Reinforcement Learning: Beyond the Basal Ganglia

Chapter · April 2021

DOI: 10.1007/978-3-030-84729-6_16

2 authors, including:

Michel Fathi
University of North Texas
113 PUBLICATIONS **762** CITATIONS

Reinforcement Learning:

beyond the basal ganglia

Yuan, Chengping

University of North Texas, chengpingyuan@my.unt.edu

Fathi, Mahdi

University of North Texas, Mahdi.Fathi@unt.edu

Abstract

In clinical psychology, reinforcement learning is one of the many forms of conditional learning that focuses upon reinforcing behavior that yields beneficial results. Similar to any other structure, the process of reinforcement learning was adopted and implemented into models of machine learning with ever-evolving complexity. This chapter will attempt to outline the philosophy and terminology associated with reinforcement learning in machines while also describing basic models and their connection to vertebrate neuroanatomy. Specifically, machine learning borrows from the basal ganglia, a cluster of subcortical nuclei that are responsible for a host of tasks including procedural learning, cognition, and emotion. Within this chapter, the parallels between the basal ganglia and current machine learning will be clearly evaluated with thoughtful deliberation of their future relationship.

Introduction to reinforcement learning

Reinforcement learning is an area of machine learning, inspired by behaviorist psychology, concerned with how agents ought to take actions in an environment so as to maximize some notion of cumulative reward. The task of reinforcement learning is to use observed rewards to learn an optimal policy for the environment. In many complex domains, reinforcement learning is the feasible way to train a program to perform at high levels. For example, in game playing, it is very hard for a human to provide accurate and consistent evaluations of large numbers of positions, which would be needed to train an evaluation function directly from examples. Instead, the program can be told when it has won or lost, and it can use this information to learn an evaluation function that gives reasonably accurate estimates of the probability of winning from any given position. Figure 1 is widely used to describe how reinforcement learning works. We can see that an Agent will take a certain action to receive the reward at the timestamp . There are many elements within reinforcement learning and it is important to know them: environment, agent, policy,

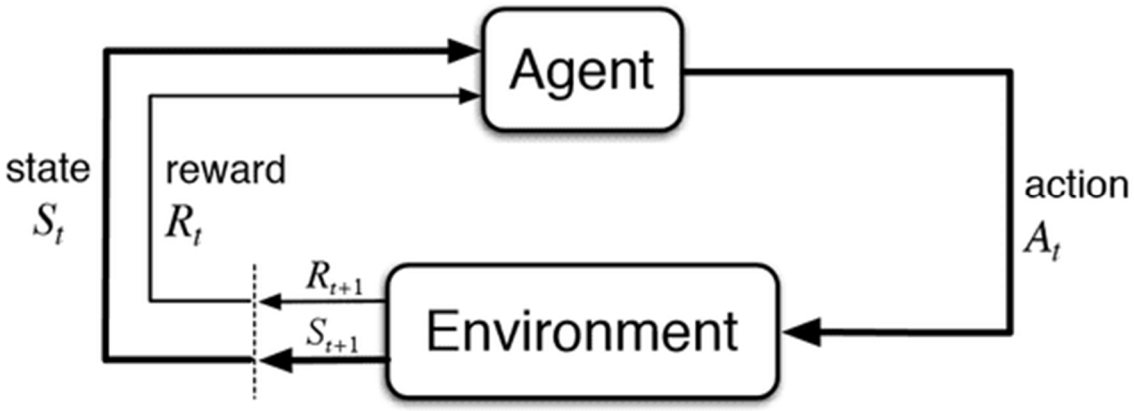

Figure 1. Reinforcement Learning Framework

Environment: The environment's task is to define a world where an agent is able to interact with. It therefore has a basic loop that can be written like this:

Produce state s and reward r

Where our state represents the current situation in the environment and the reward represents the scalar value being returned by the environment after selecting an action .

Agent: Our agent needs to learn how to achieve goals by interacting with the environment. The basis to do this is by using a basic loop.

1. Sense state and reward from the environment
2. Select an action based on this state and reward

We do note here though that the action that our agent can take can be defined under two specific categories:

1. Discrete: 1 of N actions (for example, left or down)
2. Continuous: An action as a scalar/vector of a real value (for example, the amount we need to bend our leg to be able to walk)

Policy π: Policy $\pi$is used to map the actions to the states that agents have to take. The actions can be categorized in two specific categories:

- Deterministic: Same action every time
- Stochastic: There is a probability of taking different actions (for example, we take action 1 70% of the time, and action 2 30% of the time).

Value function V **:** Value function V represents how good the state in the long run, which is calculated by our agents, in other words, what is the expected long term accumulation of reward.

There are two commonly used value functions:

1. State-Value Functions $V^{\pi}(S)$: Value of state S and following our policy π
   - It give the expected return when starting from states s and following policy πforever.
   - $V^{\pi}(s) = E_{\pi}[R_t||s_t = s]$
2. Action-Value Functions $Q^{\pi}(s, a)$: Value of state s ,taking action a ,and thereafter following policy π.

●It gives the expected return of taking action a in state s, given the policy π

●$Q^{\pi}(s,a) = E_{\pi}[R_t||s_t = s, a_t = a]$

●This is also called Q-Function, which is the core of the Q-learning algorithm.

There are two types of reinforcement learning: model-based learning vs model-free learning. In model-based learning, agents not only learn how to take actions but also learn how the environment responds to moves or actions. Model-free agents can estimate the optimal policy without using or estimating the dynamic (transaction and reward function) of the environment. Q-learning is a model-free reinforcement learning algorithm.

Here we are using the game-Tic Tac Toe to explore reinforcement learning model (Q-learning):

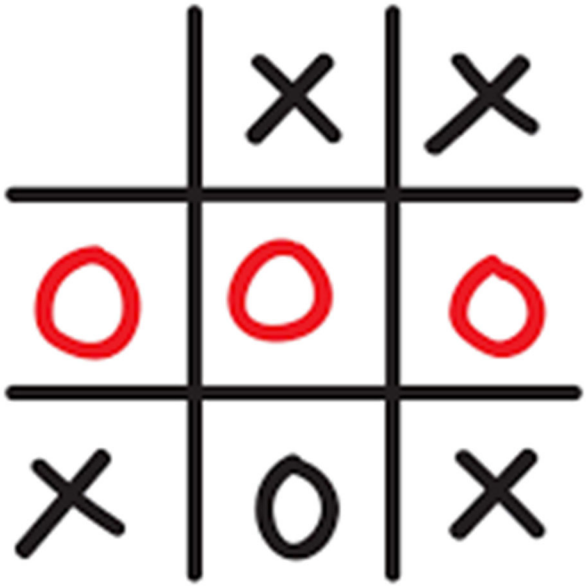

Figure2. Tic Tac Toe game board with each player played four moves

●States: States in Tic Tac Toe are the representation of the game board with moves of each player made.

●Actions (available actions): Available spaces to make move on board

●Reward: Winner will be rewarded by 100, Loser will be punished by 100 and reward is 0 when game tied.

●αValue function: Q(s,a), s is the state and a is the action (move) so that every state and action pair will have a q value from the value function. This requires a lot of game simulations.

How do we calculate the q value based on the elements above? Let's go ahead and derive Q learning by incorporating a few basic principles:

●Reward prediction error: Reward prediction error (RPE) = Actual reward - predicted reward value. An "error signal" used to adjust your value function. RPE in reinforcement learning is also from neuroscience (Schultz, 2016) - the prediction error theory. In this theory, dopamine neurons send a rapid signal that covers all three possible errors in prediction of a reward: that the reward was better than expected (a positive error); the reward was equal to expected (no error) or the reward was less than expected (a negative error). Humans can adjust their behaviors (actions) based on these three types of signal). In reinforcement learning, we use RPE and learning rate α to update the q value:

$$\text{New Value} = \text{Old Value} + \alpha * \text{RPE}$$

Now we know how to update the q value with leaning rate and RPE. We need to derive equation for predicted reward value,if $Q(s_t, a_t)$ is the sum of expected future rewards at time t ,

and $Q(s_{t+1}, a_{t+1})$is the same thing one step ahead in the future

$$Q(s_t, a_t) = r_t + r_{t+1} + r_{t+2} + \ldots \qquad \text{(estimated rewards)}$$

$$Q(s_{t+1}, a_{t+1}) = r_{t+1} + r_{t+2} + \ldots$$

$Q(s_t, a_t) = r_t + Q(s_{t+1}, a_{t+1})$assuming you take the best choice at time $t + 1$ so

$$Q(s_t, a_t) = r_t + max(Q(s_{t+1}, a))$$

So you can use Q to find the predicted reward value at time t:

predicted reward value $r_t = Q(s_t, a_t) - max_a(Q(s_{t+1}, a))$

● Learning rates: how to choose the right learning rate is also an important factor in Q learning. High learning rate leads to learning quickly to adapt to changing environments, lower learning rate learns slowly but remains stable enough to noise and stochastic rewards to avoid forgetting. The right value of learning rate depends upon how stable the environment is, a more unstable or critical environment requires a higher learning rate.

● Temporal discounting: Temporal discounting is the tendency of people to discount reward as they approach a temporal horizon in the future or the past. To put it another way, it is a tendency to give greater value to rewards as they move away from their temporal horizons and towards the

"Now". This concept is used in both neurobiology and neuroeconomics. Q learning models (and other reinforcement learning models) introduce a discount factor r between 0 and 1 to reward value. When r close to 1 represents no discounting, a reward from future will have same weight as current reward, when r close to 0 leads to "hedonistic behavior", immediate reward has more weight than long-term reward.

To sum up everything we had, we can derive Q-learning equation formula:

$$new\ Q\ = old\ Q + \alpha[reward\ prediction\ error]$$

$$new\ Q\ = old\ Q + \alpha[actual\ reward - predicted\ reward]$$

$$Q_{t+1}(s,a) = Q_{\ t} + \alpha[R_{t+1} - [Q_t - \gamma\ max_a Q_t(next\ state, a)]$$

Basal ganglia and reinforcement learning

Animals need to select the most appropriate behavior in a given environment in order to survive. An important role in this process of action selection is played in all vertebrates by a set of subcortical structures called the basal ganglia (Redgrave et al., 1999). The information processing in the basal ganglia is very strongly modulated by dopamine. The basal ganglia are critically involved both in the process of selecting actions, and in learning which actions are worth making in a given context, as demonstrated by impairments of both functions in Parkinson's disease. Death of dopaminergic neurons in Parkinson's disease leads to problems

with movements (Blandini et al., 2000) as well as difficulties in learning from feedback (Knowlton et al., 1996). The basal ganglia is organized into two main pathways: Go and No-Go. The Go pathway is related to the initiation of movements. On the other hand, the No-Go pathway is possible to be related to the inhibition of movements (Kravitz et al., 2010). Two pathways and how they work are shown in Figure 2 (Frank, 2005).

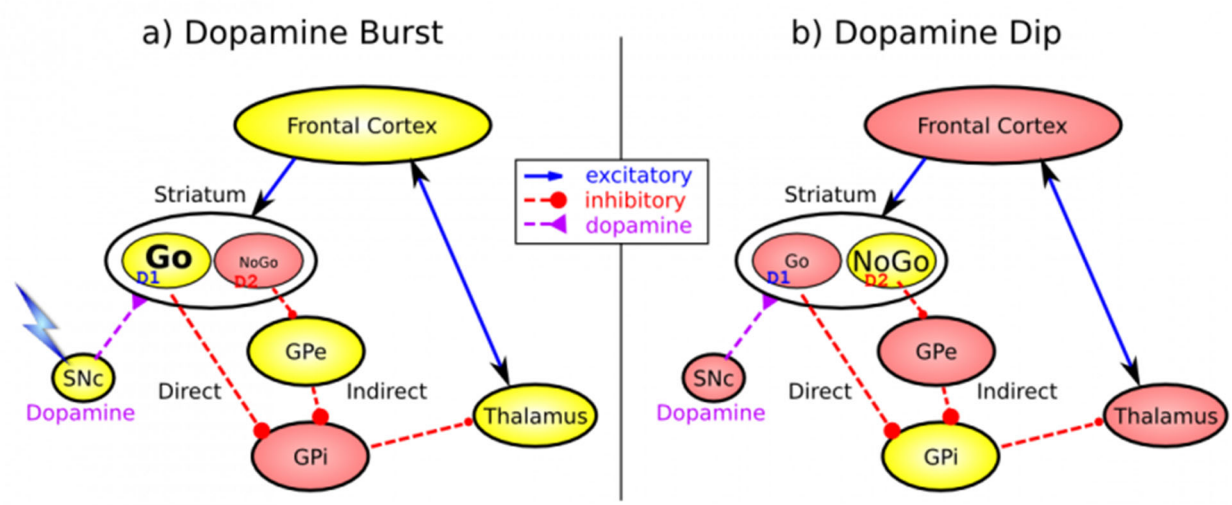

Figure 2: Biology of the basal ganglia system, with two cases shown: a) Dopamine burst activity that drives the direct "Go" pathway neurons in the striatum, which then inhibit the tonic activation in the globus pallidus internal segment (GPi), which releases specific nuclei in the thalamus from this inhibition, allowing them to complete a bidirectional excitatory circuit with the frontal cortex, resulting in the initiation of a motor action. The increased Go activity during dopamine bursts results in potentiation of corticostriatal synapses, and hence learning to select actions that tend to

result in positive outcomes. b) Dopamine dip (pause in tonic dopamine neuron firing), leading to preferential activity of indirect "NoGo" pathway neurons in the striatum, which inhibit the external segment globus pallidus neurons (GPe), which are otherwise tonically active, and inhibiting the GPi. Increased NoGo activity thus results in disinhibition of GPi, making it more active and thus inhibiting the thalamus, preventing initiation of the corresponding motor action. The dopamine dip results in potentiation of corticostriatal NoGo synapses, and hence learning to avoid selection actions that tend to result in negative outcomes.(Frank, 2005)

The competition between Go and No-Go pathway during action selection and its dopaminergic modulation inspired have been described by many computational models (e.g. Gurney et al., 2001; Humphries et al., 2012), which also lay the ground for reinforcement learning in machine learning. Dopaminergic neurons act similarly as reward functions in reinforcement learning, which will change the balance between the two pathways and promote action initiation over inhibition. The output nuclei of the basal ganglia play a similar role as value functions which will provide either positive or negative value to thalamus for action selection. The comparison of basal ganglia and reinforcement learning was described in figure 3.

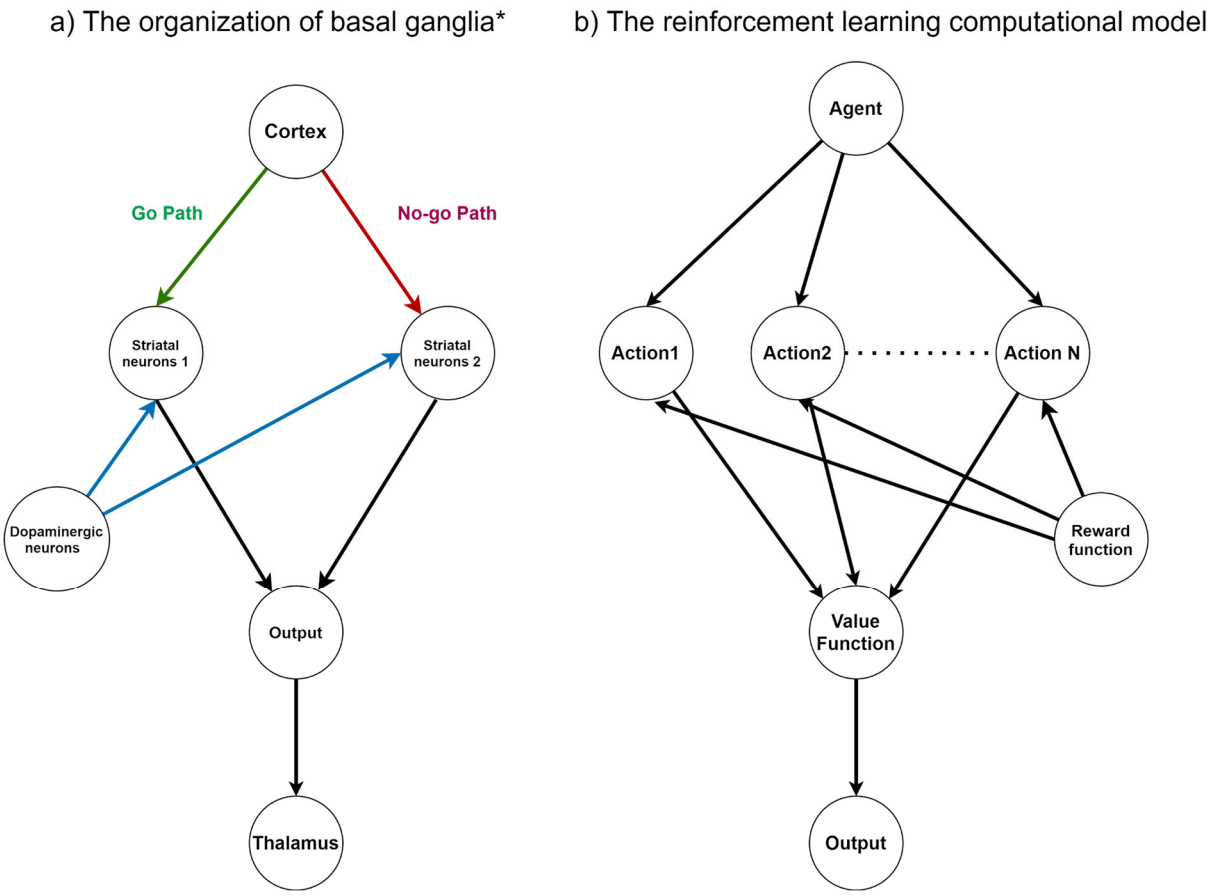

Figure 3. Comparison of basal ganglia and reinforcement learning in machine learning. a) The simplified organization of the basal ganglia diagram, components like globus pallidus internal segment (GPi) and globus pallidus neurons (GPe) are removed for better comparison. b) The reinforcement learning in machine learning model, see previous chapter for details of each component.*

Although the differences between reinforcement learning in the human brain and machine learning are tiny, the understanding of the human brain are still limited. Based on what we understand about our brain so far,

reinforcement learning seems to be a proper direction for us to take to create AI.

Future of Reinforcement learning

Deep Learning is state of the art for many challenging machine learning problems. With enough data, Deep Learning can outperform Machine learning in most scenarios. Reinforcement learning, on the other hand, has its own advantage compared to supervised and unsupervised learning. It can solve complex problems and make high level decisions. So combining deep learning and reinforcement learning become necessary to solve more challenging problems. This combination, called Deep reinforcement learning, is most useful in problems with high dimensional state-space (Francois-Lavet et al., 2018) and can apply to many real world scenarios.

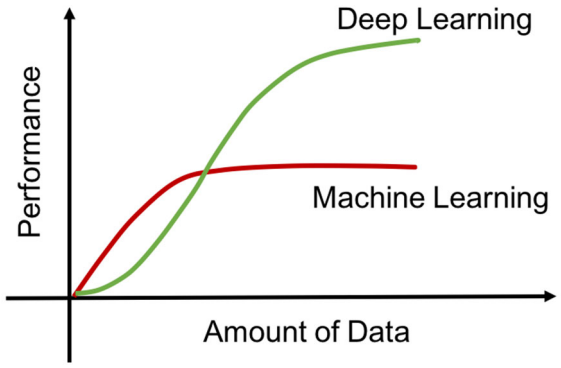

Figure 4.

Although reinforcement learning has shown its potential compared to other machine learning algorithms and techniques (Figure 5), there are still some limitations with it. For example, exploring the environment

efficiently or being able to generalize a good behavior in a slightly different context are not straightforward. Thus, researchers are proposing a large array of algorithms each year and trying to overcome these limitations.

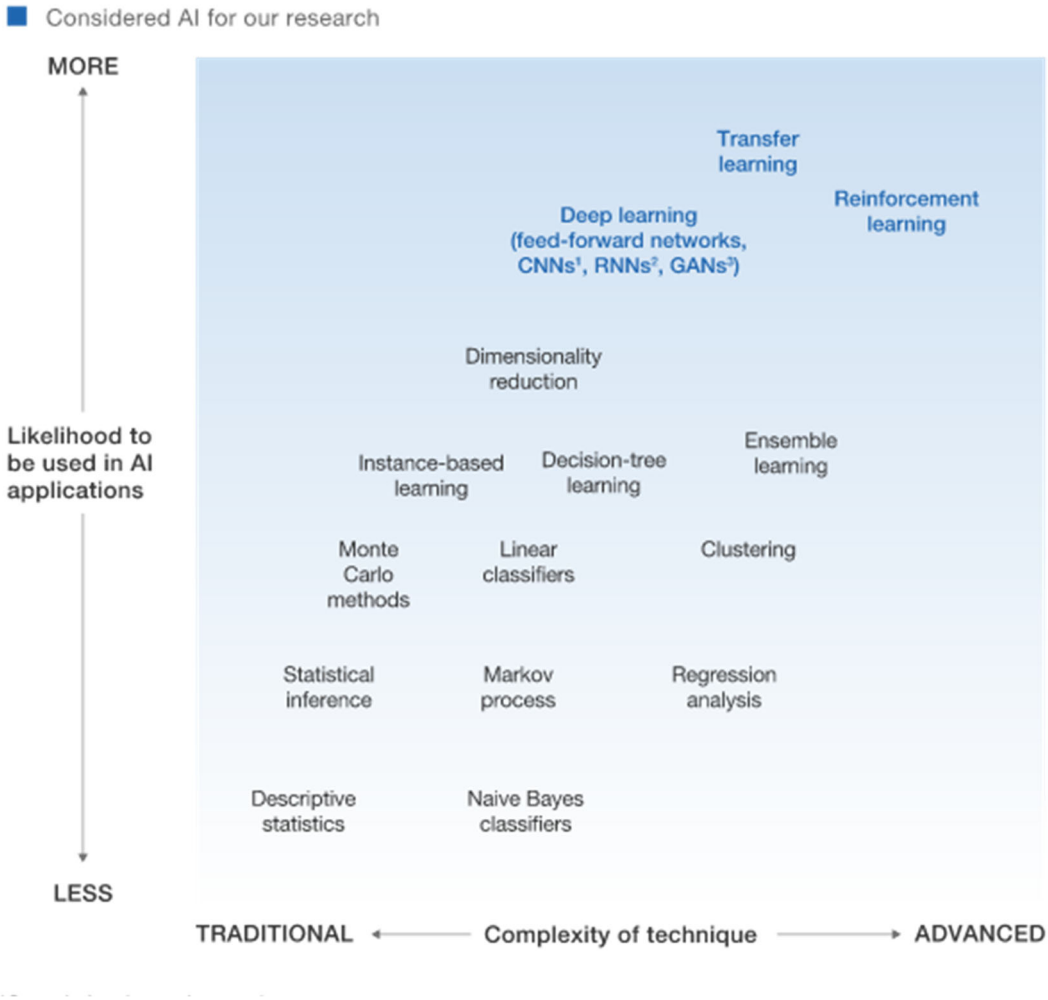

Figure 5. The machine learning algorithms and techniques with complexity and potential application in AI

Conclusion

The future of reinforcement learning is bright with so many efforts have been invested on reinforcement learning. In the foreseeable future of reinforcement learning, we can expect to see deep RL algorithms going in the direction of meta-learning and lifelong learning where previous knowledge (e.g., in the form of pre-trained networks) can be embedded so as to increase performance and training time. Another key challenge is to improve current transfer learning abilities between simulations and real-world cases. This would allow learning complex decision-making problems in simulations (with the possibility to gather samples in a flexible way), and then use the learned skills in real-world environments, with applications in robotics, self-driving cars, etc.

Finally, we expect deep RL techniques to develop improved curiosity driven abilities to be able to better discover by themselves their environment.

References:

Redgrave, Peter, Tony J. Prescott, and Kevin Gurney. "The basal ganglia: a vertebrate solution to the selection problem?." *Neuroscience* 89, no. 4 (1999): 1009-1023.

Blandini, Fabio, Giuseppe Nappi, Cristina Tassorelli, and Emilia Martignoni. "Functional changes of the basal ganglia circuitry in Parkinson's disease." *Progress in*

neurobiology 62, no. 1 (2000): 63-88.

Knowlton, Barbara J., Jennifer A. Mangels, and Larry R. Squire. "A neostriatal habit learning system in humans." *Science* 273, no. 5280 (1996): 1399-1402.

Kravitz, Alexxai V., Benjamin S. Freeze, Philip RL Parker, Kenneth Kay, Myo T. Thwin, Karl Deisseroth, and Anatol C. Kreitzer. "Regulation of parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry." *Nature* 466, no. 7306 (2010): 622-626.

Frank, Michael J. "Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism." *Journal of cognitive neuroscience* 17, no. 1 (2005): 51-72.

François-Lavet, Vincent, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. "An introduction to deep reinforcement learning." *arXiv preprint arXiv:1811.12560* (2018).

Schultz, Wolfram. "Dopamine reward prediction error coding." *Dialogues in clinical neuroscience* 18, no. 1 (2016): 23.