



Search Medium



Write



Computational Models of Cognition: Part VII: Reinforcement Learning



Alireza Dehbozorgi

26 min read · Just now



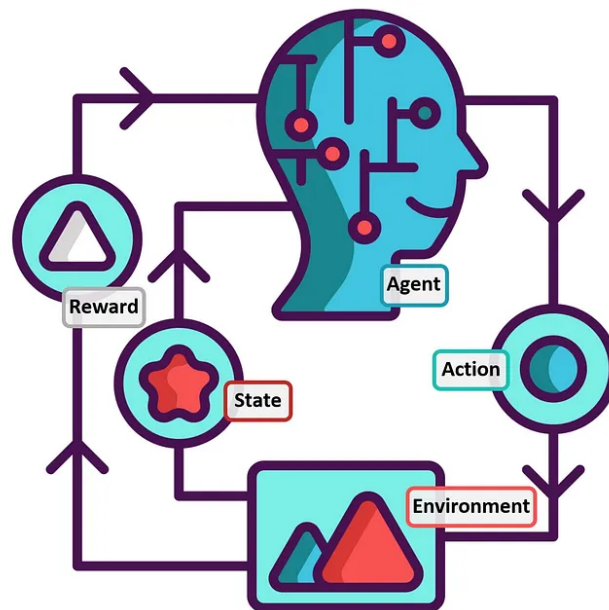


Image Source: <https://www.chrismahoney.com.au/blogs/2021-06-12-reinforcement-learning/>

Introduction

As a newborn or a novice sport player, one's actions are initially random or awkward, but with repeated experience one becomes able to achieve goals more efficiently and more reliably. Animal behavioral studies have described such processes of acquisition of behaviors by the concepts of reward and punishment. A reward promotes the execution of, or reinforces, the action that causes its delivery (Thorndike, 1898). A punishment can be considered as a negative reward signal that reduces the repetition of an action that causes, or reinforces an action that avoids its delivery. It is amazing how an

animal can acquire a variety of complex behaviors by linking its actions to consequent positive and negative rewards, either spontaneously in nature or through training by humans. This phenomenon has provided good motivation for artificial intelligence researchers to seek computer algorithms that allow machines to acquire a variety of functions simply from reward feedback signals (Barto et al., 1983).

As a newborn or a novice sport player, one's actions are initially random or awkward, but with repeated experience one becomes able to achieve goals more efficiently and more reliably. Animal behavioral studies have described such processes of acquisition of behaviors by the concepts of reward and punishment. A reward promotes the execution of, or reinforces, the action that causes its delivery (Thorndike, 1898). A punishment can be considered as a negative reward signal that reduces the repetition of an action that causes, or reinforces an action that avoids its delivery. It is amazing how an animal can acquire a variety of complex behaviors by linking its actions to consequent positive and negative rewards, either spontaneously in nature or through training by humans. This phenomenon has provided good motivation for artificial intelligence researchers to seek computer algorithms that allow machines to acquire a variety of functions simply from reward feedback signals (Barto et al., 1983).

The products of such studies are collectively called reinforcement learning and have been applied to a variety of control and optimization problems (Sutton & Barto, 2018) (SB hereafter). Since the mid-nineties, neuroscientists became aware of interesting parallels between the key signals used in reinforcement learning algorithms and what they found in neural recording and brain imaging data. The collaborations of theoreticians and experimentalists contributed to a better understanding of the functions of, most notably, the neurotransmitter dopamine and the neural circuit of the basal ganglia (Barto, 1995; Montague et al., 1995; Schultz et al., 1997). The success has now interested psychiatrists, sociologists, and economists who are trying to understand how humans make good (or bad) decisions in the real world (Doya, 2007; Glimcher & Fehr, 2013).

Reinforcement learning is one of the three major frameworks of machine learning. One is *supervised learning*, which takes explicit target output signal and minimizes the error between the learner's output and the target output. Another is *unsupervised learning*, which takes no target output but captures the statistical features of the input signal, such as clustering and dimension reduction. Reinforcement learning is positioned between supervised and unsupervised learning, by requiring scalar reward signal for a series of action outputs.

Markov Decision Process

The basic theory of reinforcement learning is developed for a Markov decision process (MDP), as shown in Figure 1. An *agent* monitors the *states* of the environment and performs an action a . The environment feeds back a scalar reward signal r and transits to a new state s' according to a probability distribution $p(r, s' | s, a)$. An agent can be an animal, a human, a robot, or a software. For animal agents, reward can be food, water, or pain. In humans, money or social fame can also be strong rewards.

The goal of the agent is to improve its action policy $P(a|s)$ so that the received reward is maximized in the long run. More specifically, the goodness of a policy is evaluated by the expected cumulative future rewards

$$E [r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots]$$

Figure 1

where $E[]$ represents the expectation (average) regarding the stochasticity of the environmental dynamics $p(r, s' | s, a)$ combined with the agent's policy

$p(a|s)$. The parameter γ is called the temporal discount factor and specifies how far into the future the agent is concerned with; only immediate reward r_t for $\gamma = 0$ and further into the future as γ increases closer to 1.

Under this framework, the aim of reinforcement learning can be formulated as finding the optimal policy, that maximizes the expected future rewards (1) starting from any state. What makes reinforcement learning interesting (and difficult) is that an action a_t does not only affect the immediate reward r_t , but also affects the next state s_{t+1} , which can affect the future rewards r_{t+1}, r_{t+2} , and so forth. Seen in another way, a given reward r_t may not be due to it immediately preceding action a_t , but may also be due to the past actions a_{t-1} , a_{t-2} , and so on. The problem of identifying which past actions at which states are responsible for a given reward is known as the temporal credit assignment problem, which is a major issue in reinforcement learning.

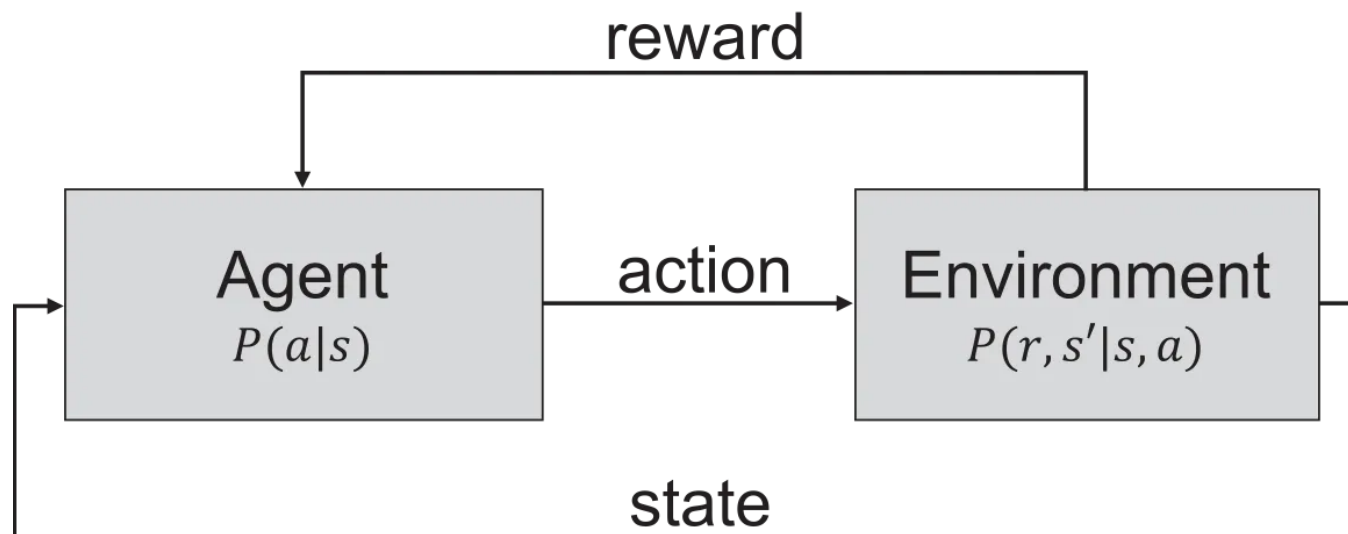


Figure 2: The interaction between the agent and the environment in reinforcement learning.

Another important problem in reinforcement learning is *exploration*. An agent should try different actions at different states to find out which is good or bad. As learning proceeds, the agent should take actions that are more likely to deliver more reward. How to balance between trying something new and focusing on known good choice is called *exploration-exploitation trade-off*.

. . .

Figure 3 shows a simple example which was used in a functional MRI study addressing the brain's mechanism of temporal discounting (Tanaka et al., 2004). It is an MDP with three states and two actions. Usually, the action $a = 1$ shifts the state to the left with a reward $r = 1$, and the action $a = 2$ shifts the state to the right with a negative reward of $r = -1$. However, from the leftmost state $s = 1$, the action $a = 1$ jumps the state to the rightmost $s = 3$ with a large negative reward $r = -5$, and from the rightmost state $s = 3$, the action $a = 2$ jumps the state to the leftmost $s = 1$ with a large positive reward of $r = 5$. Suppose you are at the middle state $s = 2$, which action would you take? If you simply follow a larger immediate reward, you would take $a = 1$ to get a positive reward, which moves you to $s = 1$, and then take $a = 2$ to avoid the large negative reward, which moves you back to $s = 2$. Thus, you will end up cycling between $s = 1$ and $s = 2$ with no net gain. A clever reader would take $a = 2$ at $s = 1$ and $s = 2$ despite immediate losses to reach $s = 3$ and then take $a = 2$ to get the larger reward. There are similar cases in real life that require costly work in order to achieve a valuable goal, such as publishing a paper or getting a PhD. Can a simple computational agent solve this task?

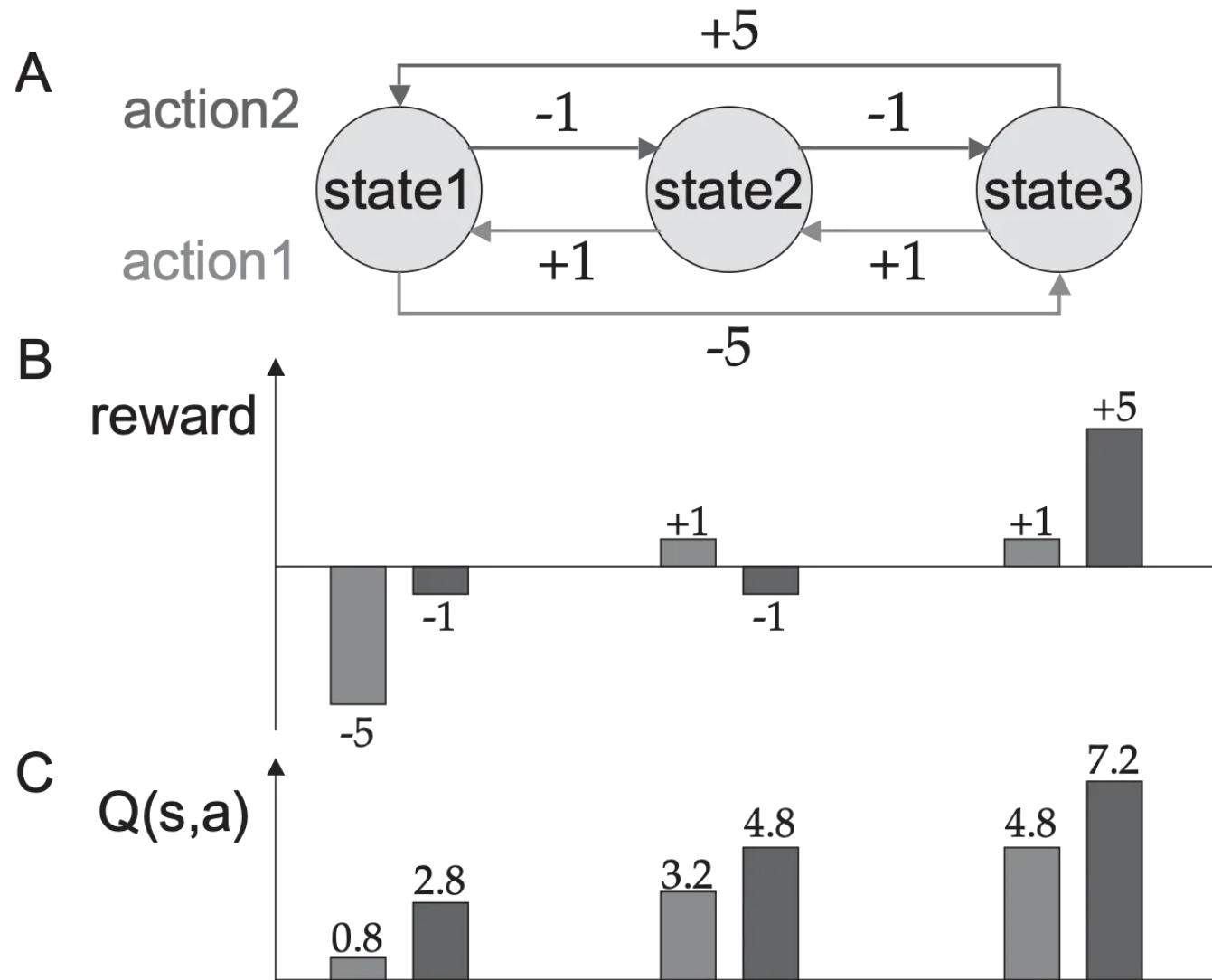


Figure 3: (A) A simple three-state Markov decision process (MDP) that requires going through immediate losses for long-term optimality (Tanaka et al., 2004). (B) The reward function and © the optimal action value function for this MDP (Doya, 2007).

• • •

Action Value Function

In order to evaluate the goodness of an action in a long run, a standard tool in reinforcement learning is the *action value function*, which is defined as

$$Q(s, a) \stackrel{\text{def}}{=} \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots | s_t = s, a_t = a]$$

The action value function $Q(s, a)$ evaluates how much future rewards the agent will get by taking an action a at state s , and then following the present policy. In psychology, it may be related to motivation or incentive to perform a certain action at a certain situation.

For an MDP with discrete states and actions, the action value function can be stored in a table of *states* \times *actions*, and its entries can be updated by a learning algorithm. For continuous or a very large number of states or actions, a function approximator like an artificial neural network (ANN) is used for representing the action value function (Mnih et al., 2015).

If the action value function has been learned for all the state-action pairs, the optimal policy is to select an action that maximizes the action value function at the present state:

$$a = \operatorname{argmax}_b Q(s, b)$$

which is called greedy policy. During learning, however, a policy has to be selected to promote exploration. A simple way is called ϵ -greedy policy, in which a random action is selected with probability ϵ and otherwise a greedy policy is taken.

Another common way of action selection using the action value function is Boltzmann or softmax selection:

$$p(a|s) = \frac{e^{\beta Q(s, a)}}{\sum_b e^{\beta Q(s, b)}}$$

where the action value function is regarded as a negative energy so that an action of larger action value is taken with higher probability. The parameter β

is called an inverse temperature and controls the randomness of choice. With $\beta = 0$, the choice is totally random and with increased β , the actions with higher action values are selected more frequently so that the choice becomes greedier.

. . .

Sarsa and Q Learning

How can an agent learn the action value function? In general, after experiencing sequences of state, action and reward, an average of discounted rewards following each state-action pair can be used as an estimate. This is called the Monte-Carlo method and is known to not be very efficient, especially when the environment dynamics are stochastic (SB, chapter 5). A more efficient way is to utilize the recursive relationship across subsequent states and actions:

$$Q(s_t, a_t) = E[r_t + \gamma Q(s_{t+1}, a_{t+1})]$$

which derives from the exponential discounting of future rewards.

The deviation from this recursive relationship can be detected by the temporal difference (TD) error:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$$

The action value function can then be updated as

$$Q(s_t, a_t) := Q(s_t, a_t) + \alpha \delta_t$$

where α is the learning rate parameter. This is known as the Sarsa algorithm, as it is based on the sequence of where α is the learning rate parameter. This is known as the Sarsa algorithm, as it is based on the sequence of $s_t, a_t, r_t, s_{t+1}, a_{t+1}$.

Another learning algorithm using the action value function is called Q-learning (Watkins, 1989; Watkins & Dayan, 1992) which uses a somewhat different TD error.

$$\delta_t = r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)$$

This means that a greedy policy is assumed from the subsequent state, even if the agent actually uses a non-greedy exploratory policy. This is called off-policy learning, while Sarsa is called on-policy learning. A benefit of off-policy learning is that the optimal value function with a deterministic policy can be learned while following a stochastic exploratory policy. Drawbacks of off-policy learning are that the performance during learning can be compromised by neglecting the effect of exploration and that learning can be unstable when combined with a function approximator (see SB, chapters 6 and 11).

. . .

Actor-Critic and State Value Function

Another class of reinforcement learning algorithm is called actor-critic architecture (Barto et al., 1983). The *actor* realizes some form of policy $p(a|s, \theta)$ with a parameter vector θ . The *critic* evaluates how well the actor's policy

is working. More specifically, the critic predicts the expected future reward from each state by following the present policy as the state value function:

$$V(s) \stackrel{\text{def}}{=} \mathbb{E}[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots | s_t = s]$$

For discrete states, the state value function can be stored in a vector, while a function approximator is used for continuous or a large number of states (Silver et al., 2016). In psychology, the state value function may be related to the prospect or mood a given situation delivers.

. . .

Dynamic Programming

The theory of Dynamic Programming provides the ways for using the reward and state transition functions to derive the *optimal value function* that an optimal policy should satisfy (Bellman, 1952)(SB, chapter 4). The recursive relationship of the state value function in Equation below can be expressed by the reward and the transition functions as

$$V(s) = \sum_a p(a|s) \left[r(s, a) + \gamma \sum_{s'} p(s'|s, a) V(s') \right].$$

This is called the *Bellman equation* for the policy $p(a|s)$. For an optimal policy, the state value function satisfies

$$V(s) = \max_a \left[r(s, a) + \gamma \sum_{s'} p(s'|s, a) V(s') \right].$$

This is called the *Bellman optimality equation* and its solution $V^*(s)$ is called the *optimal state value function*. Even though there can be multiple optimal policies, the optimal value function is unique. Once the optimal state value function is derived, an optimal policy is given by the action that maximizes the right-hand side of Equation above for each state.

The Bellman optimality equation is simultaneous nonlinear equations for the number of the states and solving it can be quite hard as the number of the states becomes large.

• • •

Action Planning

When the state transition dynamics is deterministic or near-deterministic, searching for a sequence of actions that gives a large cumulative reward is a realistic strategy. For a task that completes in a small number of steps, searching till the end of a sequence is possible. In a task with many steps, the action sequence search can be truncated by using an estimate of the state value function. For example, the expected reward for a two-step transition can be estimated as:

$$Q(s_0, a_0, a_1) = r(s_0, a_0) + \gamma \sum_{s_1} p(s_1 | s_0, a_0) \left[r(s_1, a_1) + \gamma \sum_{s_2} p(s_2 | s_1, a_1) V(s_2) \right].$$

In complex tasks like the game of Go, computing the optimal state value function for all possible states is intractable and searching through all possible action sequences till the end of the game requires an enormous amount of time. However, a good combination of an approximate value function and action search using a state transition model, such as the Monte Carlo tree search (MCTS) (Coulom, 2006)(see SB, chapter 8), can give practical solutions (Silver et al., 2016, 2018)(see SB, chapter 16).

The prediction of the future states in model-based action planning may be considered as the process of imagery or mental simulation.

• • •

Partially Observable Markov Decision Processes

The state transition model can be useful not only for planning future actions, but also for estimating the present state from previous actions when the sensory observation is subject to noise, delay, or occlusion. In the partially observable Markov decision process (POMDP); see SB, chapter 17), the agent receives stochastic observation of the environmental state as $p(o|s)$. A simple solution to POMDP is to learn a policy based on observation $p(a|o)$, but that is often suboptimal. When the agent has access to models of the sensory observation and state transition, it is possible to utilize the dynamic Bayesian framework to update the probabilistic estimate of the state. From the previous estimate of the state probability $p(s_{t-1})$ and the previous action a_{t-1} , the prior probability for the present state is given by the state transition model as

$$\sum_{s_{t-1}} p(s_t | s_{t-1}, a_{t-1}) p(s_{t-1})$$

This can be combined with the likelihood from the present observation $P(o_t | s_t)$ as

$$p(s_t | o_t, a_{t-1}) \propto p(o_t | s_t) \sum_{s_{t-1}} p(s_t | s_{t-1}, a_{t-1}) p(s_{t-1})$$

The posterior state probability $p(s_t | o_t, a_{t-1})$ is called *belief state* and can be iteratively used as the prior probability $p(s_t)$ for computing the next belief state.

A standard way of action choice under sensory uncertainty is to average the action values over possible states

$$\sum_s p(s) Q(s, a)$$

and take the action that maximizes it.

Identification of an underlying state from noisy observations is a central issue in sensory perception, or perceptual decision making, and human actions often reflect uncertainty or confidence in the perceived state.

. . .

Reinforcement Learning for Artificial Intelligence

There can be multiple approaches in creating intelligent machines. One is to analyze specific features of a given problem and come up with a domain-specific solution algorithm. Another is to mimic the skills of human experts. The third approach is to let machines discover a good solution by experience. Creating a machine that learns like a human has been a long-time dream of artificial intelligence (AI) researchers. The classic example is Samuel's checker player, which included the idea of propagating the board score across subsequent states (Samuel, 1959)(see SB, chapter 16). The modern form of TD learning was presented in (Barto et al., 1983), which demonstrated its performance by simulation of the task of cart-pole balancing. Watkins clarified the link between TD learning and dynamic programming and

derived the Q-learning algorithm (Watkins, 1989; Watkins & Dayan, 1992). The first practical demonstration of the strength of TD learning was TD-Gammon, which achieved world champion level performance (Tesauro, 1994).

. . .

Deep Reinforcement Learning

The most recent advance in reinforcement learning, and AI in general, is delivered by a combination of TD learning with deep neural networks. It has been shown that a combination of TD learning with function approximation can cause instability, because the update of the present value $V(s_t)$ can affect its target value $V(s_{t+1})$ as a side effect of generalization by the function approximator (Boyan & Moore, 1995; Tsitsiklis & Roy, 1997). Researchers at DeepMind discovered an approach to overcome this problem using two techniques (Mnih et al., 2015).

One is to keep a copy of the value function approximator network, called the target network for computing $V(s_{t+1})$ as in the TD error Equation, and update it only intermittently after the network for computing $V(s_t)$ has been updated

upon many state transitions. This avoids the inflation of the target value due to generalization over temporally adjacent states.

Another is to store the state-action-reward sequence in a memory and update the value function by randomly sampling state-action-reward-state experience from the memory, called experience replay. This avoids the difficulty in learning from temporally correlated samples. The benefit of experience replay, which has also been demonstrated in early works (Moore & Atkeson, 1993), was inspired by episodic memory mechanism of the hippocampus (Hassabis et al., 2017).

The effectiveness of the combination was demonstrated by the Deep Q-Network that takes the screen images of a computer game as the state input and the action values for the joystick and button operation as the output.

The strength of combination of TD learning with deep neural network was further demonstrated in the game of Go. In the original version of AlphaGo, learning was initially guided by the play records of a human expert (Silver et al., 2016). In the later versions, AlphaGo Zero (Silver et al., 2017), learning was solely based on the program's own simulated games. Furthermore, in Alpha Zero (Silver et al., 2018), the same algorithm achieved superhuman performances in Go, Chess, and Shogi.

• • •

Robotics

Creating a robot that can learn a variety of motor skills by trial and error has also been a dream of robotics researchers. Early efforts included building a robot that learns to walk or to stand up (Morimoto & Doya, 2001). Major issues in applying reinforcement learning to robots are the need of continuous, high-dimensional actions for fine movements and the time, cost, and danger involved with trial and error in physical environments.

The actor-critic and other algorithms using parameterized policy are commonly used for continuous control (Peters & Schaal, 2008). Using a physics simulator for early exploratory learning and then transferring to additional learning in real environments (sim-to-real) is also a common practice. Recently, the combination of deep learning with reinforcement learning is making advances in vision-based control tasks, such as the manipulation of a variety of objects (Gu et al., 2017).

• • •

Reinforcement Learning in the Brain

The concept of reinforcement learning originates from how animals learn behaviors. The developments of reinforcement learning algorithms provided some plausible mechanisms of how they might be realized in the brain. Indeed, in the last couple of decades, numerous advances have been made in the brain's mechanism of reinforcement learning.

Dopamine Coding of Temporal Difference Error

A breakthrough discovery regarding the brain's mechanism of reinforcement learning was that midbrain dopamine neurons respond to reward prediction error (Schultz, 1998; Schultz et al., 1993). Schultz and colleagues recorded dopamine neuron activities while monkeys performed tasks like reaching for food or pressing a lever for juice (Figure 4). Before learning or when there was no predictive cue, dopamine neurons responded to the reward. As the animal learned to associate a sensory cue to the delivery of reward, dopamine neurons started to respond to reward-predictive sensory cues and the response for the predicted reward was diminished. When the reward was omitted after learning, dopamine neuron firing was suppressed at the timing when reward delivery was expected. These are interesting findings on their

own, but most exciting for those who are familiar with reinforcement learning theory because it exactly matches what the TD error does.

Before learning, by assuming that the value function $V(s) = 0$ for all states, the TD signal δ_t is equal to the reward r_t . When a new state s_{t+1} allows the agent to predict the forthcoming reward, $V(s_{t+1})$ becomes positive and thus the TD error δ_t responds with a positive pulse even if the reward $r_t = 0$. When the predicted reward is presented, the value $V(s_{t+1})$ goes down to the baseline, so that the temporal difference $\gamma V(s_{t+1}) - V(s_t)$ becomes negative and cancels a positive reward r_t .

This parallel between the dopamine neuron activities and the TD signal inspired theoretical proposals that the dopamine neurons and their major projection target, the striatum, may implement TD-type reinforcement learning (Barto, 1995; Houk et al., 1995a; Montague et al., 1996; Schultz et al., 1997), as depicted in Figure 5 (below).

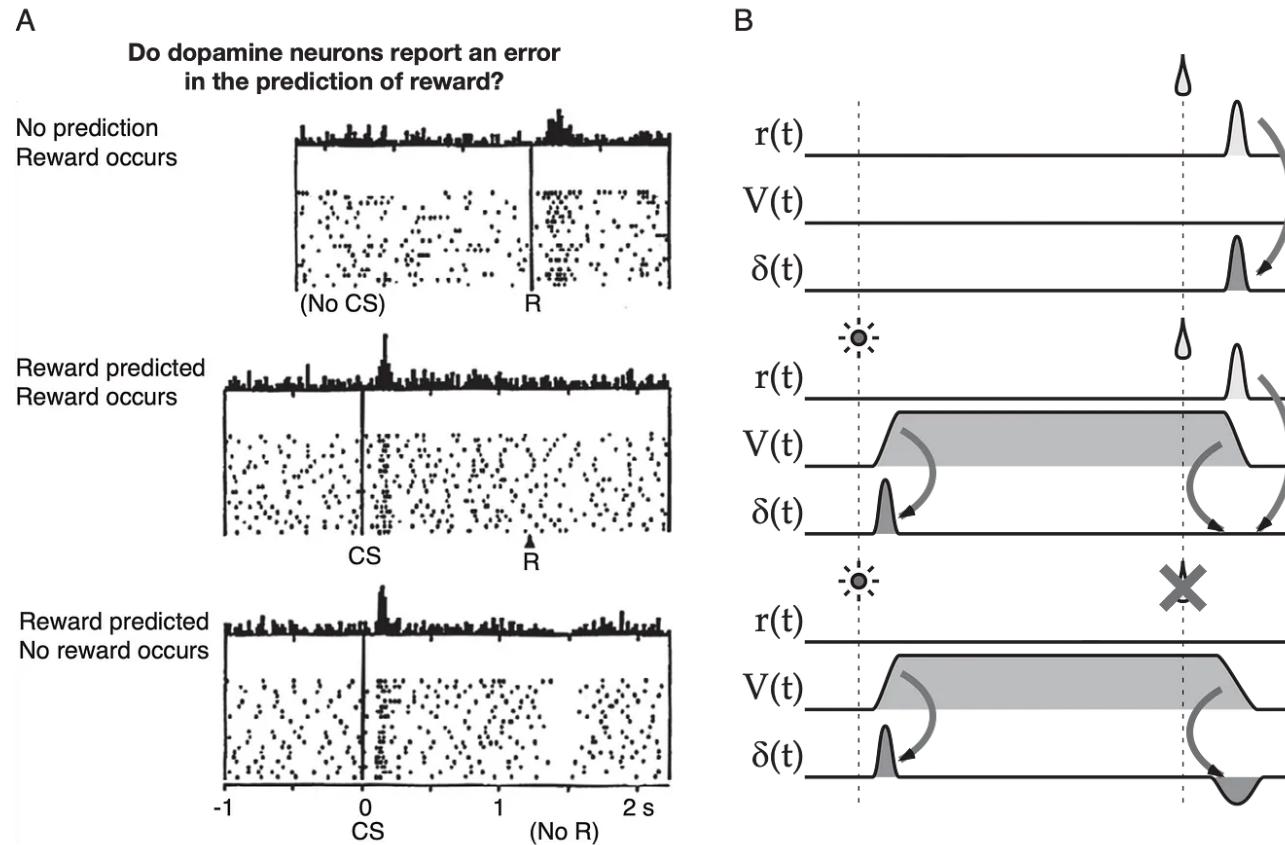


Figure 4: (A) The response of midbrain dopamine neurons to unpredicted reward, reward-predictive stimulus, and omitted reward (Schultz et al., 1997). (B) The dopamine neuron response coincides with the TD error signal in these cases.

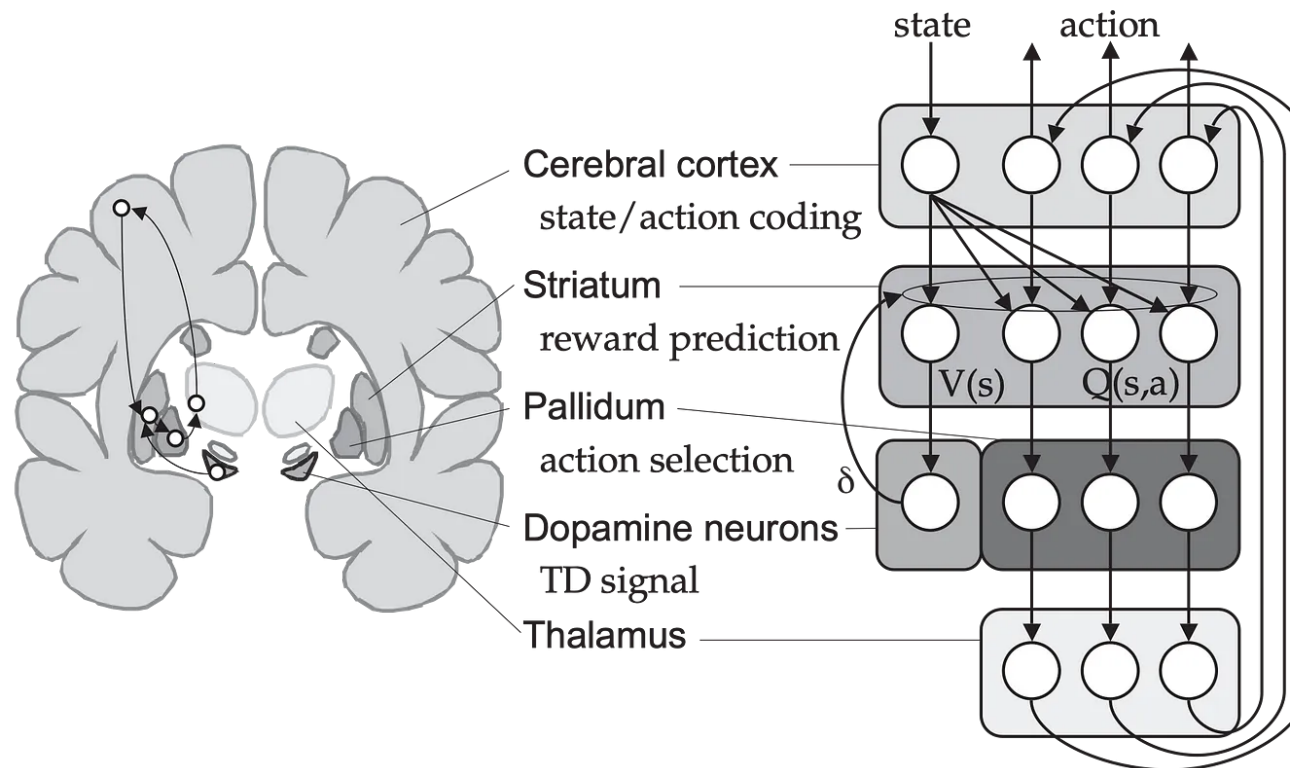


Figure 5: The anatomical organization of the basal ganglia (left) and their possible roles in reinforcement learning (right) (Doya, 1999, 2000).

More recently, Yagishita and colleagues investigated the dopamine-dependent synaptic plasticity using optical activation of presynaptic glutamate, postsynaptic activation by intracellular electrode, and optogenetic stimulation of dopamine terminals (Yagishita et al., 2014). In the striatal neurons expressing D1 type receptors, pre-post stimulation followed by dopamine input within about 1 second caused synaptic potentiation. In the striatal neuron expressing D2 type receptors, which has a higher affinity

(sensitivity) than D1 type receptors, the suppression of dopamine release caused synaptic potentiation (Iino et al., 2020).

. . .

Value and Action Coding in the Basal Ganglia

The TD error coding of the dopamine neurons and dopamine-dependent synaptic plasticity in the striatum strongly suggest that the basal ganglia play a major role in reinforcement learning in the brain (Houk et al., 1995b). The basal ganglia form parallel loop circuits with the input from the cerebral cortex and the output through the thalamus back to the cortex (Alexander & Crutcher, 1990). Given the dopamine-dependent synaptic plasticity, a specific hypothesis is that the striatal neurons are involved in learning state or action value functions (Figure 3). Samejima et al. showed in a free choice task that many of the striatal neurons represent action-specific reward prediction (Samejima et al., 2005).

In rodents, the cortico-basal ganglia loops are roughly divided into the motor loop through the dorsolateral striatum, the prefrontal loop through the dorsomedial striatum, and the limbic loop through the ventral striatum

(Voorn et al., 2004). Neural recording from the striatum of rats also showed action value coding neurons in the dorsal striatum and state-value coding neurons in the ventral striatum (Ito & Doya, 2015).

The striatum is composed of two compartments, the striosome projecting to the midbrain dopamine neurons and the matrix (or patch) projecting to the globus pallidus (Gerfen, 1992; Graybiel & Ragsdale, 1978). The globus pallidus is composed of the internal segment (GPi) that projects to the thalamus and the external segment (GPe) that projects to GPi both directly and through the subthalamic nucleus (STN), which receive inputs from the cortex. The cortical input through the basal ganglia has three pathways: the direct pathway through the striatum to GPi; the indirect pathway through the striatum, GPe, and subthalamic nucleus (STN) to GPi; and the hyperdirect pathway through STN to GPi (Nambu et al., 2002). What is the reason for such multiple pathways?

Recently, genetically encoded calcium indicators (GECI) and optogenetic manipulation enabled cell-type specific recording and manipulation of striatal neurons. In rodent striatum, D1-receptor-expressing neurons project to the direct pathway causing double inhibition, while D2-receptor-expressing neurons project to the indirect pathway involving triple inhibition. They have been hypothesized to be involved in action initiation and

suppression (Alexander & Crutcher, 1990; Delong, 1990), or learning from reward and punishment (Frank et al., 2004; Hikida et al., 2010).

Optogenetic stimulation of D1-receptor-expressing, direct pathway neurons in the dorsomedial striatum induced reinforcing effect, while stimulation of D2receptor-expressing, indirect pathway neurons induced aversive effect (Kravitz et al., 2012). Intriguingly, measurement of population activities of D1 and D2 striatal neurons by fiber photometry showed that both populations are activated at the onset of actions (Cui et al., 2013). This may be because the start of a new action is often the end of the previous action. In a sequential lever press task of repeating components (e.g., LLRR), optogenetic activation of D1 neurons induced over repetition (e.g., LLLRR) while activation of D2 neurons induced premature transition (e.g., LRR), suggesting that they are involved in sticking and switching, respectively (Geddes et al., 2018).

. . .

Model-Free/Model-Based Action and Learning

Human and animal behaviors can be classified as goal-directed, depending on the present needs, or habitual, responding routinely to given stimuli. These behaviors are dissociated by a devaluation paradigm, in which the value of a particular food is changed by satiation or poisoning. Balleine and colleagues demonstrated that the prefrontal-dorsomedial striatal loop and the motordorsolateral striatal loop are respectively involved in goal-directed and habitual behaviors (Balleine et al., 2007). Daw and colleagues further postulated that goal-directed and habitual behaviors are based on model-based predictive search and model-free reactive choice (Daw et al., 2005). While model-based strategies are often attributed to the prefrontal and the parietal cortex (Glascher et al., 2010), functional MRI studies suggested the involvement of the basal ganglia also (Daw et al., 2011) (Figure 6, below). Another study using multistep action planning showed activation of not only the cortical areas but also the cerebellum and the basal ganglia (Fermin et al., 2016), which is consistent with the view that the cerebellum predicts the resulting state of action candidates using internal models acquired by supervised learning and that the basal ganglia evaluates their goodness by the value function acquired by reinforcement learning (Doya, 1999, 2000).

The dichotomy between model-free and model-based systems has some resemblance to other dichotomies in psychology and cognitive science (Dayan, 2009), such as procedural versus declarative, System 1 versus

System 2 (Kahneman, 2011; Kahneman & Tversky, 1979), and unconscious and conscious (Bengio, 2017).

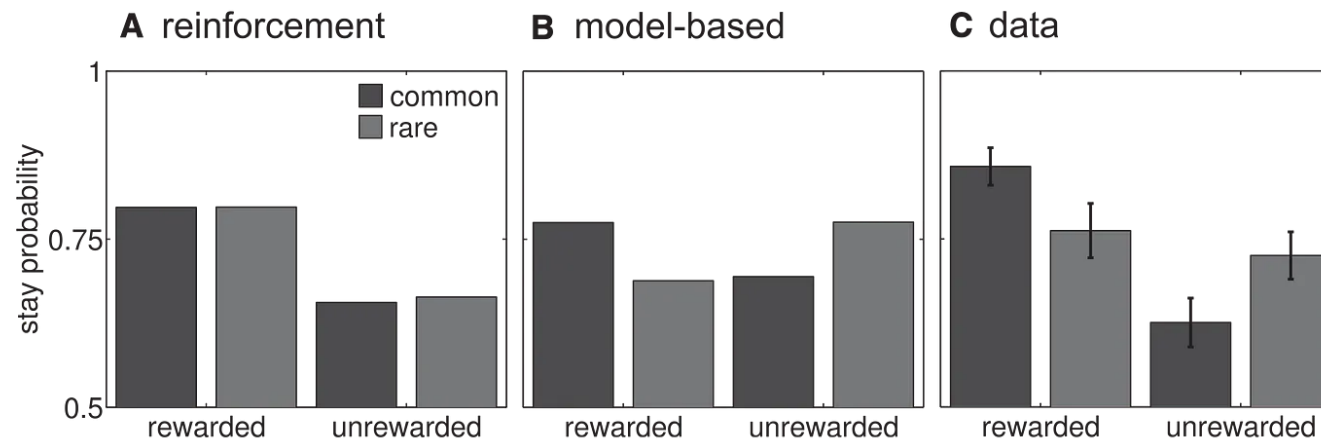
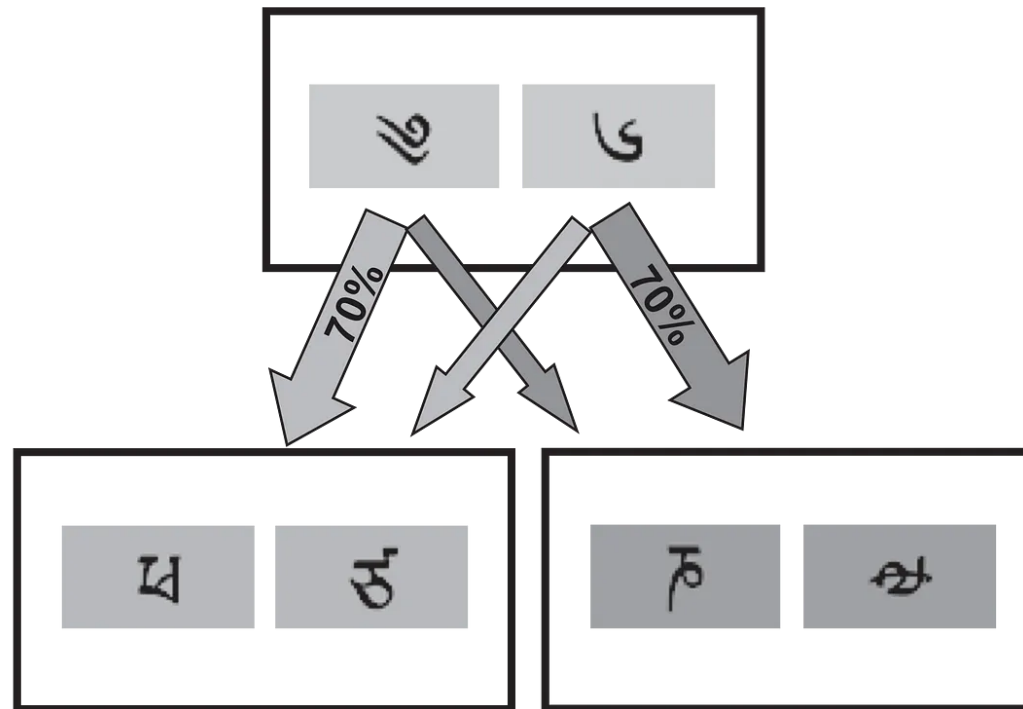


Figure 6: The “two-step task” used for dissociating model-free and model-based learning (Daw et al., 2011). If a reward is acquired after a rare transition in the first step, a model-free agent would repeat the same action, while a model-based agent would choose another action to reach to the rewarded state in the second step with a higher probability. Actual subjects tend to be between the two.

• • •

Conclusion

Reinforcement learning is a theoretical framework that has promoted fruitful interactions across neuroscience, psychiatry, psychology, sociology, and economics. This is because the problem setup of reinforcement learning captures the basic features of animal and human behaviors.

There are presently several major challenges and limitations in reinforcement learning algorithms. One is sample efficiency, meaning that learning requires a lot of data. In tasks where simulators are available, a computer agent can have limitless interactions with a stationary environment. The success of AlphaGo is based on a huge number of game plays that any human player cannot experience in a lifetime (Silver et al., 2017). In real physical environments, such as robot control or human interaction, taking actual experience can be time consuming or costly, and the environment can keep changing so that slow learners cannot catch up. Another challenge is representation learning. Efficient reinforcement learning requires good representation of states and actions. Deep reinforcement learning gives one

solution to representation learning for reinforcement learning (Mnih et al., 2015), but that still suffers from sample efficiency.

Development of robust and flexible reinforcement learning algorithms may provide helpful models for understanding the sophisticated reinforcement learning mechanisms in the brain. Also, understanding of how such algorithms can fail in certain conditions may shed light on the complex pathology of psychiatric disorders (Montague et al., 2012; Redish & Gordon, 2016).

The basal ganglia are by no means the sole locus of reinforcement learning in the brain. Even small brains of worms or flies should have the capability for reinforcement learning (Bendesky et al., 2011; Yamagata et al., 2014). In the vertebrate brain, the amygdala is also known to be critical for learning from reward and punishment (Belova et al., 2007). Recent developmental study revealed that the lateral amygdala neurons have the same origin as those of the cortex, while the central amygdala neurons have their origin as basal ganglia neurons (Soma et al., 2009). The amygdala is an evolutionarily older brain structure than the basal ganglia; it may be considered as a prototype of the cortico-basal ganglia circuit (Cassell et al., 1999). Reward-dependent activities are also found in a variety of cortical areas, such as the orbitofrontal cortex (Schultz et al., 2000), the prefrontal cortex (Matsumoto et al., 2003;

Watanabe, 1996), and the parietal cortex (Dorris & Glimcher, 2004; Platt & Glimcher, 1999; Sugrue et al., 2004). The computation of state, value, and action may not happen step-wise in separate brain areas but may be realized by the dynamics of the cortico-basal ganglia loop (Cisek, 2007).

. . .

References (in order of citation):

- Thorndike, E. L. (1898). *Animal intelligence: an experimental study of the associate processes in animals*. Psychological Review, Monograph Supplements, 2(8), 1–109.
- Barto, A. G., Sutton, R. S., & Andersen, C. W. (1983). *Neuronlike adaptive elements that can solve difficult learning control problems*. IEEE Transactions on Systems, Man, and Cybernetics, 13(5), 834–846.
- Sutton, R. S., & Barto, A. G. (2018). **Reinforcement Learning: An Introduction** (2nd ed.). Cambridge, MA: MIT Press.
- Barto, A. G. (1995). *Adaptive critics and the basal ganglia*. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), **Models of Information**

Processing in the Basal Ganglia, (pp. 215–232). Cambridge, MA: MIT Press.

- Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). *Bee foraging in uncertain environments using predictive Hebbian learning*. *Nature*, 377, 725–728.
- Schultz, W., Dayan, P., & Montague, P. R. (1997). *A neural substrate of prediction and reward*. *Science*, 275, 1593–1599.
<https://doi.org/10.1126/science.275.5306.1593>
- Doya, K. (2007). *Reinforcement learning: computational theory and biological mechanisms*. *Frontiers in Life Science*, 1(1), 30–40.
<https://doi.org/10.2976%2F1.2732246>
- Glimcher, P. W., & Fehr, E. (2013). **Neuroeconomics: Decision Making and the Brain** (2nd ed.). London: Elsevier Academic Press.
- Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). *Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops*. *Nature Neuroscience*, 7(8), 887–893.
- Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). *Human-level control through deep reinforcement learning*. *Nature*, 518(7540), 529–533.

- Watkins, C. J. C. H. (1989). **Learning from delayed rewards**. Ph.D. Thesis, University of Cambridge.
- Watkins, C. J. C. H., & Dayan, P. (1992). Q-Learning. *Machine Learning*, 8(3–4), 279–292.
- Silver, D., Huang, A., Maddison, C. J., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489.
- Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 38, 716–719.
- Coulom, R. (2006). Efficient selectivity and backup operators in Monte-Carlo tree search. 5th International Conference on Computer and Games. Turin, Italy.
- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3, 210–229.
- Tesauro, G. (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6, 215–219.

- Boyan, J. A., & Moore, A. W. (1995). Generalization in reinforcement learning: safely approximating the value function. In T. K. Leen (Ed.), **Advances in Neural Information Processing Systems 7** (pp. 369–376). Cambridge, MA: MIT Press.
- Tsitsiklis, J. N., & Roy, B. V. (1997). An analysis of temporal-difference learning with function approximation. IEEE Transactions on Automatic Control, 42, 674–690.
- Silver, D., Schrittwieser, J., Simonyan, K., et al. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354–359.
- Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: reinforcement learning with less data and less time. Machine Learning, 13(1), 103–130.
- Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. Neuron, 95(2), 245–258.
- Silver, D., Hubert, T., Schrittwieser, J., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362 (6419), 1140–1144.
<https://doi.org/10.1126/science.aar6404>.
- Morimoto, J., & Doya, K. (2001). Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning. Robotics and

Autonomous Systems, 36, 37–51.

- Peters, J., & Schaal, S. (2008). Reinforcement learning of motor skills with policy gradients. Neural Networks, 21(4), 682–697.
- Gu, S., Holly, E., Lillicrap, T., & Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In IEEE International Conference on Robotics and Automation (ICRA 2017).
- Schultz, W. (1998). Predictive reward signal of dopamine neurons. Journal of Neurophysiology, 80, 1–27.
- Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. Journal of Neuroscience, 13, 900–913.
- Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), **Models of Information Processing in the Basal Ganglia**, (pp. 215–232). Cambridge, MA: MIT Press.
- Houk, J. C., Adams, J. L., & Barto, A. G. (1995a). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.),

Models of Information Processing in the Basal Ganglia, (pp. 249–270). Cambridge, MA: MIT Press.

- Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012). Computational psychiatry. Trends in Cognitive Sciences, 16(1), 72–80.
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275, 1593–1599.
- Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science, 345(6204), 1616–1620.
- Iino, Y., Sawada, T., Yamaguchi, K., et al. (2020). Dopamine D2 receptors in discrimination learning and spine enlargement. Nature (online).
- Houk, J. C., Adams, J. L., & Barto, A. G. (1995b). **Models of Information Processing in the Basal Ganglia**. Cambridge, MA: MIT Press.
- Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. Trends in Neuroscience, 13, 266–271.

- Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. Science, 310(5752), 1337–1340.
- Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. M. (2004). Putting a spin on the dorsal-ventral divide of the striatum. Trends in Neurosciences, 27(8), 468–474.
- Gerfen, C. R. (1992). The neostriatal mosaic: multiple levels of compartmental organization in the basal ganglia. Annual Review of Neuroscience, 15, 285–320.
- Graybiel, A. M., & Ragsdale, C. W., Jr. (1978). Histochemically distinct compartments in the striatum of humans, monkeys, and cats demonstrated by acetylthiocholinesterase staining. Proceedings of the National Academy of Sciences, 75(11), 5723–5726.
- Nambu, A., Tokuno, H., & Takada, M. (2002). Functional significance of the cortico–subthalamo–pallidal ‘hyperdirect’ pathway. Neuroscience Research, 43(2), 111–117.
- DeLong, M. R. (1990). Primate models of movement disorders of basal ganglia origin. Trends in Neurosciences, 13, 281–285.
- Frank, M. J., Seeberger, L. C., & O’Reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science,

306(5703), 1940–1943.

- Hikida, T., Kimura, K., Wada, N., Funabiki, K., & Nakanishi, S. (2010). *Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior*. *Neuron*, 66(6), 896–907.
- Kravitz, A. V., Tye, L. D., & Kreitzer, A. C. (2012). *Distinct roles for direct and indirect pathway striatal neurons in reinforcement*. *Nature Neuroscience*, 15(6), 816–818.
- Cui, G., Jun, S. B., Jin, X., et al. (2013). *Concurrent activation of striatal direct and indirect pathways during action initiation*. *Nature*, 494(7436), 238–242.
- Geddes, C. E., Li, H., & Jin, X. (2018). *Optogenetic editing reveals the hierarchical organization of learned action sequences*. *Cell*, 174(1), 32–43, e15.
- Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). *The role of the dorsal striatum in reward and decision-making*. *Journal of Neuroscience*, 27(31), 8161–8165.
- Daw, N. D., Niv, Y., & Dayan, P. (2005). *Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control*. *Nature Neuroscience*, 8(12), 1704–1711.

- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69 (6), 1204–1215.
- Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4), 585–595.
- Fermin, A. S., Yoshida, T., Yoshimoto, J., Ito, M., Tanaka, S. C., & Doya, K. (2016). Model-based action planning involves cortico-cerebellar and basal ganglia networks. *Scientific Reports*, 6, 31378.
- Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks*, 22(3), 213–219.
- Kahneman, D. (2011). **Thinking, Fast and Slow**. New York, NY: Farrar, Straus and Giroux.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2), 263–291.
- Bengio, Y. (2017). The consciousness prior. <https://doi.org/10.48550/arXiv.1709.08568>.

- Redish, A. D., & Gordon, J. A. (2016). **Computational Psychiatry**. Cambridge, MA: MIT Press.
<https://doi.org/10.7551/mitpress/9780262035422.001.0001>
- Bendesky, A., Tsunozaki, M., Rockman, M. V., Kruglyak, L., & Bargmann, C. I. (2011). Catecholamine receptor polymorphisms affect decision-making in C. elegans. Nature, 472(7343), 313–318.
<https://doi.org/10.1038/nature09821>
- Yamagata, N., Ichinose, T., Aso, Y., et al. (2014). Distinct dopamine neurons mediate reward signals for short- and long-term memories. Proceedings of the National Academy of Sciences (online).
<https://doi.org/10.1073/pnas.1421930112>
- Belova, M. A., Paton, J. J., Morrison, S. E., & Salzman, C. D. (2007). Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. Neuron, 55(6), 970–984.
<https://doi.org/10.1016/j.neuron.2007.08.004>
- Soma, M., Aizawa, H., Ito, Y., et al. (2009). Development of the mouse amygdala as revealed by enhanced green fluorescent protein gene transfer by means of in utero electroporation. Journal of Comparative Neurology, 513(1), 113–128.
<https://doi.org/10.1002/cne.21945>

- Cassell, M. D., Freedman, L. J., & Shi, C. (1999). *The intrinsic organization of the central extended amygdala*. Annals of New York Academy of Sciences, 877, 217–240.
- Cisek, P. (2007). *Cortical mechanisms of action selection: the affordance competition hypothesis*. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1485), 1585–1599.
<https://doi.org/10.1098/rstb.2007.2054>
- Matsumoto, K., Suzuki, W., & Tanaka, K. (2003). *Neuronal correlates of goal-based motor selection in the prefrontal cortex*. Science, 301(5630), 229–232.
- Watanabe, M. (1996). *Reward expectancy in primate prefrontal neurons*. Nature, 382, 629–632.
- Platt, M. L., & Glimcher, P. W. (1999). *Neural correlates of decision variables in parietal cortex*. Nature, 400, 233–238.
- Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). *Matching behavior and the representation of value in the parietal cortex*. Science, 304(5678), 1782–1787.
<https://doi.org/10.1126/science.1094765>

. . .

Thank you so much for your kind attention! Stay tuned for the next endeavor!

. . .

Alireza Dehbozorgi

alirezadehbozorgi83@yahoo.com

alireza@lingoai.io

<https://www.linkedin.com/in/alireza-dehbozorgi-8055702a/>

Twitter: @BDehbozorgi83

Cognitive Science

Computation

Statistical Learning

Machine Learning

Mathematical Psychology



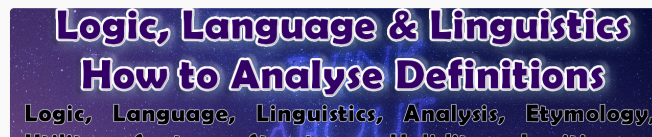
Written by Alireza Dehbozorgi

[Edit profile](#)

97 Followers

NMT researcher at <https://lingoai.io/> | Content Creator at <https://www.deks.app/> | <https://twitter.com/BDehbozorgi83> | LinkedIn: alireza-dehbozorgi-8055702a/

More from Alireza Dehbozorgi



Alireza Dehbozorgi

Unleashing the Power of Mathematical Linguistics: Part I...



Alireza Dehbozorgi

Unleashing the Power of Mathematical Linguistics: A...

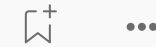
Language and meaning have an intrinsic logical structure that can be captured...

4 min read · 3 days ago



Language, as the pinnacle of human communication, has long intrigued linguist...

3 min read · 5 days ago



Alireza Dehbozorgi

Modern Schools of Linguistics (Resources)

Part I: Generativism

3 min read · Oct 31, 2022

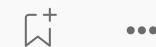


Alireza Dehbozorgi

Some Notes on Fourier Analysis


Introduction

4 min read · Oct 10, 2022



See all from Alireza Dehbozorgi

Recommended from Medium

 May Pang in Better Humans

Statistically, You Will Marry the Wrong Person. Here's Why.

Luckily, statistics also tell you how not to.

★ · 9 min read · Jul 14



5.3K



107



 Christina Sa in UX Planet

The UX Design Case Study That Got Me Hired

Getting a job in UX design is tough, but one particular case study helped me stand out...

★ · 8 min read · Mar 16



8.5K



111



Lists

Predictive Modeling w/ Python

18 stories · 139 saves

Natural Language Processing

417 stories · 62 saves

Practical Guides to Machine Learning

10 stories · 152 saves

The New Chatbots: ChatGPT, Bard, and Beyond

13 stories · 55 saves

 Carlyn Beccia  in Sexography

The Latest Studies on Sexual Attraction Would Make Darwin...



But it explains why evolutionary psychology is problematic.

🌟 · 6 min read · Jul 10

 4.97K

 73



 Michal Malewicz 

There are FIVE levels of UI skill.


Only level 4+ gets you hired.

🌟 · 6 min read · Apr 25

 4.6K

 53



 Unbecoming

How My Husband's Sexuality Defined Our Marriage

And the lies I told myself

★ · 5 min read · May 17



7.7K



73

 Microsoft Design in Microsoft Design

A change of typeface: Microsoft's new default font has arrived

Introducing Aptos, our modern successor to Calibri

5 min read · Jul 13



3.9K



79



See more recommendations