Reinforcement Learning: A Biological Perspective

Intermediate article

P Read Montague, Baylor College of Medicine, Houston, Texas, USA
David M Eagleman, Salk Institute, La Jolla, California, USA
Samuel M McClure, Baylor College of Medicine, Houston, Texas, USA
Gregory S Berns, Emory University School of Medicine, Atlanta, Georgia, USA

CONTENTS

Introduction
The three basic components of reinforcement learning
Trial and error
TD learning
Midbrain dopamine systems
Predictability, reward and human brain response
Summary

Reinforcement learning is an approach to learning problems that takes account of the interaction of the learner with their reactive environment. Despite its early connections to animal learning, reinforcement learning has been an active field of study in engineering and computer science. New research is now starting to connect reinforcement learning to identifiable biological systems. These connections may well provide direct insights into goal-directed learning and decision-making in biological organisms.

INTRODUCTION

Brain research during the last 50 years has focused on the metaphor of learning through interactions with an environment. The central idea here is interaction. All mobile creatures, including humans, are embedded in a rich, reactive environment. We move our eyes – the visual scene changes. We push against a tree branch – we feel its roughness, and it even pushes back. We shift our attention – suddenly a sight or sound becomes clearer and easier to interpret. Even our bodies act as part of the reactive environment, as mere consideration of an action can cause a change in an animal's physiological state whether or not it actually moves. The learning mechanisms that are embedded in our nervous systems have evolved to deal with just such reactive aspects of environments.

Reinforcement learning takes account of the idea that learners are situated in real-world settings which react to and may even adapt to the actions of the learner. In the words of modern reinforcement learning pioneers Sutton and Barto 'Reinforcement learning is learning what to do – how to map situations to actions – so as to maximize a numerical reward signal.'

In this framework, learning is specified as a complete problem of an agent interacting with an environment. Thus the properties of the agent and the environment and the dynamics of interaction must be specified in a reinforcement learning problem. As a class of problems, reinforcement learning problems are most closely related to mathematical problems of optimal control. One of the most important ideas that is built into all reinforcement learning systems is the concept of a *goal-seeking agent*. In reinforcement learning, the learner possesses goals that influence their selection of actions, the way they update their memory, and so on. From a biological perspective, the critical issue is to determine the physical mechanisms that define and control the goals and actions of the learner.

In this article we shall outline one computational approach to reinforcement learning and its biological realization in living systems. Reinforcement learning appeals to the computational community because it can be formally written down as equations or simulations. It appeals to biologists because of its plausibility as an architecture and its growing connection to real biological data. The discovery that reinforcement learning is a powerful approach to machine-learning should come as no surprise. Biological systems appear to have long exploited the approach to solve many problems of real-time adaptation in complex environments.

THE THREE BASIC COMPONENTS OF REINFORCEMENT LEARNING

Every reinforcement learning system has three basic components: (1) a reward function, (2) a value function and (3) a policy. These relatively abstract terms capture the idea of immediate evaluation (reward function), long-term judgment (value function) and action selection (policy).

The *reward function* formalizes the idea of a goal for a reinforcement learning system. It assigns to each state of the agent a single numerical quantity – the reward. The reward function defines what is good 'right now', and can be viewed as a built-in assessment of each state that is available to the agent (learner). It is also used to define the agent's goal – to maximize the total reward.

The *value function* formalizes the notion of longer-term assessments (judgments) of each state of the agent. It provides a valuation of the current state of the agent, taking into account the succession of states that could follow. Formally, for each state, value is defined as the total amount of reward that the agent can expect from that state onward into the distant future. These values would have to be stored in some fashion within the agent. In practice, the learner uses the reward function to improve their internal estimate of the value function.

In shorthand, rewards are immediate and values are long-term. For example, a rat may take many steps across an electrified grid (low reward) to reach food (high reward). All of those intermediate states (steps on the grid) have a very low reward, but possess high value because they lead directly to future states with food (high reward).

A *policy* formalizes exactly what the word implies – 'given this, do that'. Formally, a policy maps states to actions. In both biological and machine-learning examples, a policy is usually probabilistic. For a given state, a policy defines the probability of taking one of many possible actions in order to end up in one of many succeeding states. For example, a rat seeking food in a maze encounters a three-way junction. The policy assigns probabilities separately to each of the three actions that are available to the rat.

TRIAL AND ERROR

The idea of trial-and-error learning is familiar to everyone – to solve a problem, try some solution, assess how well it performed, and try again if it did not perform up to expectations. Reinforcement is intimately linked to trial-and-error learning. The basic idea of reinforcement derives from psychology, took its clearest form initially as Thorndike's law of effect, and depends on the concept of reinforcers. There are two types of reinforcers – positive and negative. The law of effect is simple. Actions or internal states followed by positive reinforcers are later more likely to occur, and actions or internal states followed by negative reinforcers are later less likely to occur. In one form or another, this idea has become one basis (almost a principle) for thinking about trial-and-error learning. An animal tries an action, assesses its success in terms of positive and negative reinforcement, and adjusts its later likelihood of taking that action.

The idea of trial-and-error learning is also central to reinforcement learning. Trial-and-error learning systems have been investigated from the earliest days of artificial intelligence. In fact, in 1954 one of the pioneers of artificial intelligence, Marvin Minksy, wrote his PhD thesis (*Theory of Neural-Analog Reinforcement Systems and its Applications to the Brain-Model Problem*), at Princeton University on reinforcement learning. In addition to his many other contributions, Minsky was one of the first to specify clearly the central problem in trial-and-error learning where some type of reward signal is used to criticize the outcome of a series of actions. This problem is known as the *temporal credit assignment problem*.

Temporal credit assignment is intuitively familiar. For example, an animal seeking food (high reward) may make many decisions before actually acquiring the food and receiving a large reward signal. A natural question then arises. How does the animal's brain assign credit to each of the individual decisions? Some decisions are critical to eventually obtaining food, while others may have been completely irrelevant. This type of problem arises whenever the rewards received are delayed in time from the events that lead to reward. In any real-world setting, an animal must *learn associations through time* (e.g. 'I go left here' is associated with 'I get three units of reward' an hour later).

Depending on the situation, there are numerous ways to solve the temporal credit assignment problem. One way to achieve trial-and-error learning using a delayed reward signal is *temporal difference (TD) learning*.

TD LEARNING

Here we shall describe in detail the way in which temporal difference (TD) learning frames and solves the temporal credit assignment problem described above. We shall begin with a description in terms of animal learning – that is, at the level of behavioral learning. In the later sections of this article we shall show that this same formal description can account for the pattern of activity that is recorded in midbrain dopamine neurons while alert animals are actively learning. The latter connection is important because it suggests that dopamine neurons may be one part of a mechanism that implements reinforcement learning in mammals.

Animals are taken as living in what is technically known as a Markov decision problem. This construct formalizes the characteristics of problems such as maze tasks, in which there are different states (e.g. different locations in the maze), actions (e.g. directions in which to move) and rewards. The rewards can either depend on the animal's actions or be provided without regard to what the animal does (as in classical conditioning). This framework is general enough to model many standard behavioral tasks. The most important assumptions underlying the TD approach involve the nature of the presumed computational task solved by an organism. There are two main assumptions in TD.

Assumption 1

First, the computational goal of learning is to use a set of sensory cues $\boldsymbol{x}(t) = \{x_1(t), {}_2(t), x_3(t), \ldots\}$ (e.g. characterizing the current state of an organism) to fit a 'value' function $V^*(\boldsymbol{x}(t))$ that 'values' the current state as the *average discounted sum of all future rewards from time t onward*:

$$V^*(\boldsymbol{x}(t)) = E\{\gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + \cdots\}$$

E is the expected value operator (the average), $r(t)$ is the reward at time t, $r(t+1)$ is the reward at time $t+1$, and so on; γ is a discount factor that ranges between 0 and 1 and captures the idea that rewards in the near future are more valuable than rewards in the distant future. If the true (optimal) $V^*(\boldsymbol{x}(t))$ could be estimated by a system, then the system could use such an estimate to update its internal model of future rewards and future actions predicated on the expected receipt of those rewards. This would give the system a way to simulate possible future action sequences and value them according to their expected long-term returns. In this description, the indices on the x's denote different sensory cues.

Assumption 2

This is the Markovian assumption – that is, the appearance of future sensory cues and rewards depends only on the immediate (current) sensory cues and not on the past sensory cues. This is a relatively restrictive assumption about the structure of the environment. However, it has proved to be useful even in real-world settings.

Adjusting the Predictions (Weights)

The strategy of TD learning is to use a set of sensory cues $\boldsymbol{x}(t) = \{x_1(t), x_2(t), x_3(t), \ldots\}$ present in a learning trial along with a set of adaptable weights $\boldsymbol{w}(t) = \{w_1(t), w_2(t), w_3(t), \ldots\}$ to make an estimate $V(\boldsymbol{x}(t))$ of the true $V^*(\boldsymbol{x}(t))$. In this formulation, the weights act as predictions of future reward. For completeness, we shall include a remark about the weights. The weight associated with each sensory cue (e.g. $w_1(t)$ associated with sensory cue 1) is actually a collection of weights – one for each time point following the appearance of sensory cue 1.

Local Data Anticipate Long-Term Reward

The difficulty in actually adjusting weights to estimate $V(\boldsymbol{x}(t))$ is that the system (i.e. the animal) would have to wait to receive all of its future rewards in a trial $r(t+1), r(t+2), r(t+3)$ etc. in order to assess its predictions. The latter constraint would require the animal to remember over time which weights need changing and which weights do not. Fortunately, there is information available at each instant in time that can act as a surrogate prediction error. This possibility is implicit in the definition of $V^*(\boldsymbol{x}(\boldsymbol{t}))$, since it satisfies a condition of consistency through time:

$$V^*(\boldsymbol{x}(t)) = E\ [r(t) + \gamma V^*(\boldsymbol{x}(t+1))]$$

Since the estimate V satisfies the same condition, an error δ in the estimated value function V (estimated predictions) can now be defined using information available at successive timesteps (i.e. taking the difference between both sides of the above equation and ignoring the expected value operator E for clarity).

$$\delta(t) = r(t) + \gamma V(\boldsymbol{x}(t+1)) - V(\boldsymbol{x}(t))$$

δ is called the *TD error* and it acts as a surrogate *prediction error* signal which is instantly available at time $t+1$. If the estimated predictions are correct, then $V(\boldsymbol{x}(t)) = V^*(\boldsymbol{x}(t))$, and the average prediction error is 0 (i.e. $E[\delta(t)] = 0$). In other words, if the system can adjust its weights (predictions) appropriately, then it can learn to expect future rewards predicted by the collection of sensory cues.

This section is somewhat technical, but it does emphasize the fact that local information can be used to make estimates of reward in the distant future. That is, it presents a method by which the temporal credit assignment problem is solved. It is not the only method for such a task, but its importance has recently been enhanced because this theoretical description of TD learning appears to match the information that is encoded in the spike trains of midbrain dopamine neurons.

MIDBRAIN DOPAMINE SYSTEMS

In the mammalian brain, midbrain dopamine systems play a major role in reward processing and reward-dependent learning. Recent physiological and computational research strongly suggests that the information which is constructed and broadcast by midbrain dopamine systems represents a prediction error analogous to that described above.

Some of the basic electrophysiological findings are shown in Figure 1. After repeated pairings of visual and auditory cues followed by reward, dopamine neurons change the time of their phasic activation from just after the time of reward delivery to the time of cue onset. In one task, a naive monkey is required to touch a lever following the appearance of a small light. Before training, most dopamine neurons show a short burst of impulses following reward delivery. After training, the animal learns to reach for the lever as soon as the light is illuminated, and this behavioral change correlates with two remarkable changes in the dopamine neuron output. First, the primary reward no longer elicits a phasic response, and secondly, the onset of the (predictive) light now causes a phasic activation in dopamine cell output.

In trials where the reward is not delivered at the appropriate time following the illumination of the light, dopamine neurons are depressed dramatically below their basal firing rate at precisely the time when the reward should have been delivered. This well-timed decrease in spike output shows that the expected time of reward delivery based

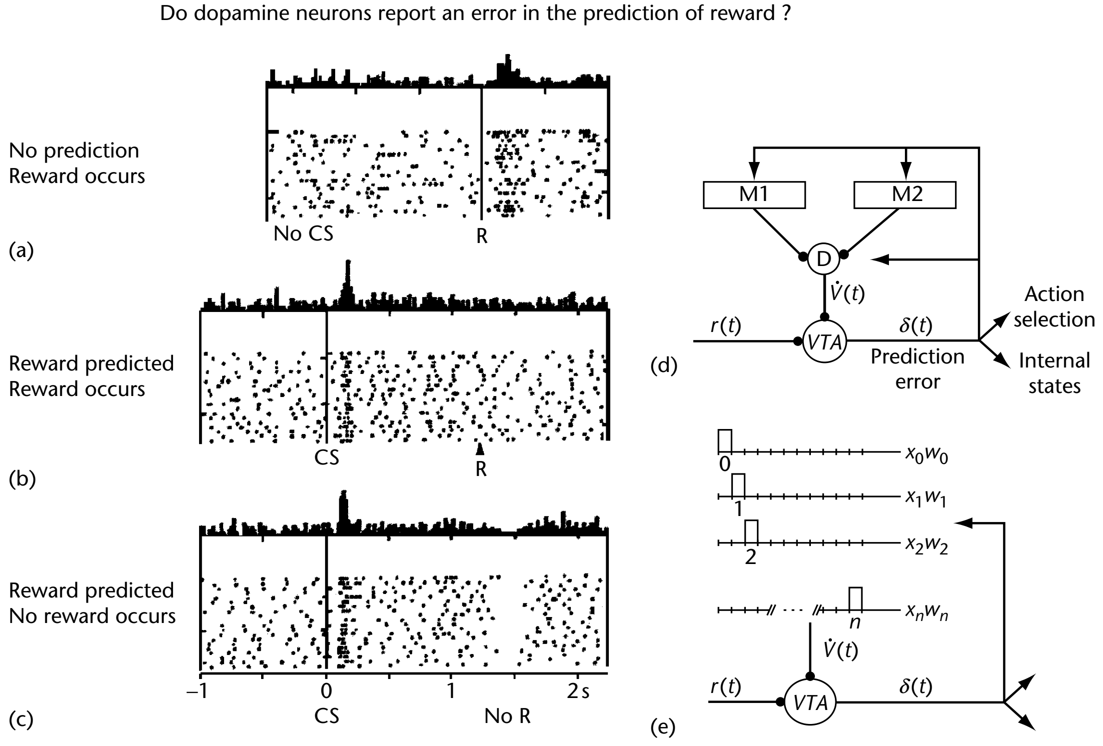

Figure 1. Left-hand panel. Dopaminergic predictor neurons during simple classical conditioning: plots of spike output of midbrain dopaminergic neurons recorded in alert primates during simple conditioning tasks. (a) Unpredicted delivery of reward (juice) causes a phasic increase in activity. This occurs for no conditioned stimulus (CS) (e.g. the light) or for a neutral CS (meaningless to a naive monkey). (b) After a number of trials involving a CS followed by juice delivery, the neurons stop responding to the presentation of the juice and instead give a phasic response just after the onset of the CS. (c) On error trials the cue is presented, but no reward is delivered. The neuron firing rate drops to 0 at the time when the reward would have been delivered if no mistake had been made. Right-hand panel. (d) Architecture of computational model mimics architecture of neuromodulatory systems in bees and humans. (e) Representation of a stimulus over time. There must be some representation of a stimulus over time in order to learn predictions. (Adapted from Schultz *et al.* (1997) *Science* **275**: 1593–1599.)

on the illumination of the light is also encoded in the fluctuations in dopaminergic activity.

From the results shown in Figure 1, it can be seen that dopamine neurons do not simply report the occurrence of rewarding events. Instead, their outputs appear to code for an error between the actual reward received and predictions of the time and magnitude of the reward. These neurons are only activated if the time of the reward is uncertain (i.e. unpredicted by any preceding cues). Dopamine neurons are therefore excellent feature detectors of the scalar 'goodness' of environmental events relative to learned predictions about those events. They emit a positive signal (increased spike production) if an appetitive event is better than predicted, no signal (no change in spike production) if an appetitive event occurs as predicted, and a negative signal (decreased spike production) if an appetitive event is worse than predicted. These observations have led to the hypothesis that these neurons are *predictor neurons* which distribute a prediction error signal (like $\delta(t)$ described above) to target neural structures in the form of changes in dopamine delivery.

Details of the way in which the dopamine signal is used neurally to implement full TD learning are beyond the scope of this article. The important point is that the electrical recordings from dopamine neurons show that they are part of a circuit in the brain that is capable of making predictions about future reward. This is one of the few instances where brain activity can be clearly connected to a well-understood computational learning procedure. Confirmation of the degree to which this description is incomplete or incorrect awaits future experimental and computational research. However, this work has provided a new way to interpret activity in dopamine neurons.

Dopamine neurons have long been known to be involved in reward processing. However, until recently it was thought that dopamine delivery was equivalent to reward. That is, the dopamine delivery itself was a positive reward signal. The experimental results described above, when combined with the computational interpretation in terms of TD learning, show that this cannot be the case. After training, there is no increased dopaminergic activity in response to the delivery of the rewarding fluid. The theory that dopamine is equivalent to reward is not consistent with this clear finding. However, an interpretation in terms of a prediction error is consistent, although this is not the end of the story. Dopaminergic activity is also seen in other contexts involving attention and orienting behavior. The connection to reinforcement learning has helped to frame the experimental issues in more sophisticated and formal terms.

One consequence of the TD model of dopaminergic function has recently been explored in human subjects using functional magnetic resonance imaging (fMRI).

PREDICTABILITY, REWARD AND HUMAN BRAIN RESPONSE

Reinforcement learning represents a general framework for describing how animals learn to interact with their environment. Therefore it is reasonable to extend this to the study of human reward pathways. Although these mechanisms are not thought to be fundamentally different in humans, the challenge lies in separating the core processes associated with reward prediction from secondary higher cognitive functions. For example, a temporal-difference model of reinforcement learning might predict that dopamine is released in proportion to prediction errors. While a monkey might learn a particular task by trial and error, humans are quite adept at using complex and abstract strategies that circumvent such laborious methods.

One relevant approach to studying humans is suggested directly by TD models. The core property that drives learning in TD models is predictability. Learning only occurs because there is some regular relationship between sensory cues and future reward. If there were no such relationship, then there would be nothing to learn. This can be used to advantage in humans. By simply altering the probability that one type of cue predicts a reward, one should be able to manipulate the response of dopaminergic systems in the brain. Furthermore, these manipulations can be subtle and therefore not consciously detectable by humans. These issues are important factors in experimental design if the aim is to prevent people from using conscious learning strategies.

It is worth noting that reward for a human may be different to reward for a monkey. We assume that human 'rewards' represent a lifetime of various types of conditioning that leads to the abstract assignment of reward to neutral objects (e.g. money). Understanding the way in which constructs such as money acquire reinforcing value are important but complex questions to address experimentally. However, it is worth first understanding how humans respond to the predictability of primary rewards, such as food and water.

In a simple experiment using fMRI, it has been shown that, just as in monkeys, the activity of dopaminergic projection sites in the human brain

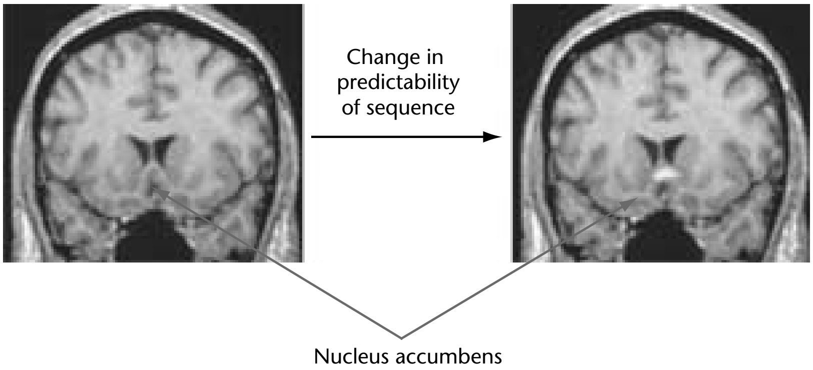

Figure 2. Change in predictability of a sequence of gustatory stimuli activates targets of dopamine projections. A sequence of squirts of fruit juice and water was delivered to the mouths of subjects. When the sequence of squirts changed from predictable to unpredictable, there was a larger activation in the nucleus accumbens, a region that is known to be involved in reward processing. (Adapted from Berns *et al*. (2001) *Journal of Neuroscience* **21**: 2793–2798.)

is modulated by the predictability of rewarding events. When subjects received squirts of fruit juice and water in their mouths, the activity in the nucleus accumbens was significantly amplified when the pattern of squirts was unpredictable (Figure 2). Since the participants were unaware of the differences in predictability, the fMRI activity changes cannot be ascribed to other top-down attentional processes.

SUMMARY

Reinforcement learning is one approach to the essential elements of learning problems that are encountered in real-world situations. There are numerous approaches to reinforcement learning, but here we have singled out temporal difference (TD) learning because it has growing connections to real biological data. We have reviewed one connection to the mammalian dopamine system and shown how a TD model accounts for changes in spike activity in dopaminergic neurons in alert monkeys during learning tasks. The same model provided an excellent starting point for the design of similar experiments in humans. One example of such an experiment using fMRI has been outlined, and the response of a target of dopamine projections, namely the nucleus accumbens, was shown to be activated by changes in the predictability of stimuli. This result was anticipated from the reinforcement learning model of the dopamine system, and it provides an excellent example of the way in which computational approaches to learning problems are informing the design and interpretation of experiments.

Further Reading

Bertsekas DP (1995) *Dynamic Programming and Optimal Control*. Belmont, MA: Athena Scientific.

Kaelbling LP, Littman ML and Moore AW (1996) Reinforcement learning: a survey. *Journal of Artificial Intelligence Research* **4**: 237–285.

Montague PR, Dayan P and Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience* **16**: 1936–1947.

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

Sutton RS and Barto AG (1998) *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.