Reinforcement learning: Computational theory and biological mechanisms

a discussion on Doya's paper

Andrei Florea Adrian Florea

15th Meetup of Papers We Love (Bucharest Chapter),
7 April 2017

Agenda

- How is information relayed in the brain?
- Conditioning as a primal formal learning
- Reinforcement learning introduction, Bellman equation, TD error
- Dopamine, the center piece in the puzzle of learning without a teacher
- Lessons from mental disease and altered states of mind

Definition of a Neuron

- **Biology**: A neuron is an electrically excitable cell that possesses and transmits information through electrical and chemical signals.
- **Neuroscience**: Neurons are the basic information processing structures in the Central Nervous System.
  Everything above the level of neurons qualifies as information processing too. But nothing below the level of neurons does.

Chemical vs. Electric Conduction of Information

Q: Why do we need both means of chemical and electrical conduction?

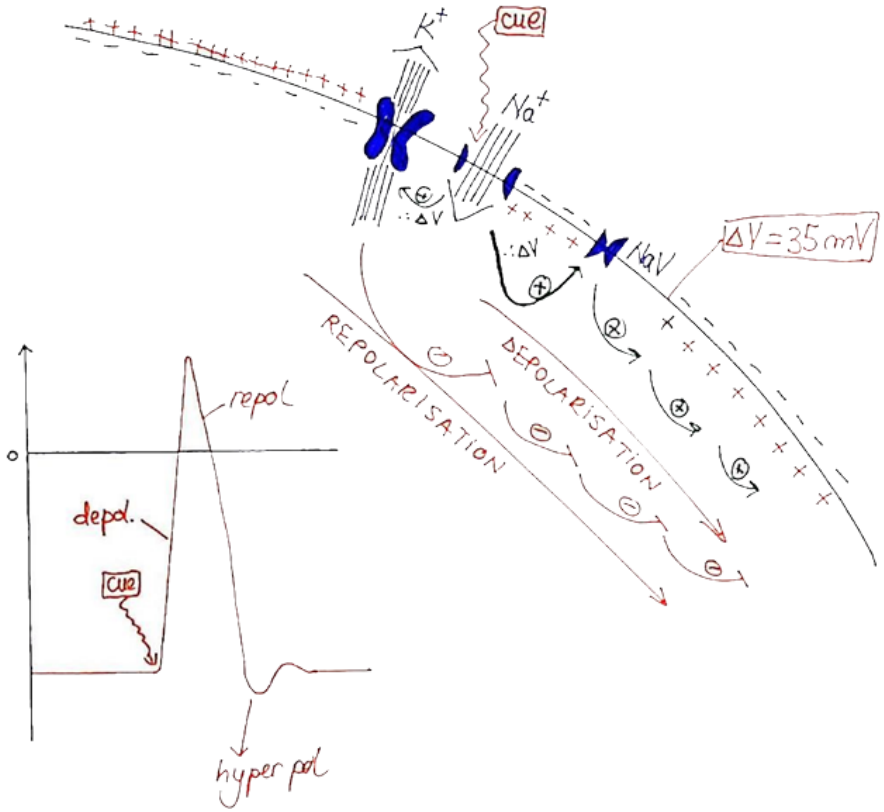
- **E**: Transmitting information chemically is very slow, limited by the diffusion limit (a.k.a. the speed things spread evenly within a solution through Brownian motion) a few cm/s for chemical conduction vs. a few hundreds of m/s for electrical conduction.
- **E**: Chemical conduction creates a gradient of concentration in molecules, the concentration of molecules decaying exponentially the further you go from the origin of the chemical signal. Very low signal/noise ratio
- **E/C**: Chemical conduction requires a 1 signal : 1 neurotransmitter relationship, whereas electrical conduction allows for the conveying of a single signal corresponding to the weighted average of all inputs received.
- **C**: Electric conduction in very similar neurons (much unlike electrical circuits) will lead to a homogeneous spread of electric impulses. Instead, chemical conduction helps to isolate specific signals and to maintain directionality of the signal.

Action Potentials

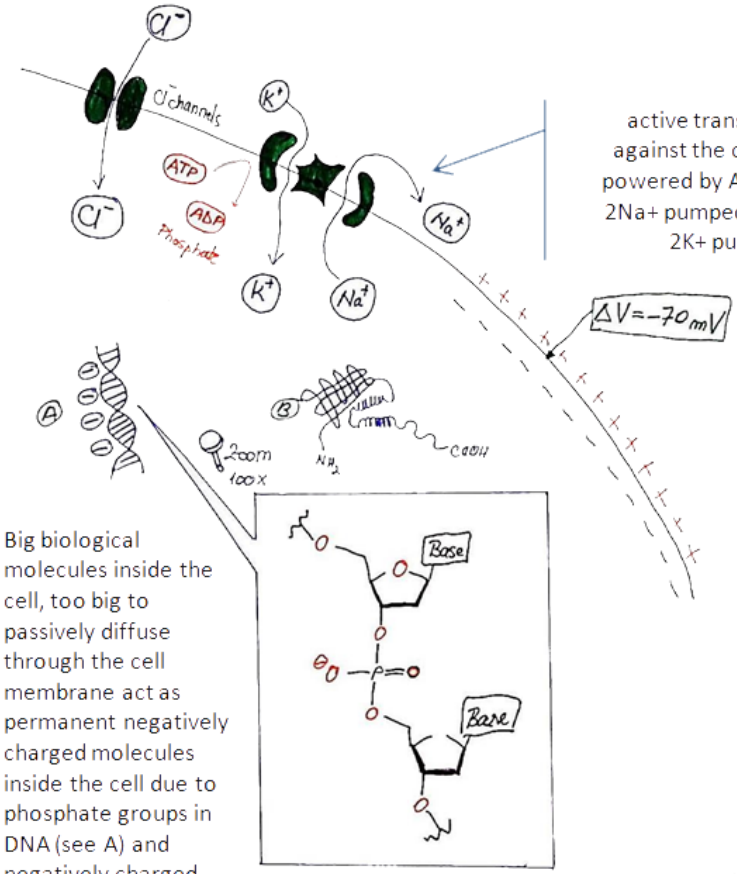
Q: How is information relayed?

- The cellular membrane of the neuron separates the interior from the exterior and allows only passage of some chemical species.
- Ions are normally trapped either in or out (Na^+, K^+) and their relative densities of charge on both faces of the membrane make it act like a *de facto* capacitor.
- In the resting state (no signal), there is more negative charge on average from inside the cell than the outside, thus the membrane being formally negatively charged on the inside and positively on the outside, in the lack of an excitatory signal.
- In the resting state, the ATP dependent Na^+/K^+ exchanger removes 3 Na^+ ions from inside the cell in exchange for 2 K^+ ions, against the concentration gradient.

Action potential



Resting potential



active transport of ions against the conc. gradient powered by ATP. Exchanges 2Na+ pumped out for every 2K+ pumped in

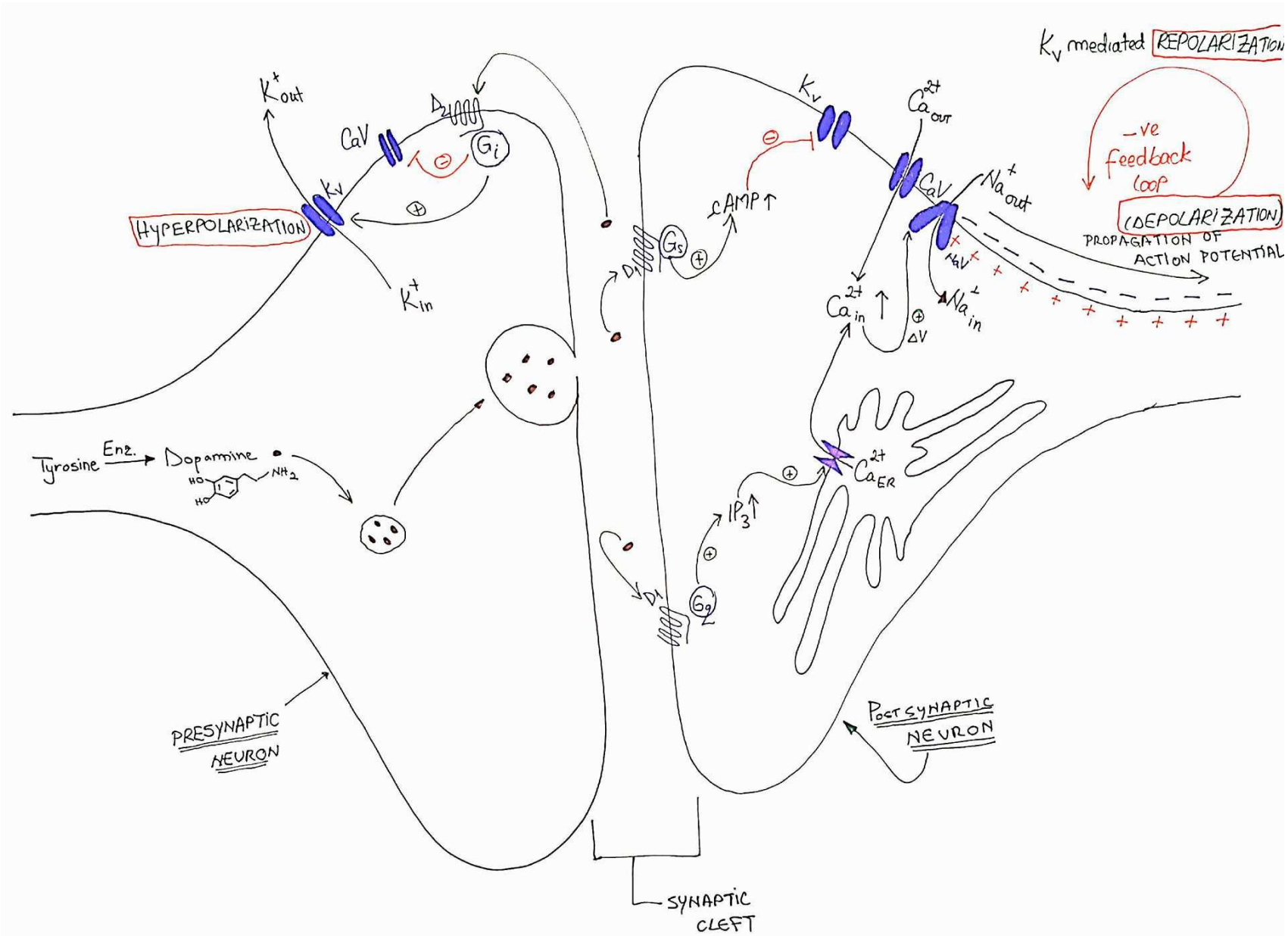
Big biological molecules inside the cell, too big to passively diffuse through the cell membrane act as permanent negatively charged molecules inside the cell due to phosphate groups in DNA (see A) and negatively charged groups on polypeptides and proteins (see B)

- Neurons have dendrites that receive inputs from other neurons, cell bodies that process these inputs and axons that transmit unified outputs.
- Synapses occur at the junctions of two neurons and classically involve the pre-synaptic axon of one neuron and the post-synaptic dendrite of another neuron.
- Synapses involve a synaptic cleft, a tight junction of extracellular medium, where neurotransmitters are released and act on specific receptors of the post-synaptic neuron.

Principles of a synapse

1. Neurotransmitters are released by the pre-synaptic neuron and can act on receptors on both neurons involved in the synapse
2. One neurotransmitter can act on a multitude of receptors, with responses varying from slightly different to complete opposites (i.e. inhibition instead of activation in the case of the D2R compared to the D1R - see later)
3. Mechanisms exist in place so the signal is unidirectional
4. Signals past the threshold are characterised by a signal strength.

Action Potentials



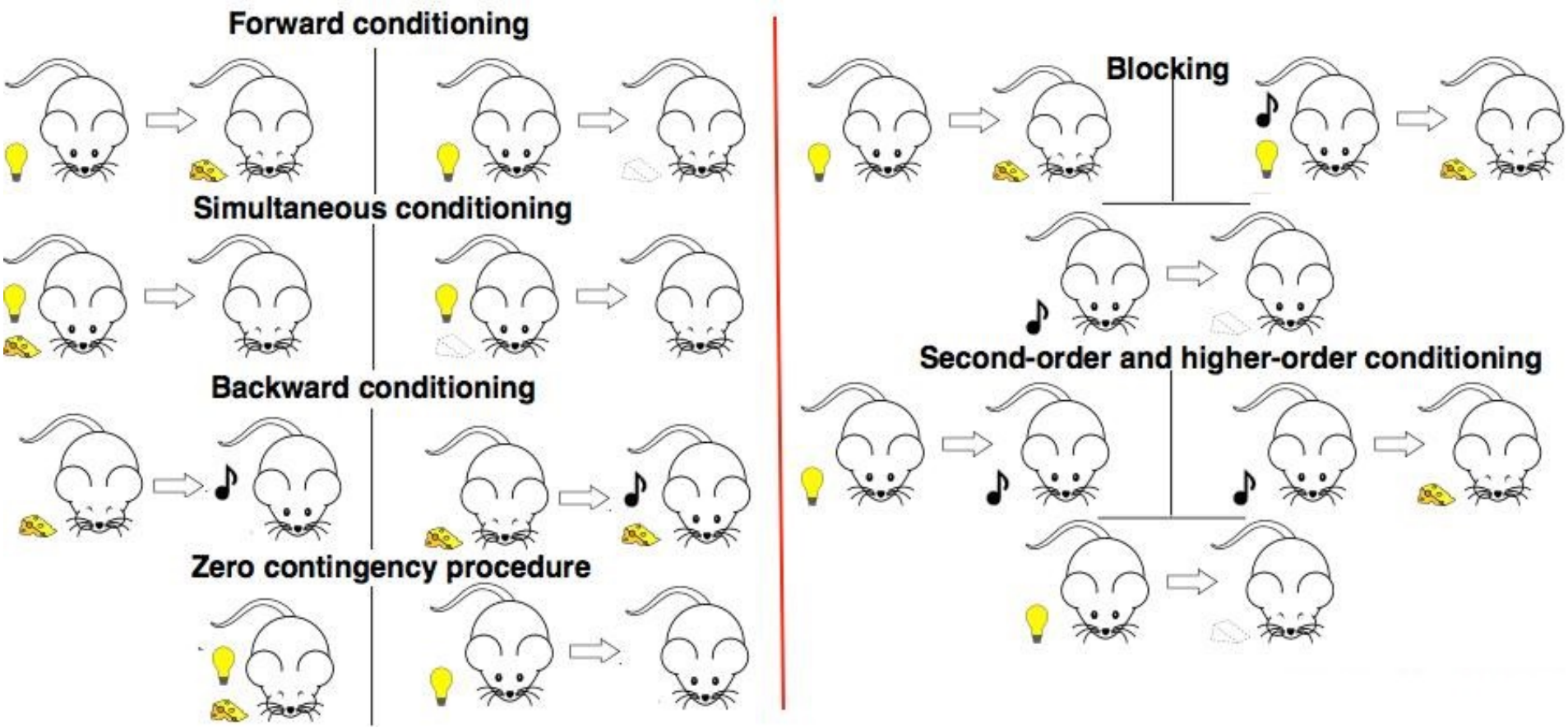
Classical conditioning

- Conditioning, or developing responses as a consequence of the presence of a conditional stimulus **CS** (i.e. light), is a well established neuroscience experiment based on pairing an unconditional stimulus **US** (i.e. food) with a conditional one (i.e. light). If light is shone repeatedly before feeding the animal, salivation will be triggered upon shining the light alone.
  - CS cannot succeed US (creates backward conditioning, which is inhibitory in nature and suggests US has ended)
  - CS has to be followed by US or else the response will fail to appear (also known as extinction, the frequency of the conditional response (CR) will return to pre-training levels if the frequency of application of US is not high enough)
  - Pairing CS1 with US prompts a CR. Overlapping CS1 with CS2 and then subtracting CS1 does not trigger a CR. (see quote below)

A Neural Substrate of Prediction and Reward, Schultz W., 1997

Some theories ... suggest that learning is driven by the unpredictability of the reward by the sensory cue. One of the main ideas is that no further learning takes place when the reward is entirely predicted by a sensory cue. ... If, after such training, the light is paired with a sound and this pair is consistently followed by food, then something unusual happens – the rats behavior indicates that the light continues to predict food, but the sound predicts nothing. This phenomenon is called blocking.

Classical Conditioning, Blocking and Higher Order Conditioning



There seems to be a link in between learning, reward and dopamine.

Evidence

- Dopaminergic Neurons in the VTA and Substantia nigra send their axons to the brain structures involved in motivation and goal-directed behavior, for example, the striatum, nucleus accumbens, and frontal cortex. (**from anatomy and dissections**)
- Drugs such as cocaine and amphetamine which boost the pleasure and reward centres of the brain block molecular structures responsible for degrading/removing synaptic dopamine. (**from pharmacology**)
- Electrodes attached to dopaminergic neurons give a pleasure and reward feeling to live animals, enough so that they disregard food and sex. (**from direct brain stimulation studies**)
- Rats treated with dopamine blockers (an antagonist of the D1R receptor earlier discussed) less likely to press lever for food (reward) (**from sad over-medicated rats**)

The body of evidence points towards dopaminergic pathways in the midbrain

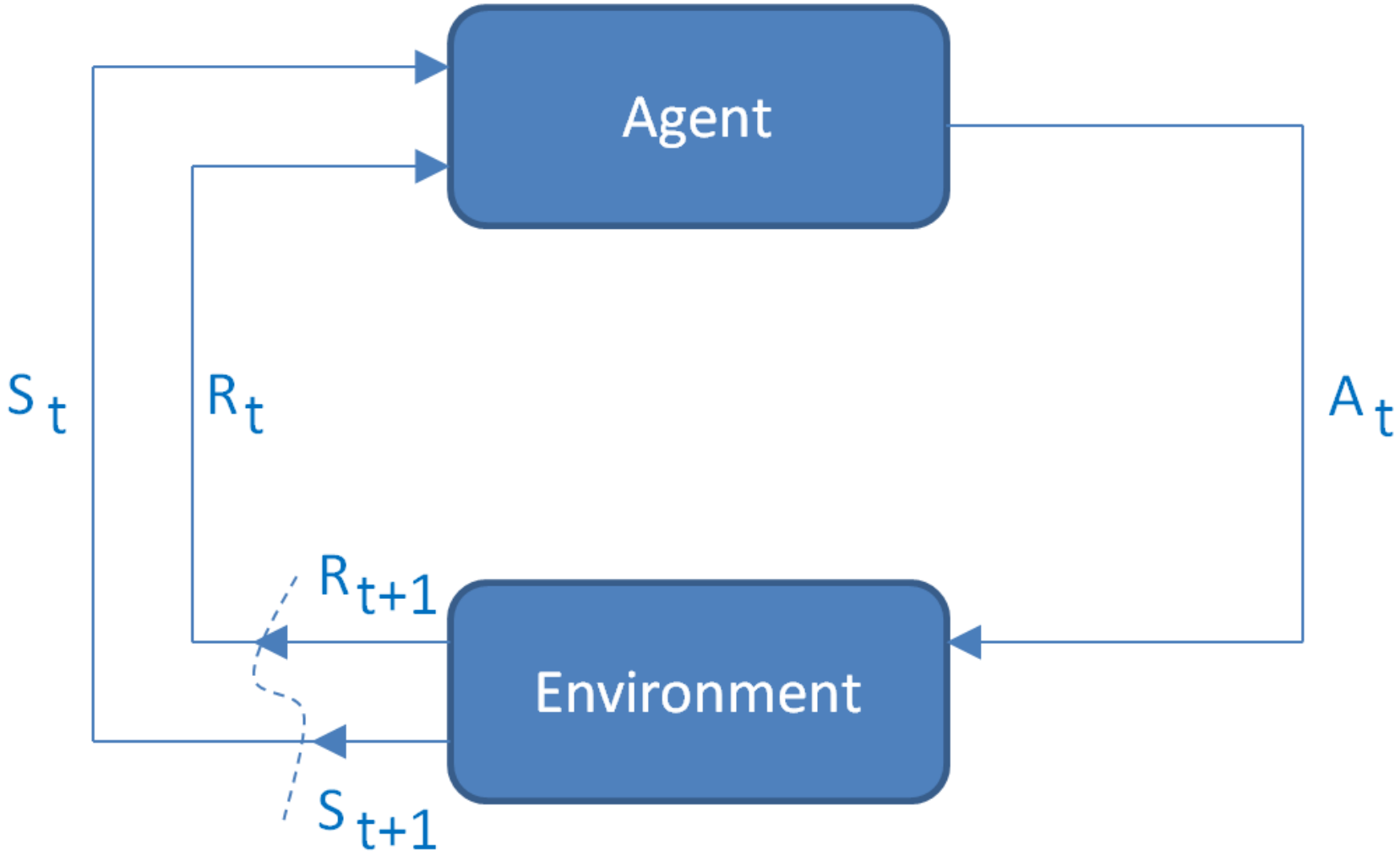
Reinforcement learning: another ML paradigm

- An agent is learning from interaction with the environment to achieve an explicit goal.
- The agent must be able to sense the state of the environment and learn how to take actions that affect the state.
- Closed-loop problem (learning actions influence the agent's later inputs over an extended time period, feedback not instantaneous).
- No teacher, the agent is learning from its own experience (trial & error paradigm).
- Trade-off between exploration (actions not selected before, to find more about the environment) and exploitation (actions selected before, to maximize reward).
- The closest ML form to the kind of learning that humans and other animals do.

Main elements of reinforcement learning

- The agent is the learner and decision maker.
- Anything that cannot be changed arbitrarily by the agent can be part of its environment.
- A policy is a mapping from perceived states of the environment to actions to be taken by the agent in those states (it's the agent's behaviour, similar with *stimulus-response rules* in psychology).
- At each time step, the environment sends a reward signal (a scalar) to the agent, reward signal that is the primary basis for altering the policy (it *immediately* tells what is good/bad); the agent's goal is to *maximize* the sum of reward signals received over the long run.
- The value of a state is the sum of all reward signals an agent expects to receive over the future, starting from that state; the agent selects actions that bring states of highest value.
- The model of the environment is optional and predicts the next state and next reward signal.

Agent-environment interaction



- Agent and environment interact at discrete *time steps* (not necessarily fixed intervals), $t = 0, 1, 2, \ldots$; the problem can be extended to continuous time.
- *Return*: $G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \ldots$, where $R_t \in \mathcal{R} \subset \mathbb{R}$ and to avoid an infinite value of the return we can define:
- *Discounted return*: $G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$, that has a finite value if the reward sequence is bounded and:
- *Discount rate*: $\gamma \in [0, 1]$ (uncertainty of the future! $\gamma \ll 0.5$ is *myopic* and $\gamma \gg 0.5$ is farsighted)
- *State*: $S_t \in \mathcal{S}$, $\mathcal{S}$ is the set of all possible states
- *Action*: $A_t \in \mathcal{A}(S_t)$, $\mathcal{A}(S_t)$ is the set of actions available in S_t
- *History*: $H_t \doteq S_0, A_0, R_1, \ldots, S_{t-1}, A_{t-1}, R_t, S_t, A_t$

We assume that $\mathcal{S}$ and $\mathcal{R}$ are finite so we work with sums and probabilities instead of integrals and probability densities

Markov states and Markov Decision Processes

Ideally, a state S_t should retain all relevant information from the history H_t. Such a state is called a *Markov state* and has the property:

$$\mathbf{Pr}\{S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a\} = \mathbf{Pr}\{S_{t+1} = s', R_{t+1} = r | H_t\}$$

for all s, a, r, and s'. We can define:

$$p(s', r|s, a) \doteq \mathbf{Pr}\{S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a\}$$

Given a Markov state and action we can predict the next state and the next reward and by extension all the future states and rewards.
It's useful to assume/approximate that all the states in a process are Markov states. Such a process is called a *Markov Decision Process* (*MDP*).

$$\mathbb{E}[dice] = 1\frac{1}{6} + 2\frac{1}{6} + 3\frac{1}{6} + 4\frac{1}{6} + 5\frac{1}{6} + 6\frac{1}{6} = 3.5$$

$$r(s,a) \doteq \mathbb{E}[R_{t+1}|S_t = s, A_t = a] = \sum_{r\in\mathcal{R}} r \sum_{s'\in\mathcal{S}} p(s', r|s, a)$$

$$p(s'|s,a) \doteq \mathbf{Pr}\{S_{t+1} = s'|S_t = s, A_t = a\} = \sum_{r\in\mathcal{R}} p(s', r|s, a)$$

$$r(s,a,s') \doteq \mathbb{E}[R_{t+1}|S_t = s, A_t = a, S_{t+1} = s'] = \frac{\sum_{r\in\mathcal{R}} rp(s', r|s, a)}{p(s'|s,a)}$$

Policy: $\pi_t(a|s) \doteq \mathbf{Pr}\{A_t = a|S_t = s\}$
State-value function for policy π:

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t|S_t = s] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s]$$

Action-value function for policy π:

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t|S_t = s, A_t = a] = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s, A_t = a]$$

State transition and state value function

We can define the *state transition probability*:

$$p(s, s') \doteq \mathbf{Pr}\{S_{t+1} = s' | S_t = s\}$$

State transition matrix (each row of $\mathcal{P}$ sums to 1):

$$\mathcal{P} \doteq \begin{bmatrix} p(s_1, s_1) & \dots & p(s_1, s_n) \\ \vdots & \ddots & \vdots \\ p(s_n, s_1) & \dots & p(s_n, s_n) \end{bmatrix}$$

$$r(s) \doteq \mathbb{E}[R_{t+1} | S_t = s]$$

State value function:

$$v(s) \doteq \mathbb{E}[G_t | S_t = s]$$

$$v(s) = \mathbb{E}[G_t | S_t = s] = \mathbb{E}[R_{t+1} + \gamma G_{t+1} | S_t = s] = \mathbb{E}[R_{t+1}] + \gamma \mathbb{E}[G_{t+1} | S_t = s]$$

Bellman equation

$$v(s) = r(s) + \gamma \sum_{s' \in \mathcal{S}} p(s, s') v(s')$$

We have obtained the *Bellman equation* in a *MRP* in a linear, matricial form:

$$\vec{v} = \vec{r} + \gamma \mathcal{P} \vec{v}$$

so:

$$\vec{v} = (I - \gamma \mathcal{P})^{-1} \vec{r}$$

It can be shown that the Bellman equation in a *MDP* is:

$$v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) \sum_{s' \in \mathcal{S}, r \in \mathcal{R}} p(s', r|s, a)[r + \gamma v_\pi(s')], \forall s \in \mathcal{S}$$

and is non-linear. We can solve it using iterative methods.

First-visit Monte-Carlo policy evaluation

Total return $S(s) \leftarrow 0, \forall s \in \mathcal{S}$
Repeat forever:

Generate an episode, using the policy π that we need to evaluate

For each s in episode

$t \leftarrow$ the first time-step s is visited in the episode
$N(s) \leftarrow N(s) + 1$
$S(s) \leftarrow S(s) + G_t$
$V(s) \leftarrow \frac{S(s)}{N(s)}$

It can be proved that $V(s) \xrightarrow[N(s) \to \infty]{} v_\pi(s)$

But $S(s)/N(s)$, the expression of $V(s)$, is an average. It would be nice to compute it incrementally:

$$\mu_k = \frac{1}{k}\sum_{i=1}^{k} x_i = \frac{1}{k}(x_k + \sum_{i=1}^{k-1} x_i) = \frac{1}{k}(x_k + (k-1)\mu_{k-1}) = \mu_{k-1} + \frac{1}{k}(x_k - \mu_{k-1})$$

That means:

$$V(s_t) \leftarrow V(s_t) + \frac{1}{N(s_t)}(G_t - V(s_t))$$

or:

$$V(s_t) \leftarrow V(s_t) + \alpha(G_t - V(s_t))$$

Because of G_t, the update waits until the return following the visit is known. But, from Bellman equation, we can have the update as:

$$V(s_t) \leftarrow V(s_t) + \alpha(R_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

that means the update needs to wait only until the next time step.

TD error

We define as *TD error* the quantity:

$$\delta_t \doteq R_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

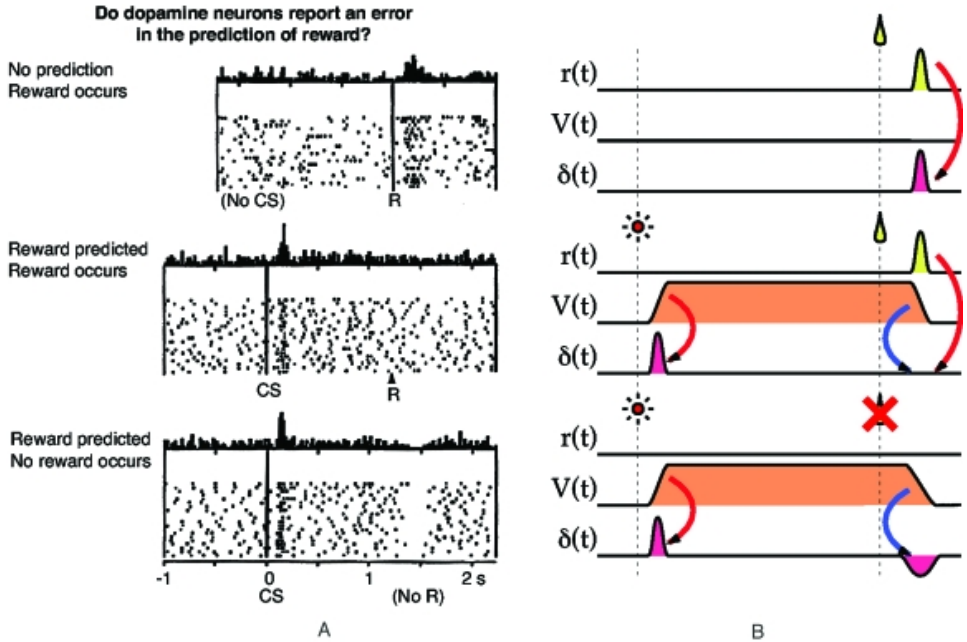
and we have:

$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

Single Dopaminergic Neuron Electrophysiology 1

Experimental Set-Up

Work by **Schultz et al. (1997)**, where awake monkeys bearing a microelectrode brain implant capable of discerning individual neuron signals were subjected to behavioural tests. The monkeys were locked in a room with a lever that, upon pulling, released a specific amount of juice, on the monkeys' liking. The brain activity of the dopaminergic neurons was recorded self-trained and un-trained monkeys respectively.



Single Dopaminergic Neuron Electrophysiology 2

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

no reward predicted, reward occurs

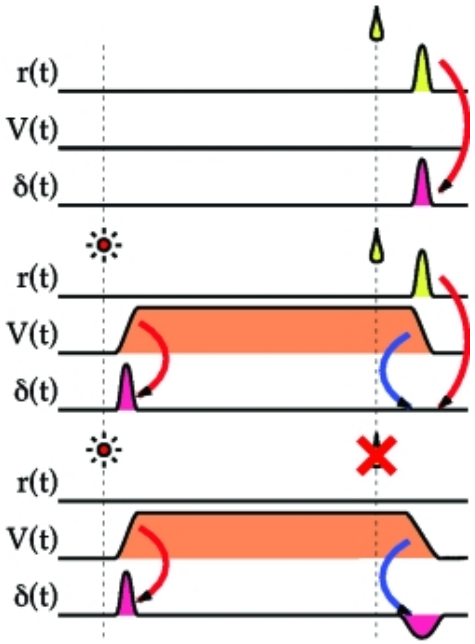
$$V(t) = 0, \forall t \implies \delta(t) = 0, \forall t$$

reward predicted, reward occurs

$$\delta_{cue} = 0 + \gamma V - 0 = \gamma V$$
$$V = \mathbb{E}[G] = R \implies \delta_{cue} = \gamma R$$
$$\delta_{reward} = R + \gamma 0 - V = R - V$$
$$V = \mathbb{E}[G] = R \implies \delta_{reward} = 0$$

reward predicted, no reward occurs

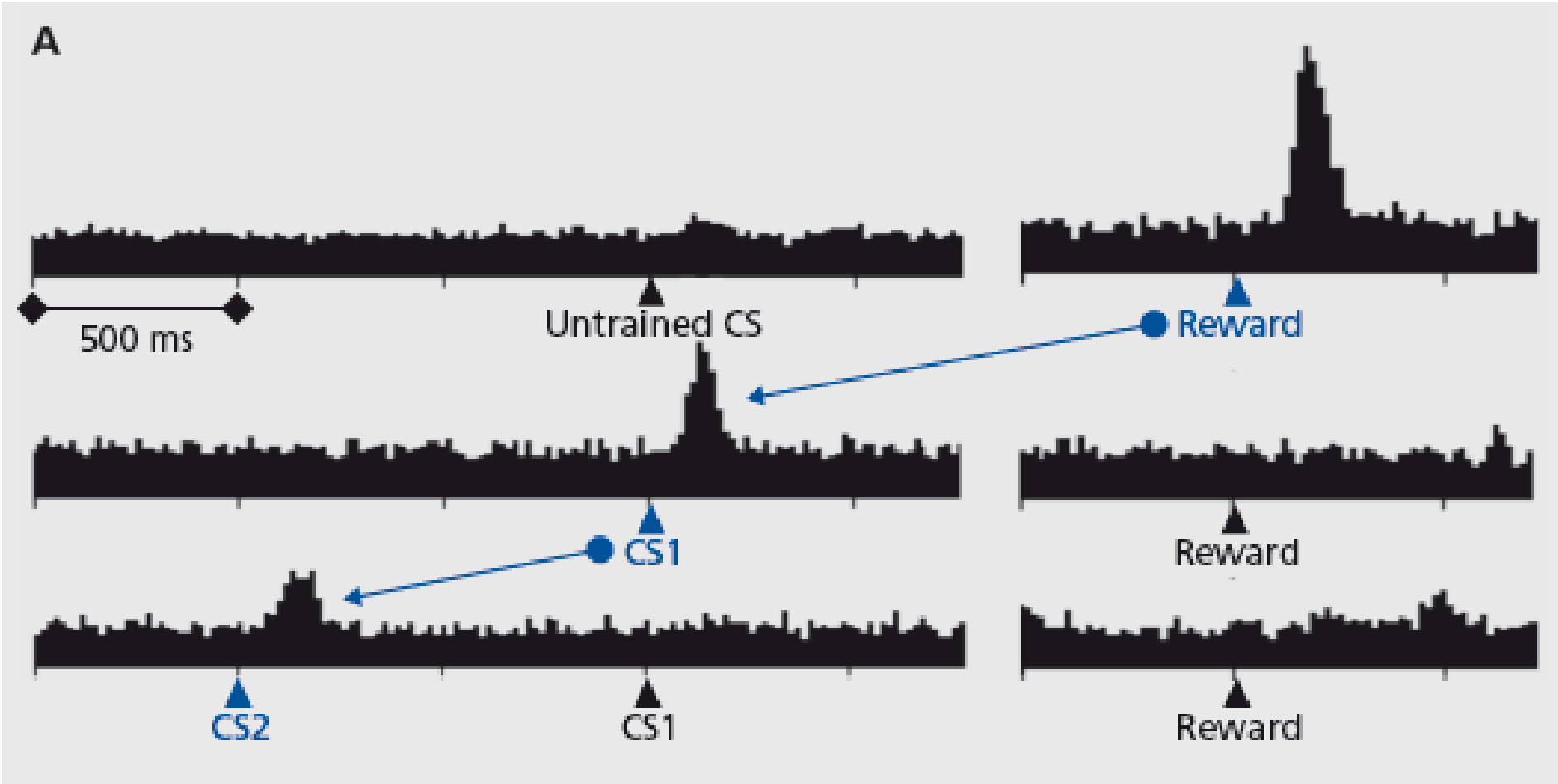
$$r(t) = 0, \forall t$$
$$\delta_{cue} = 0 + \gamma V - 0 = \gamma V$$
$$\delta_{reward} = 0 + \gamma 0 - V = -V$$



Single Dopaminergic Neuron Electrophysiology 3

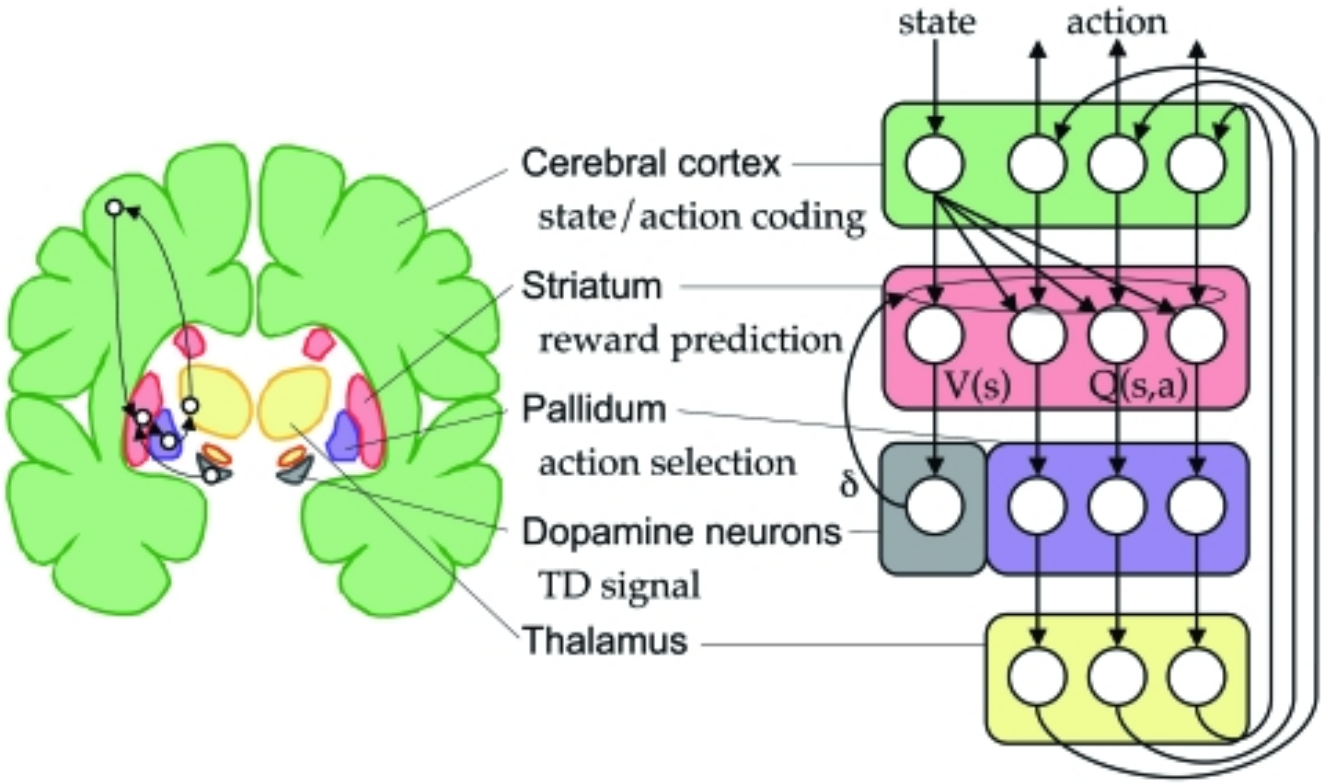
- In untrained monkeys V(t) shows no increase as there is no expectation for reward. Indeed, in the case of $\delta(t)$, the temporal difference error, we can see a surge corresponding to the time of an unexpected reward.
- Conversely, for the trained brain, the expectation will occur upon being subjected to a CS, thus agreeing with the electrophysiological data that show a dopamine spike at the time of the CS but not US in the trained brain.
- The same reasoning can apply to justify the fall seen in dopamine activity after the monkey does not receive the juice. It is a mechanism that serves to adjust the future expectations (V_{n+1}) of the monkey when subjected to CS_{n+1}

Changes in dopamine signalling during training



from Enomoto et al. (2011) Dopamine neurons learn to encode the long-term value of multiple future rewards.

Our current understanding of the TD error transmission in the brain is that the abstract state is represented in the cortex. It is interesting to note that the state value is represented on a universal scale in the human mind and the different nature of the expected rewards in the real world (i.e. political, social, economic, emotional, physical) is not a factor for the cortex, everything being judged as a scalar on a universal axis in the brain and compared as such. The information then follows the following route:

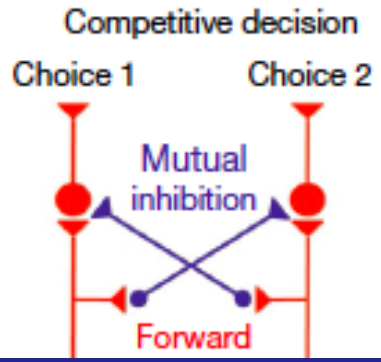


Winner-take-all decision making

When faced with a choice in between multiple potential courses of action, studies carried *in sillico* by Berridge (2012) as well as animal work by Yager & Robinson (2010) or Saunders & Robinson (2011) have pointed out that our brains are wired so as to experience a "motivational magnet" phenomenon.

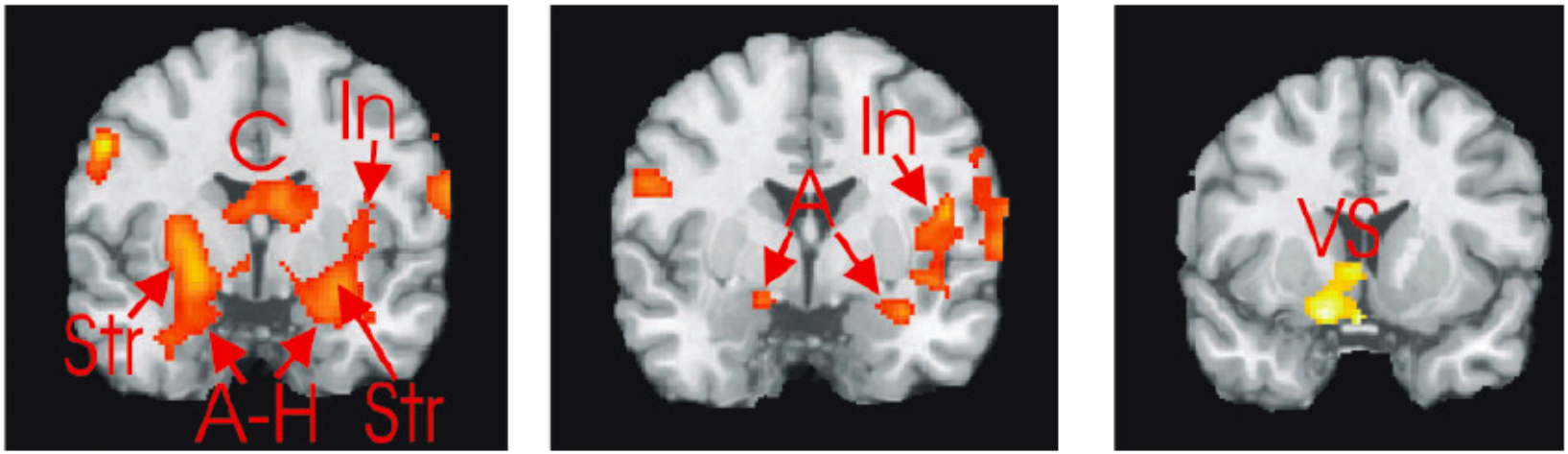
In other words, at the neural level, individual neurons coding for separate decisions will adopt a winner-take-all mechanism of action, with the neuron transmitting the decision of the winning course of action inhibiting all others.

For 2 hypothetical decision paths, this inhibition originally starts as an equal mutual inhibition of 2 separate neurons. Local fluctuations in the strength of the action potentials relayed through them will determine which one takes over and wins.



fMRI studies of disease states with regard to reward processing

Currently fMRI studies are used to map blood flow to particular areas of the brain associated with the various components of reinforcement learning. These can be used to better understand the components involved in signalling as well as for better understanding of the pathologies with an outlook towards future treatments



Normal | Depression | Schizophrenia

Psychedelic drugs and Reward processing

- Comparing healthy individuals with people in a state of mental impairment can be very informative as diseases such as depression, anxiety or schizophrenia are all thought to alter dopamine signalling in a way or another. Downregulation of signalling in depression is able to explain the lack of motivation and the inability of depressed people to react to reward stimuli in the same way control subjects do.
- The complete lack of dopamine signalling in the midbrain, cortex and striatum in schizophrenia explains the altered reality of a schizophrenic person, as dopamine modulating drugs have been implicated in hallucinatory states (i.e. psilocybin, mescaline, LSD)
- These substances can allow a top-down approach that helps isolate the important parameters and variables existent in complex mechanisms such as reward processing and greatly contribute to the understanding of the human brain and psyche.

frametitle¡presentation¿Bibliography


D. Silver
Reinforcement Learning Course
University College London, 2015


R.S. Sutton, A.G. Barto
Reinforcement Learning: An Introduction, 2nd ed. (draft)
MIT Press, 2017


K.C. Berridge
From prediction error to incentive salience: mesolimbic computation of reward motivation
Eur J Neurosci. 35(7): 11241143, 2012


K. Doya
Reinforcement learning: Computational theory and biological mechanisms
HFSP Journal, 1(1):3040, 2007


V.B. Gradin, P. Kumar, G. Waiter, T. Ahearn, C. Stickle, M. Milders, I. Reid, J. Hall, J.D. Steel
Expected value and prediction error abnormalities in depression and schizophrenia
Brain. A Journal of Neurology, 134, 1751176 (2011)


G. Morris, A. Nevet, D. Arkadir, E. Vaadia, H. Bergman
Midbrain dopamine neurons encode decisions for future action
Nature Neuroscience 9, 1057 - 1063 (2006)


W. Schultz
Reward in A.W. Toga (ed.), *Brain Mapping: An Encyclopedic Reference*, vol. 2, pp. 643-651
Elsevier, 2015