatory power and have demonstrated excellent ability in domains such as playing Go and autonomous driving [6,7].

This study aims to explore the evolution of reinforcement learning algorithms inspired by neuroscience through conducting a systematic review. The proposed study will shed light on how neuroscience and reinforcement learning mutually reinforce each other and suggest some priorities for future reinforcement learning research.

The paper is organized as follows: in section 2 the researcher introduces the original RL framework. In section 3 the researcher gives some evidence of the association of RL with neurons. In section 4 the researcher further introduces the variants of reinforcement learning algorithms to address the limitations of the original one and briefly discuss the potential neural correlates of some of them. Finally, the researcher ends this paper with a summary to highlight some new directions for future research.

2. Reinforcement learning

2.1 Notations and terminologies

Before introducing the algorithms, the study will first introduce some notation and terminology for the framework of RL. The agent is the learning system trained to perform actions a_t in the environment. With each time the agent takes a different action, the state s_t of the environment changes and a reward r_t is fed back. Before the specific action is selected, A_t, S_{t+1} and R_t are all random variables. The density of A is called policy function

$$\pi(\mathrm{a} \mid \mathrm{s}) = \mathrm{P}(\mathrm{A} = \mathrm{a} \mid \mathrm{S} = \mathrm{s}) \tag{1}$$

and there is a state transition function to represent the randomness from the environment which is

$$\mathrm{p}(\mathrm{s}_{t+1} \mid \mathrm{s}_t, \mathrm{a}_t) = \mathrm{P}(\mathrm{S}_{t+1} = \mathrm{s}_{t+1} \mid \mathrm{S}_t = \mathrm{s}_t, \mathrm{A}_t = \mathrm{a}_t) \tag{2}$$

In different situations, whether the information of these two functions is known cannot be determined. A cumulative future reward is further defined by

$$U_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 R_{t+3} + \dots \tag{3}$$

where the γ (less than 1) is the discount factor since humans prefer immediate rewards and this tendency is also consistent with the economic concept of the value of time. With the cumulative reward, the action value function can be further defined.

$$Q_\pi(s_t, a_t) = \mathbb{E}[U_t \mid S_t = s_t, A_t = a_t] \tag{4}$$

and state value function

$$V_\pi(s_t) = \mathbb{E}_A[Q_\pi(s_t, A)] \tag{5}$$

In general, reinforcement learning is to obtain action value function or state value function through training to guide the selection of actions.

2.2 Temporal difference learning

As mentioned in the introduction, the RL model was originally proposed for prediction learning in Pavlovian conditioning. In the case that the Rescorla-Wagner model could not solve the second-order prediction, Sutton and Barto proposed the temporal difference model [4, 8].

The goal of temporal difference (TD) learning is to approximate the $V_\pi(s_t)$, which represent the cumulative future rewards or punishment. Here S_t is a Markov process and in Pavlovian condition the participants don't need to take actions (so the policy function is not needed). So from the definition of state value, the researcher can simplify the recursive relationship of $V(s_t)$ as

$$V(s_t) = E[r_t \mid s_t] + \gamma E[r_{t+1} \mid s_t] + \gamma^2 E[r_{t+2} \mid s_t] + \cdots$$
$$= r_t + \gamma \sum_{S_{t+1}} P(S_{t+1} \mid s_t) V(S_{t+1}) \tag{6}$$

This formula is the core of the TD algorithm, if the correct $V(s_t)$ is obtained, above formula must hold, otherwise the temporal difference prediction error can be defined by

$$\delta_t = r_t + \gamma \sum P(S_{t+1} \mid S_t) V(S_{t+1}) - V(S_t) \tag{7}$$

This error can further be used as a signal to update the approximation value of $V(s_t)$ by

$$V(S_t)_{\text{new}} = V(S_t)_{old} + \eta \cdot \delta_t \tag{8}$$

where η is the learning rate.

In practice, instead of using $P(S_{t+1} \mid S_t)$ which normally cannot be known in model-free cases, the state value function can be stochastically updated with stochastic prediction error

$$\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t) \tag{9}$$

Since the animal will obtain the samples from the environment, the state value function will eventually converge to the correct value.

2.3 Actor-Critic methods

In the case of instrumental conditioning, the actions need to be considered, as **Fig. 1**, the model can be departed into three parts, first of all, the whole learning takes place in an environment, then an actor (the agent), which is actually represented by the policy function π, receives state data from the environment and takes actions. Then a critic is consist of the state value function and TD error. After executing the action, the critic is responsible for receiving rewards from the environment and using the TD error to update both the policy and the state value, as introduced in the last section. After a certain number of training sessions, the critic will get better at assessing the situation and the actor will get better at choosing the right action to improve the cumulative rewards.

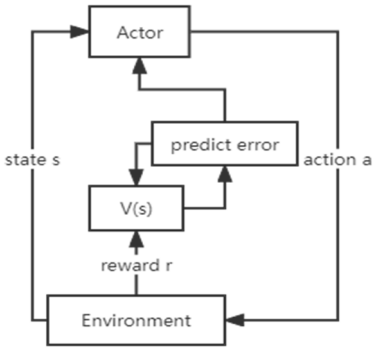

Fig. 1 Actor-Critic model

When the critic obtains a positive TD error, by its definition the positive value shows that the current action has improved the cumulative rewards, so the probability of the current action should be increased and vice versa. This observation implies the updating formula of policy

$$\pi(S,a)_{\text{new}} = \pi(S,a)_{\text{old}} + \eta_\pi \delta_t \tag{10}$$

The convergence of the above algorithm is given by Dayan and Abbott in some specific cases [9]. Finally, for the case of instrumental conditioning, a simpler algorithm design idea is introduced in the next subsection.

2.4 Q value learning

If the model can explicitly learn the action value function, the agent actually can choose actions according to the action value function without policy, so the Q value learning is the alternative to Actor-Critic methods. Wakins suggested Q-learning under the framework of TD algorithm, which is very similar to the above methods [10].

$$Q(S_t, a_t)_{\text{new}} = Q(S_t, a_t)_{\text{old}} + \eta\delta_t \tag{11}$$

where

$$\delta_t = r_t + \max_a \gamma Q(S_{t+1}, a) - Q(S_t, a_t) \tag{12}$$

Or a_{t+1} can just be substituted to the action value function instead of using the Max function as long the policy is known. Bertsekas and Tsitsiklis prove that both of these two forms of TD error allow the Q learning to converge if the learning rate satisfies proper conditions [11].

3. Potential neural correlates of reinforcement learning

RL algorithms such as the ones mentioned above are used to interpret data from a variety of neuroscience studies. One that has received the most attention is the dopamine reward mechanism. The dopamine system is essential to the proper functioning of human cognition. It has been shown to be involved in working and learning memory, and disorders resulting from the malfunction of the dopamine system can lead to addiction, depression and Parkinson's disease.

Initially, dopamine was thought by behaviorists to be the signal of reward; however, in the experiment conducted by Schultz, Apicella and Ljungberg, behaviorists recorded the dopaminergic responses of monkeys performing in the Pavlovian conditioning experiments [12]. At first, when the monkey received the reward (eating the fruit), the electrical signal was detected, and then a sound was presented before the reward was given. After the training was repeated many times, the dopaminergic response was actually shifted to the sound. The dopaminergic response does not equal the reward in the real world; instead, it is found by Barto to coincide with the temporal difference prediction error in the TD model.

Later, Schultz et al. further investigated the similarities between RL and the dopamine system in the Pavlovian conditioning experiment and reported that dopaminergic activity does shift from the Unconditioned Stimulus (US) to the Conditioned Stimulus (CS) [13]. Moreover, when the shifting happens, removing the US will result in a negative temporal difference prediction error at the time when the US would be delivered. Besides, if training with an earlier CS 2 after the signal has shifted to CS, then the signal will further shift to CS 2. The above phenomena all point to the consistency of dopaminergic neural activity with temporal difference prediction error. The reward prediction error hypothesis of dopamine was formally put forward by Montague et al [14]. Gradually, a lot of research related has been done and the neural data indeed further support this hypothesis, even in more complex conditioning tasks.

In particular, for the Actor-Critic model mentioned in the previous section, Barto suggests that the neural underpinning of the Critic in the model is related to the ventral tegmental, ventral striatal and frontal areas [15]. While Houk et al. found that a signal in the substantia nigra pars compacta to dorsal striatal target areas is similar to the TD error in the Actor-Critic model to train the actor [16].

However, the experimental data don't always match the hypothesis, researchers found that while in many cases the data showed support for the Actor-Critic model, in some cases Q-learning or other reinforcement learning models that are not covered in the present study were better at interpreting the data. This phenomenon could indicate that the brain may not be using just one simple reinforcement learning model, but rather choosing between different situations.

The decision-making processes in the brain are complex: some of which are not dopamine-dependent, and there are hypotheses linking RL models to other neural decision-making mechanisms, but they won't be discussed any further here. No matter what kind of hypothesis is based on, the exploration technology of the brain is very important to the corresponding research. Among them, functional magnetic resonance imaging (fMRI) has been widely used in the study of the relationship between the RL model and brain structure. This is a non-invasive but limited precisive way of probing neural activity by measuring signals that are dependent on the oxygen content in the blood.

Berns, McClure, Pagnoni, and Montague are the first researchers who use fMRI to detect the TD error signal in the brain and their finding suggests that the nucleus accumbens and the orbitofrontal

cortex may be involved [17]. Later, more studies used this technology to conduct related experiments and a lot of data pointed to the correlation between temporal difference prediction error and the dorsal and ventral striatum. However, researchers have been unable to distinguish reward from punishment signals, and the neural mechanisms underlying punishment learning remain unclear.

Overall, RL models provide a convictive quantitative framework for decision-making and learning in the brain, which is supported by data obtained based on different experimental methods. But on the one hand, these detection methods have their own limitations. For example, researchers do not know exactly what the signal in fMRI means. In addition, there are many hypotheses about RL models related to the brain, and a more complex and unified RL model is required to reflect how different models fit the data in different experiments.

4. Variants of reinforcement learning algorithms

In response to new problems discovered in neuroscience research, there have been many wonderful and powerful developments of RL models from their original foundations.This section will briefly introduce new developing RL models which are further inspired by cognitive science and neuroscience.

The problem of generalization, which is mentioned in the previous section, focuses on how behavior learned from the past tasks would affect future decisions. Inspired by the cognitive hierarchy theory, Botvinick , Niv and Barto introduce hierarchical reinforcement learning (HDL) to model different levels of behaviors so that learning from a single task can transfer to a lower level and affect future decisions [2]. In addition, the model has many other benefits such as it can also greatly simplify the complexity of the algorithm through the method of state abstraction or time abstraction. As a result, the model can deal with the curse of dimensionality in more complex tasks. Different from the general RL agent, which can only choose actions, the HRL agent can also choose sub-paths which have different sub-goals.

With the development of deep machine learning, RL is also combined with deep neural networks. For example, the policy function, the state value function and the action value function in its algorithm can all be modeled using deep neural networks. For example, Volodymyr et al. used the deep Q-network to train the agent to play Atari games and achieved good results [18].

Another notable development in RL modeling is the AlphaGo algorithm, as mentioned in section 2. The decision-making mechanism of the brain may be a combination of model-free and model-based algorithms. AlphaGo algorithm combined a pre-trained model-based Critic and a model-free actor, which makes it both cheap to be trained and flexible enough to respond to changes in the game.In addition, the policy function and action-value function of the algorithm are approximated by deep neural networks, which gives it enough degrees of freedom to deal with complex tasks. In the end, the algorithm performed well and defeated the Go world champion in 2016.

Finally, there are many other interesting problems in reinforcement learning, such as the Markov decision problem, signal detection, and the exploration and exploitation problem. These questions can be linked to the optimal stopping problem. Peter and Nathaniel proposed a unified and coherent Bayesian approach to unify the above problems [19]. Under the framework of the Bayesian approach, the algorithmic nature of various problems can be understood and studied more clearly.

5. Conclusion

To conclude, there has long been a passion for the exploration of decision-making processes, from the initial qualitative analysis to the requirement of mathematical models, the proposed RL framework provides a concise and powerful paradigm for studying decision-making problems and effectively explains a lot of neuroscience data.

Meanwhile, the research based on neuroscience and the development of imaging technology further revealed the potential connection between RL models and various neural mechanisms. In

particular, the connection between dopamine function and the RL model has attracted much attention and a large number of research data support the association between the two. Related research can not only promote the understanding of dopamine function and help treat diseases caused by dopamine dysfunction, but also promote the development of the RL model in turn.

Inspired by neuroscience and cognitive science, RL algorithms have also developed more and more powerful. In order to answer how past learning can play an impact in the future, HRL was proposed. In order to deal with more complex learning behaviors, researchers further use deep neural networks to approximate key functions in RL models. These new algorithms have been widely and powerfully applied in various fields. However, the focus of many algorithms is starting to shift from neuroscience to solving real-world problems.

Problems such as whether these newly proposed algorithms can be placed under a larger decision-making framework and whether they have a corresponding neural basis and can shed further light on human cognitive mechanisms are still to be studied, but in any case, RL is already a very important and fruitful in the field of decision-making.

References

[1] Sutton R S, Barto A G. Reinforcement learning: An introduction[J]. Robotica, 1999, 17(2): 229-235..

[2] Botvinick M M, Niv Y, Barto A G. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective[J]. Cognition, 2009, 113(3): 262-280.

[3] Yerkes R M, Morgulis S. The method of Pawlow in animal psychology[J]. Psychological Bulletin, 1909, 6(8): 257.

[4] Sutton R S, Barto A G. Time-derivative models of pavlovian reinforcement[J]. 1990.

[5] Redish A D, Jensen S, Johnson A, et al. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling[J]. Psychological review, 2007, 114(3): 784.

[6] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge[J]. nature, 2017, 550(7676): 354-359.

[7] Shalev-Shwartz S, Shammah S, Shashua A. Safe, multi-agent, reinforcement learning for autonomous driving[J]. arXiv preprint arXiv:1610.03295, 2016.

[8] Fanselow M S. Pavlovian conditioning, negative feedback, and blocking: mechanisms that regulate association formation[J]. Neuron, 1998, 20(4): 625-627.

[9] Dayan P, Abbott L F. Theoretical neuroscience: computational and mathematical modeling of neural systems[M]. MIT press, 2005.

[10] Watkins C J C H. Learning from delayed rewards[J]. 1989.

[11] Bertsekas D, Tsitsiklis J N. Neuro-dynamic programming[M]. Athena Scientific, 1996.

[12] Schultz W, Apicella P, Ljungberg T. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task[J]. Journal of neuroscience, 1993, 13(3): 900-913.

[13] Schultz W, Apicella P, Ljungberg T. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task[J]. Journal of neuroscience, 1993, 13(3): 900-913.

[14] Montague P R, Dayan P, Sejnowski T J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning[J]. Journal of neuroscience, 1996, 16(5): 1936-1947.

[15] Barto A G. 1 '1 Adaptive Critics and the Basal Ganglia,.'[J]. Models of information processing in the basal ganglia, 1995, 215.

[16] Houk J C, Adams J L. 13 A model of how the basal ganglia generate and use neural signals that[J]. Models of information processing in the basal ganglia, 1995: 249.

[17] Berns G S, McClure S M, Pagnoni G, et al. Predictability modulates human brain response to reward[J]. Journal of neuroscience, 2001, 21(8): 2793-2798.

[18] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[J]. arXiv preprint arXiv:1312.5602, 2013.

[19] Dayan P, Daw N D. Decision theory, reinforcement learning, and the brain[J]. Cognitive, Affective, & Behavioral Neuroscience, 2008, 8(4): 429-453.