Paper: Schultz, W., Dayan, P., & Montague, P. R. (1997). **A neural substrate of prediction and reward.**

Manoosh Samiei

This article reviews how temporal difference (TD) learning algorithm provide an interpretation of the activity of dopamine neurons thought to mediate reward-processing and reward-dependent learning.

Introduction

- Fluctuating output of dopaminergic neurons in the primate, signals changes or errors in the predictions of future salient and rewarding events.

- One clear connection between reward and prediction derives from a wide variety of conditioning experiments.

- **CS (Conditioned Stimulus):** The arbitrary cue, such as the **light** or a **tone**. It has no intrinsic value but comes to **predict** the reward.

- **US (Unconditioned Stimulus):** The thing with intrinsic, biological value, such as **food** or a **juice reward**. This is what the animal truly wants.

- We call the appetitive CS the sensory cue and the US the reward.

- To establish classical conditioning, the sensory cue must reliably precede the reward. Once learned, the cue predicts the reward's timing and magnitude and the value of the reward is transferred to the cue.

- Some theories suggest that unpredictability of the reward by the sensory cue drives learning.

- No further learning takes place when the reward is entirely predicted by a sensory cue (or cues).

- Once a rat is conditioned to expect food following a light, the subsequent pairing of that light with a sound preceding the food results in the rat relying exclusively on the light as the predictive cue, effectively ignoring the sound. This phenomenon is called “blocking.”

- Because the light perfectly predicts the food, the sound provides no new information, and therefore, no association is learned with it.

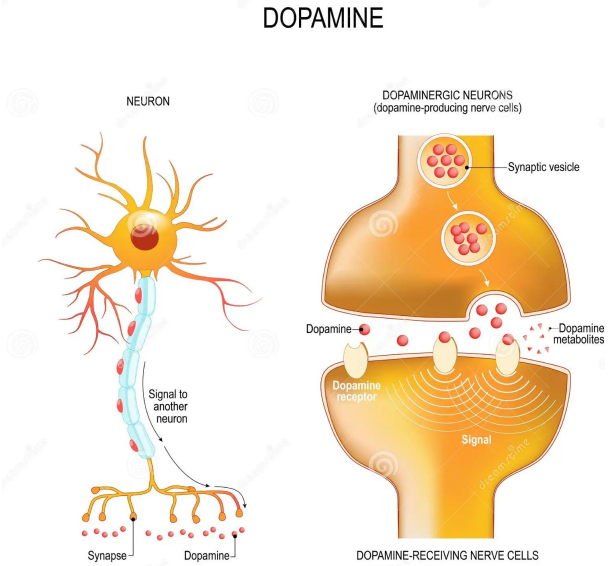
- It appears therefore that learning is driven by deviations or “errors” between the predicted time and amount of rewards and their actual experienced times and magnitudes.

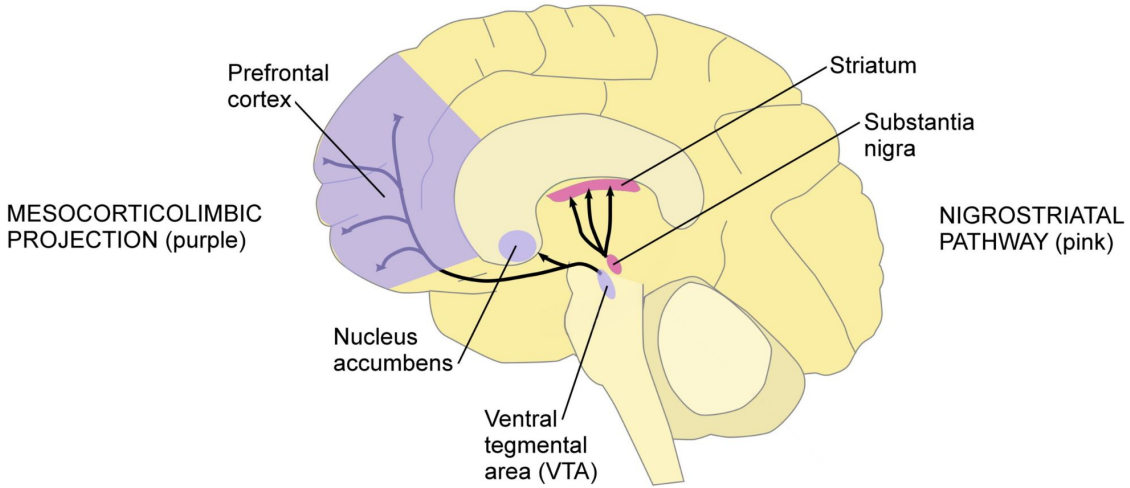
- In reinforcement learning, systems learn to predict through temporal difference (TD) algorithm. This algorithm was originally inspired by behavioral data on how animals actually learn predictions.

Anatomy:

Dopamine neurons of the **ventral tegmental area (VTA)** and **substantia nigra** have long been identified with the processing of rewarding stimuli.

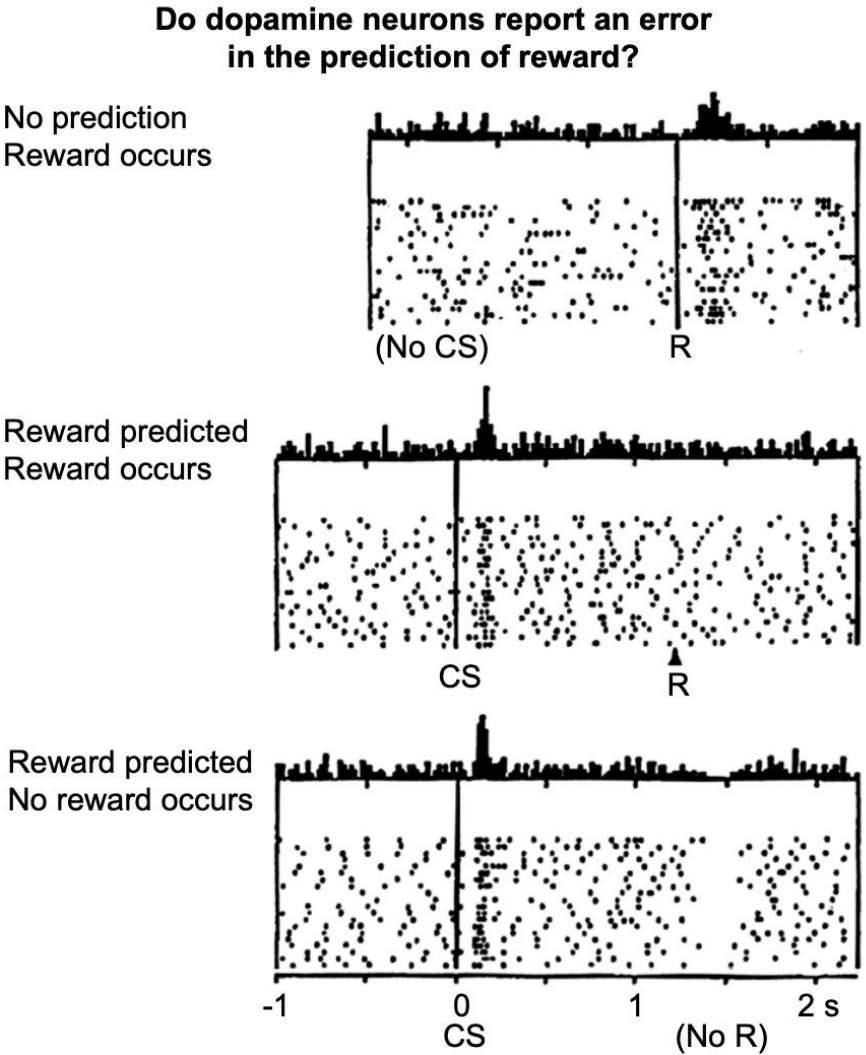
These neurons send their axons to brain structures involved in motivation and goal-directed behavior, for example, the **striatum**, **nucleus accumbens**, and **frontal cortex**.





After an animal repeatedly experiences a visual and auditory cue followed by a reward, its dopamine neurons shift their peak activation from the moment the reward arrives to the moment the cues begin.

When the expected reward is omitted following the light cue, dopamine neurons show a marked decrease in firing (depression) precisely at the moment the reward was anticipated.



Dopamine neurons code for a **prediction error**, the difference between the **actual reward** and the **predicted time and magnitude** of that reward.

- They show **increased firing** (a positive signal) if the reward is **better than predicted** (or entirely uncertain/unpredicted).
- They show **no change in firing** if the reward occurs **exactly as predicted**.
- They show **decreased firing** (a negative signal) if the reward is **worse than predicted** (or omitted).

Essentially, these neurons act as **feature detectors** for how "good" environmental events are relative to what has been learned and expected.

TD algorithm

- Temporal difference methods were introduced into the psychological and biological literature by Richard Sutton and Andrew Barto in the early 1980s.

- The main assumption in **Temporal Difference (TD) learning** is that the computational goal is to use **sensory cues** to predict V(t), the **discounted sum of all future rewards** in a trial.

$$V(t) = E[\gamma^0 r(t) + \gamma^1 r(t+1) + \gamma^2 r(t+2) + \cdots]$$

- The second main assumption of TD learning is the **Markovian assumption**: **future cues and rewards depend only on the immediate, current cues**.

The overall strategy is to use a **vector of sensory cues** (x(t)) combined with a **vector of adjustable weights** (w) to generate V^(t), an **estimate of the true predicted future reward** (V(t)).

To assess its predictions, this 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

$$V(t) = E[r(t) + \gamma V(t+1)]$$

An error (TD error) in the estimated predictions can now be defined with information available at successive time steps:

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

- A simple model using only **one adaptable weight** per sensory cue is **insufficient** because a cue can predict a reward at **arbitrary times** into the future. Experimental data show that when the time between the cue and reward is changed, the cue learns to predict this **new time** of delivery.

- We assume that each sensory cue is represented by a **vector of signals** x(t) = {x_1(t), x_2(t), …}. Each component x_i(t) acts as a **temporal feature**, becoming 1 exactly **i time steps after the cue's onset** and 0 otherwise. This allows the cue's representation to be distributed across future time points.

- If the light comes on at time s, x_1(s+1)=1, x_2(s+2)=1, . . . represent the light at 1, 2,... time steps into the future and w_1, w_2, . . . are the respective weights. The net prediction for cue x(t) at time t takes the simple linear form:

$$\hat{V}(t) \equiv \hat{V}(\mathbf{x}(t)) = \Sigma_i w_i x_i(t)$$

This type of temporal representation, where a sensory cue is encoded as a vector of signals active at successive time steps, is referred to as a **complete serial-compound stimulus** by Sutton and Barto, and is functionally related to Grossberg's spectral timing model.

The adaptable weights w are improved according to the correlation between the stimulus representations and the prediction error.

$$\Delta w_i = \alpha_x \Sigma_t x_i(t) \delta(t)$$

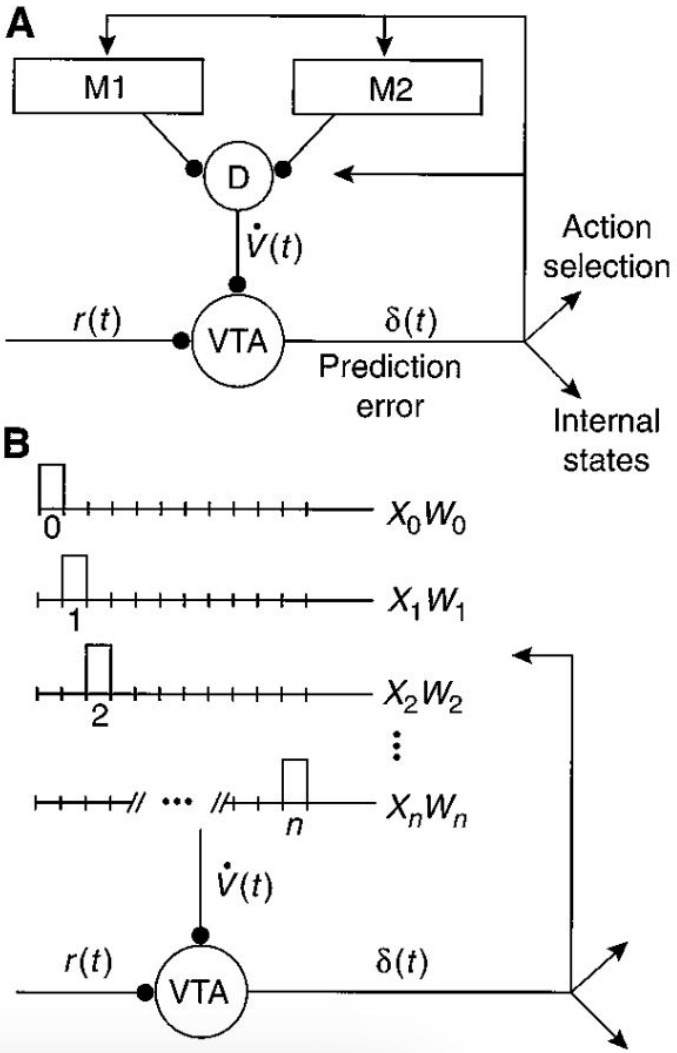
Where alpha is the learning rate for cue x(t) and the sum over t is taken over the course of a trial.

If there were many different sensory cues, each would have its own vector representation and its own vector of weights, and the above equation would be summed over all the cues.

To construct and use an error signal similar to the TD error above, a neural system would need to possess four basic features:

- Access to a measure of reward value **r(t)**;
- TD error: **gamma*Vˆ(t + 1) - Vˆ(t)**;
- A site where these signals could be **summed**
- **Delivery of the error signal** to areas constructing the prediction in such a way that it can control plasticity.
- It has been shown that midbrain dopamine neurons satisfy the first three features.

M1 and M2 represent two different cortical modalities whose output is assumed to arrive at the VTA in the form of a temporal derivative (surprise signal) $\dot{V}(t)$, which reflects the degree to which the current sensory state differs from the previous sensory state.

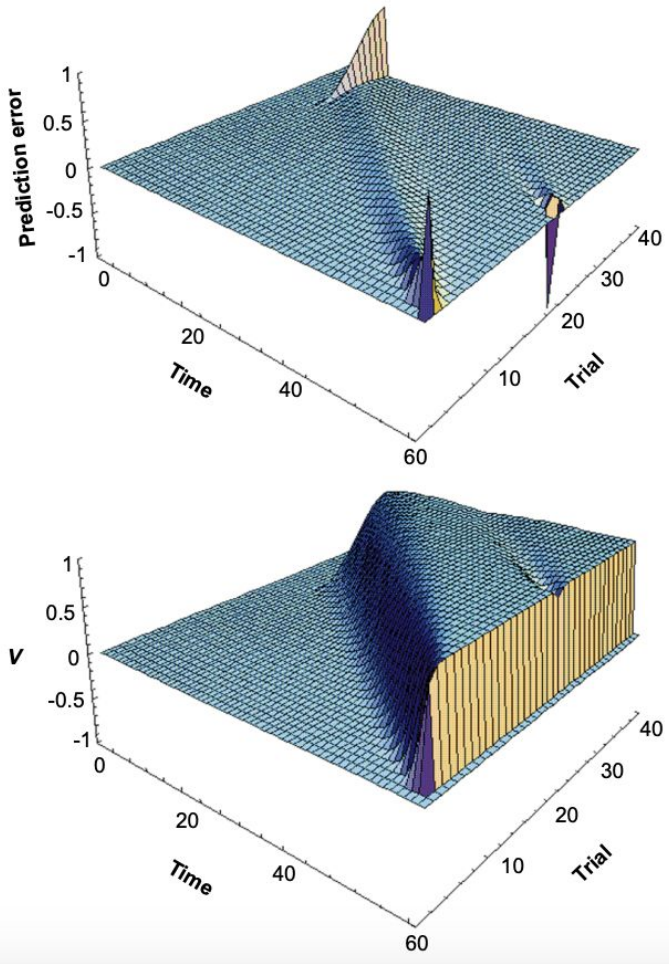


We assume that the dopamine neurons' output actually reflects:

$$\delta(t) + b(t)$$

Where b(t) is a basal firing rate.

Figure shows the training of the model on a task where a single sensory cue predicted the future delivery of a fixed amount of reward 20 time steps into the future. The model's **prediction error signal** successfully replicates the activity of **dopamine neurons** during learning. The resulting pattern of adaptable weights explains key behavioral findings like **blocking** and **secondary conditioning**, as well as dopaminergic changes when the **reward timing is altered**.



The model makes two main predictions regarding the phasic dopamine response in the presence of a cue sequence:

1. **Transfer to the Earliest Cue:** When multiple cues precede a reward, the phasic dopamine activation (the prediction signal) will **transfer entirely to the earliest consistent cue** in the sequence.

2. **Omission Signal at Intermediate Cue:** After training on a cue sequence (e.g., Light 1 -> Light 2 -> Reward), **omitting an intermediate cue** (Light 2) will cause a **phasic decrease** (depression) in dopamine activity precisely at the moment the omitted cue was expected to occur.

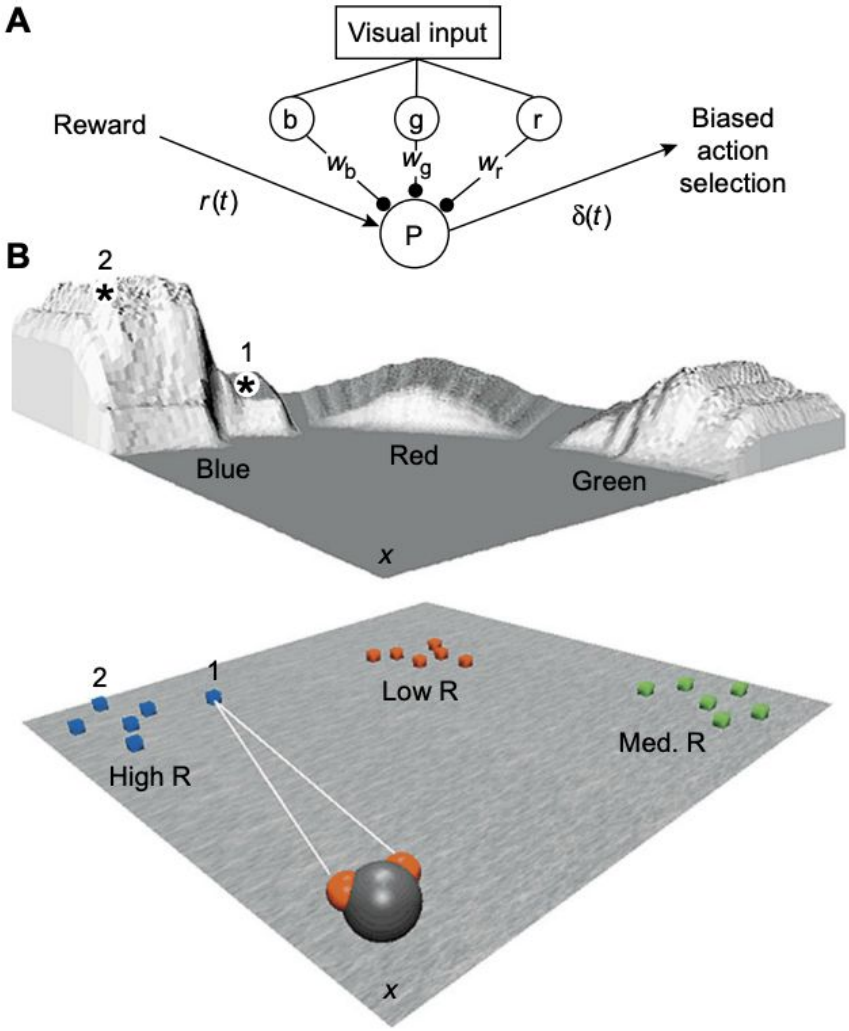
Challenge of temporal credit assignment: How does the rat, after hitting a dead end many steps later, correctly identify which specific action or "wrong turn" far in the past was responsible for the negative outcome?

One solution to the credit assignment problem is for the animal to develop a **policy** a rule dictating its actions for every **state** (the sensory cues at each maze position). To improve this policy, the animal must learn the value of each state. This value, used in dynamic programming, is the summed future reward expected from that state if the policy is followed. Since the TD method learns this exact expected future reward, it is suggested to be the mechanism underlying the brain's dopamine signal.

Bee foraging behavior on flowers can be constructed by a TD model.

Three color-sensitive units (b, g, r) report, respectively, the percentage of blue, green, and red in the visual field. Each unit influences neuron P (VTA analog) through a single weight.

The colored blocks contain varying amounts of reward with blue> green > red. After training, the weights (wb, wg, wr) reflect this difference in reward content.



Using only a single weight for each sensory cue, the model can make only one-time step predictions; however, combined with its capacity to move its head or walk about the arena, a crude “value-map” is available in the output of neuron P.

The height of the surface codes for the value V(x, y) of each location when viewed from the corner where the “creature” is positioned. All the creature needs to do is look from one location to another (or move from one position to another), and the differences in value V(t + 1) - V(t) are coded in the changes in the firing rate of P.

Summary

The dopamine neurons in the **VTA** and **substantia nigra** are proposed to report an **ongoing prediction error signal** for reward. This scalar (single-valued) error signal is delivered to target structures to influence how the brain **processes predictions** and **chooses reward-maximizing actions**.

Several issues for future work:

1. One issue is **temporal representation**: learning **how** a stimulus is encoded across time to make **precise, time-based predictions** of future events. Although animals clearly demonstrate this ability, the specific neural location and mechanism (the "temporal labels") used by the brain remain unknown.

2. Second issue is that dopamine system is mainly associated with **rewards**, not punishments. Authors suggest that the **omission of an expected reward** serves as a form of **"punishment,"** which the dopamine neurons signal via a **negative prediction error**. This signal is then processed by target structures, supporting the role of **opponent processes** (rewards vs. punishments) in learning.

3. A third issue is the required **relation between scalar and vector signals** in reward learning. The model uses a simple **scalar** (single-value) prediction error, which is insufficient for complex environments. Realistic behavior requires **vector signals** (multi-component) to represent the **specific type of reward** (e.g., food vs. water) and the **physical attributes of predictive cues**.

Because dopamine neurons only emit a non-specific **scalar appetitive error signal** (a "teaching" signal without details), other brain structures must be responsible for the **analysis and discrimination** of specific rewards.

4. The model currently fails to account for **attentional functions** in target structures like the **nucleus accumbens** and **frontal cortex**, which are crucial when varying amounts of attention are paid to different stimuli.

There's evidence suggesting **attentional mechanisms** might operate directly at the **dopamine neuron level**, as their responses decrease with repeated novel stimuli and they will generalize their responses to non appetitive stimuli that are physically similar to appetitive stimuli.

In the mammalian brain, the **striatum** is one site where this kind of scalar evaluation could have a direct effect on action choice, and activity relating to conditioned stimuli is seen in the striatum

Dopamine's influence extends to the **cerebral cortex**, dramatically affecting functions like **working memory** (in prefrontal cortex) and cognitive activation (in anterior cingulate cortex).

The **prediction error signal** (carried by dopamine) must be delivered specifically to the **local cortical regions** that actually made the faulty prediction, rather than broadcasting it globally. This mechanism of **precisely timed, specific information delivery** by the dopamine system is a significant shift from the traditional view that neuromodulators only provide slow, global state modulation, demonstrating their vital role in rapid and targeted cognitive functions.