Reinforcement Learning

Lecture 12

Connections to Psychology and Neuroscience

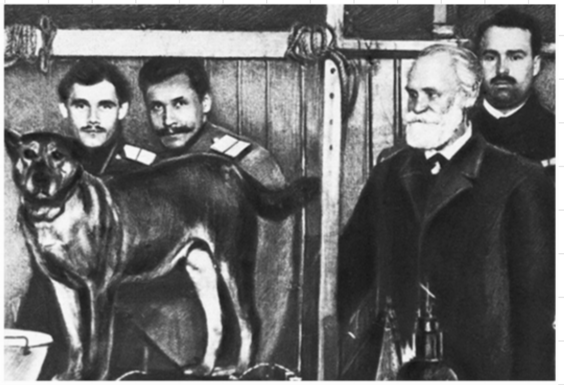
Robert Peharz

Institute of Theoretical Computer Science
Graz University of Technology
Winter Term 2023/24

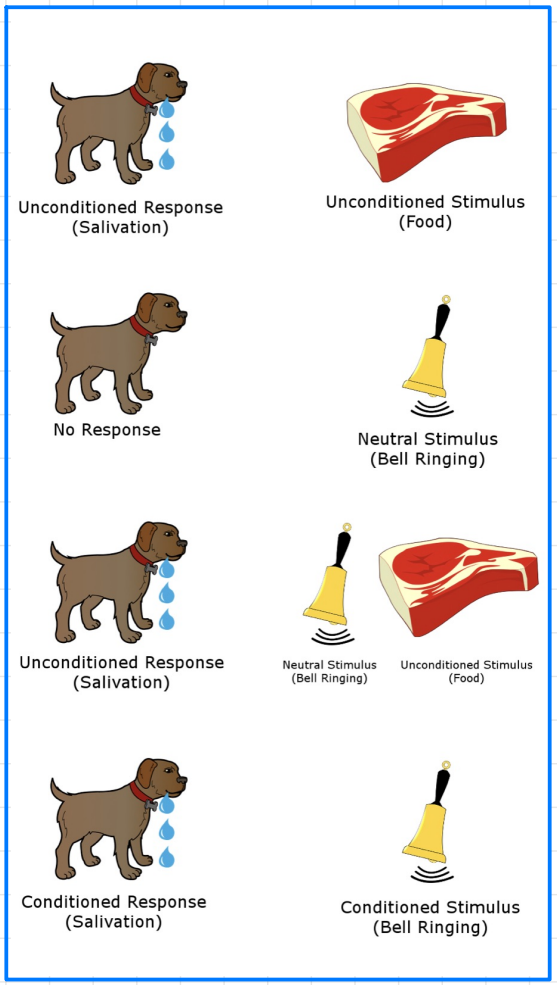
Reinforcement Learning and Psychology

Classical Conditioning

- classical conditioning is a widely studied type of learning in psychology
- Pavlov's dog (1927)
  - a dog salivates when presented food
  - it does not respond with salivation to a neutral stimulus, like a bell ringing
  - however, when the bell ringing occurs consistently before being presented food, the dog will produce salvia when hearing the bell alone
  - the dog has learned to associate the bell with food, it gets conditioned on the bell



thepsychologist.bps.org.uk



wikipedia.org

Classical Conditioning

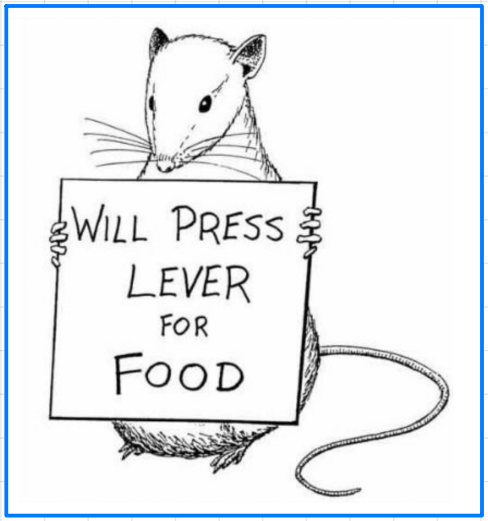
- classical conditioning observed in practically all animals, at varying time scales
- unconditioned stimulus (US)
  stimulus causing a natural response, e.g. food, electric shock, etc.
- unconditioned response (UR)
  natural response following US, e.g. salivation
- conditioned stimulus (CS)
  initially neutral stimulus (e.g. bell) which is associated to US
- conditioned response (CR)
  natural response (UR) getting associated to CS

pinterest.ie

Instrumental Conditioning

- classical conditioning investigates ability to predict US from CS
- instrumental conditioning: experiment set up such that the animal receives reward (or punishment) depending on its behavior
- in terms of reinforcement learning

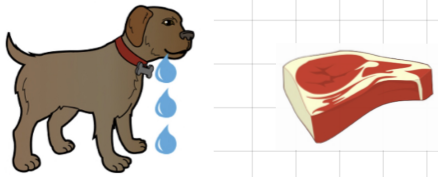
> classical conditioning ≈ prediction
>
> instrumental conditioning ≈ control



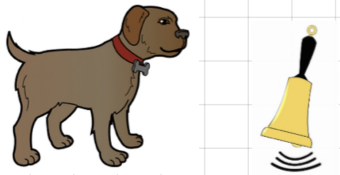
- "reinforcement" in psychology used to describe both classical and instrumental setting

Higher-Order Conditioning

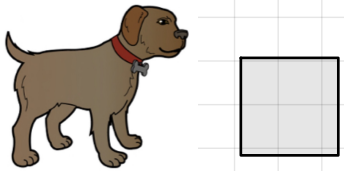
Food - US, UR



Bell - neutral



Black Square - neutral



(First order) conditioning

Second order conditioning

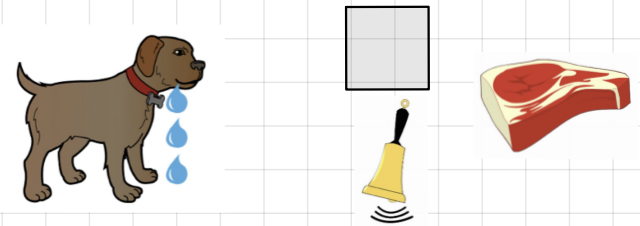
animal produces CR on secondary CS, even though US never followed !

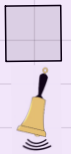
- Higher order conditioning can also be demonstrated
- (challenging, since lower CS lose their reinforcing properties when not followed by original UR)

Blocking

First phase: condition on first neutral stimulus (bell)

Second phase: add simultaneous second neutral stimulus (square)



compound stimulus: 

Third phase: no (or little) conditioning on second stimulus!
first stimulus has blocked second stimulus

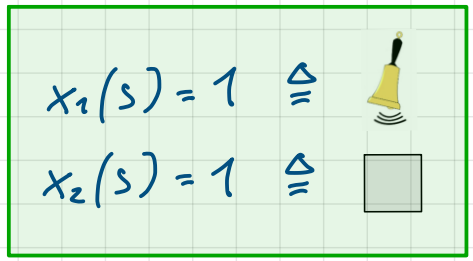
How to explain this?

Rescorla-Wagner Model

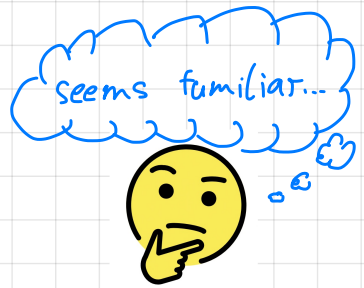
(adapted notation)

- highly influential in animal learning
- "mechanistic" model explaining blocking (no higher cognitive model)
- Key assumption: animal learns when expectations violated ("surprise")
- trial-type (state): s
- introduce binary features, indicating stimulus:

$$x_i(s) = \begin{cases} 1 & \text{if } i^{th} \text{ stimulous is present in } s \\ 0 & \text{o.w.} \end{cases}$$



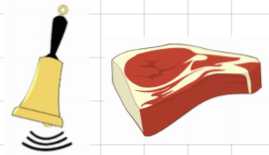
- R: "associative strength" between US and UR
- w_i: associative strength between CS and CR (= UR)
- aggregate associative strength $\hat{v}(s, \underline{w}) = \underline{w}^T \underline{x}(s)$
- update: $\underline{w} \leftarrow \underline{w} + \alpha \underbrace{(R - \hat{v}(s, \underline{w}))}_{\delta}$ with stepsize α

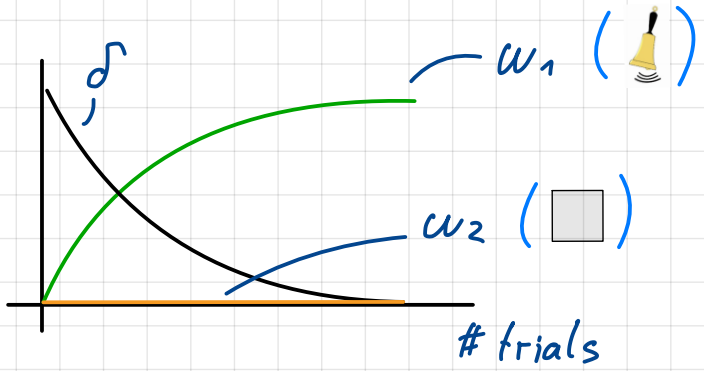


Rescorla-Wagner Model: Explaining Blocking

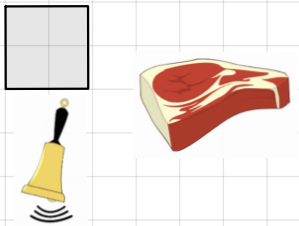
$$\underline{w} \leftarrow \underline{w} + \alpha \delta$$

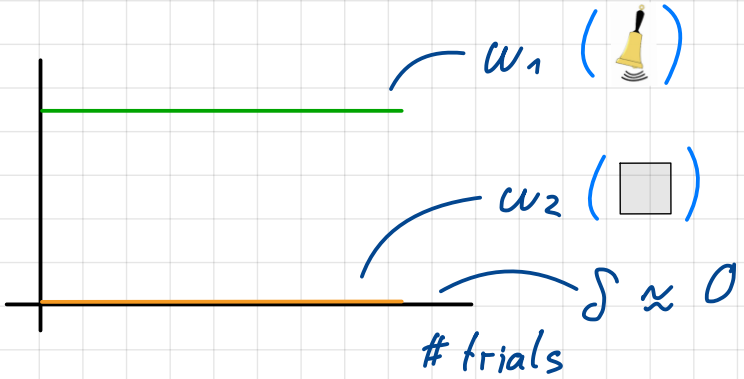
First phase





Second phase

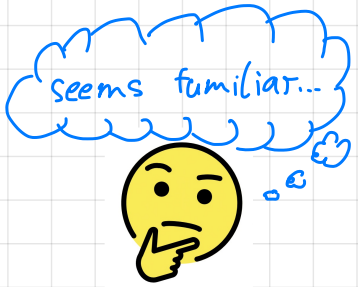




Third phase

Does not explain higher-order conditioning, though

Temporal Difference Model



Rescorla-Wagner $\underline{w} \leftarrow \underline{w} + \alpha (R - \hat{v}(s, \underline{w}))$

Temporal Differences Model

introduces discretized time t

- uses eligibility traces $\underline{z}$
- $\delta = R + \gamma \hat{v}(S_{t+1}, \underline{w}) - \hat{v}(S_t, \underline{w})$

  $\underline{z} \leftarrow \gamma \lambda \underline{z} + \underline{x}(S_t)$

  $\underline{w} \leftarrow \underline{w} + \alpha \delta \underline{z}$
- for $\gamma = 0$, same as Rescorla-Wagner, but with different "meaning of time"

A Temporal-Difference Model of Classical Conditioning

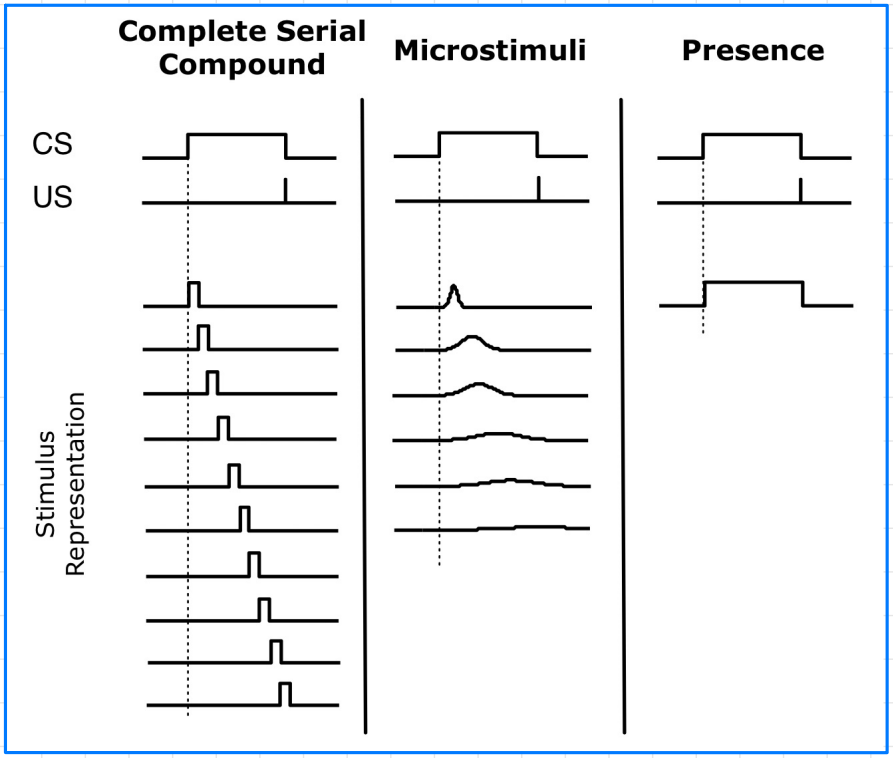
Richard S. Sutton
GTE Laboratories Incorporated

Andrew G. Barto
University of Massachusetts

Abstract—Rescorla and Wagner's model of classical conditioning has been one of the most influential and successful theories of this fundamental learning process. The learning rule of their theory was first described as a learning procedure for connectionist networks by Widrow and Hoff. In this paper we propose a similar confluence of psychological and engineering constraints. Sutton has recently argued that adaptive prediction methods called *temporal-difference methods* have advantages over other prediction methods for certain types of problems. Here we argue that temporal-difference methods can provide detailed accounts of aspects of classical conditioning behavior. We present a model of classical conditioning behavior that takes the form of a temporal-difference prediction method. We argue that it is an improvement over the Rescorla-Wagner model in its handling of within-trial temporal effects such as the ISI dependency, primacy effects, and the facilitation of remote associations in serial-compound conditioning. The new model is closely related to the model of classical conditioning that we proposed in 1981, but avoids some of the problems with that model recently identified by Moore et al. We suggest that the theory of adaptive prediction on which our model is based provides insight into the functionality of classical conditioning behavior.

Stimuli Representations

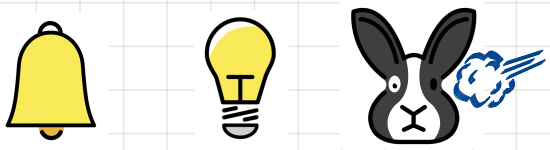
- assuming TD, how does an animal (brain) represent stimuli?
- commonly assumed representations:



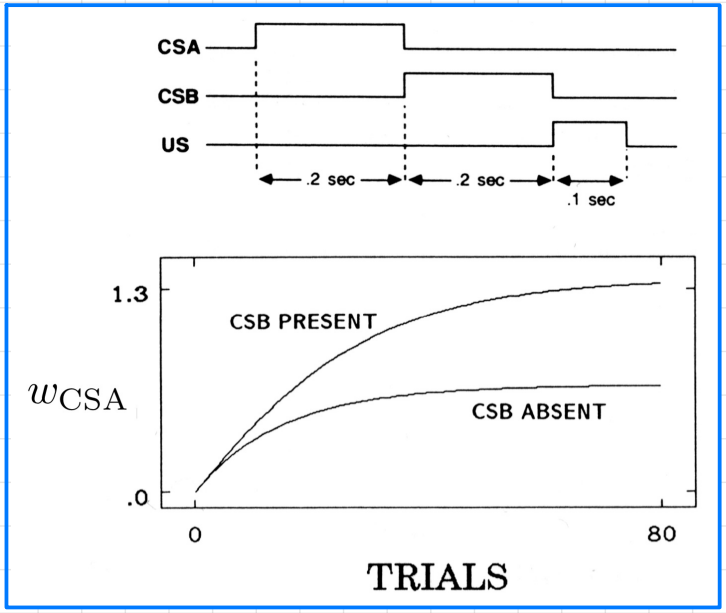
- CSC and Presence encodings biologically not very plausible

TD Simulations 1: Remote Associations

- remote association is facilitated by an intermediate stimulus
- experiment (Kehoe, 1982)
  - rabbit gets exposed to sound (CSA)
  - rabbit gets exposed to light (CSB)
  - puffed air in rabbit's eye (US)
  - rabbit closes eye (UR)
- conditioning on CSA is stronger, when CSB is present, and weaker when absent
- consistent with simulated TD



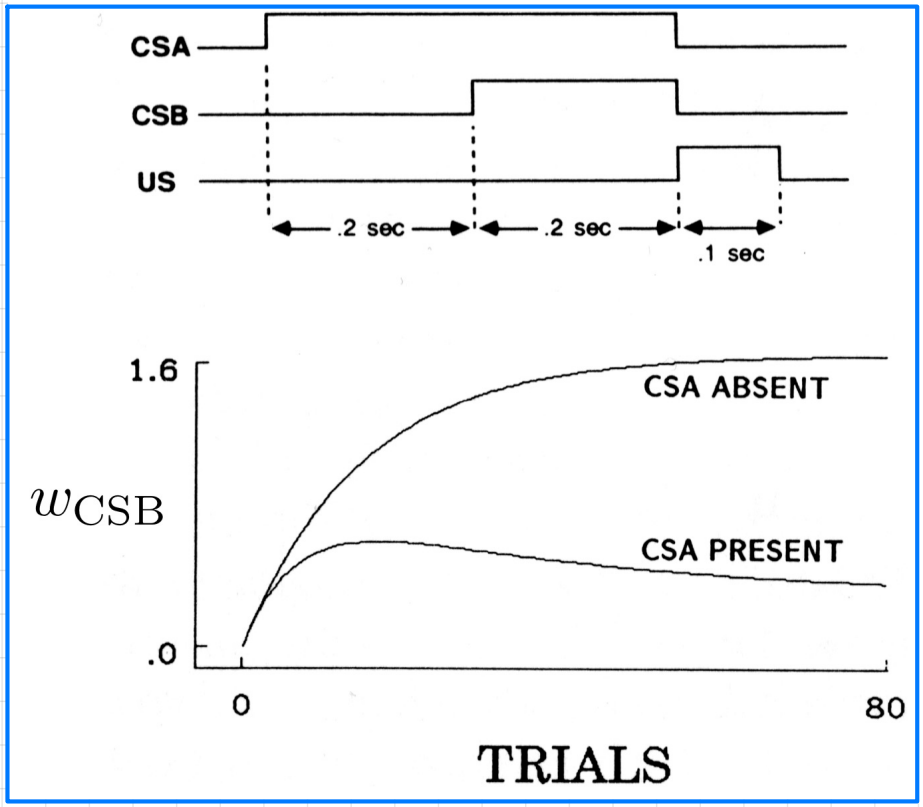
CSA CSB US



Simulation

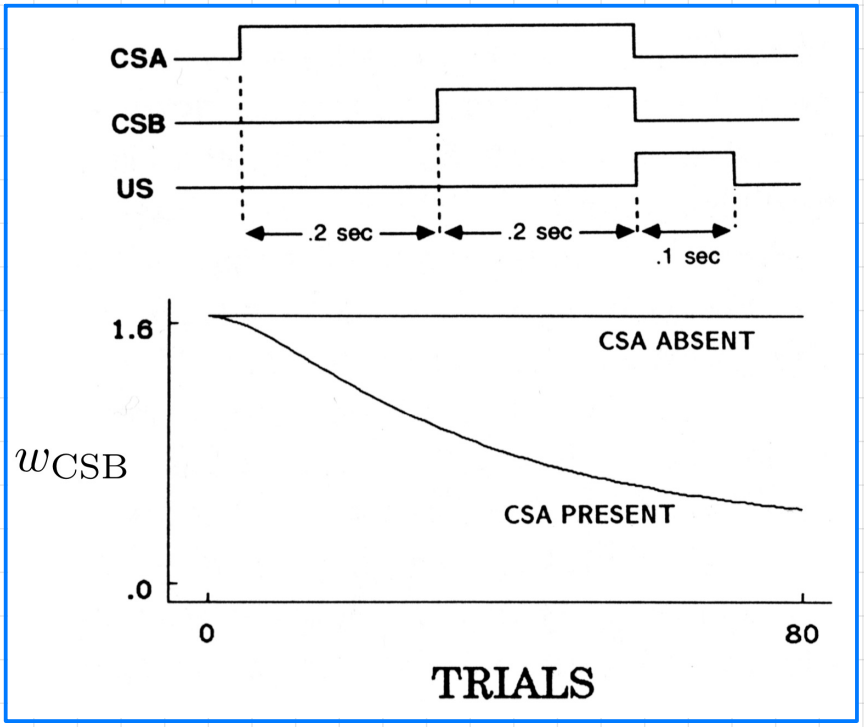
TD Simulations 2: Egger-Miller Effect

- Egger and Miller (1962) found that conditioning on a stimulus is weaker when preceded by another predictive stimulus
- this effect is also consistent with TD simulations:



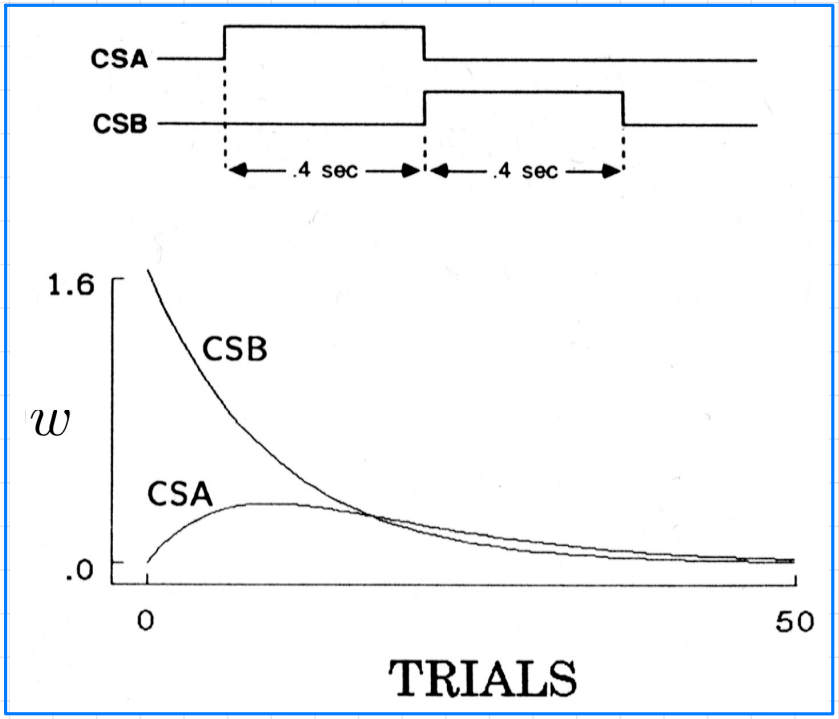
TD Simulations 3: Blocking and Reversed Blocking

- TD is consistent with blocking, i.e. when the association between CSB and US has been fully learned, a simultaneous CSA will be blocked by CSB
- however, TD predicts that blocking is reversed, when CSA is moved before CSB, since CSA becomes predictive for CSB
- different from Egger-Miller, since CSB was conditioned before introducing CSA
- effect was unknown before and experimentally confirmed by Kehoe et al. (1987)!



TD Simulations 4: Higher-order Conditioning

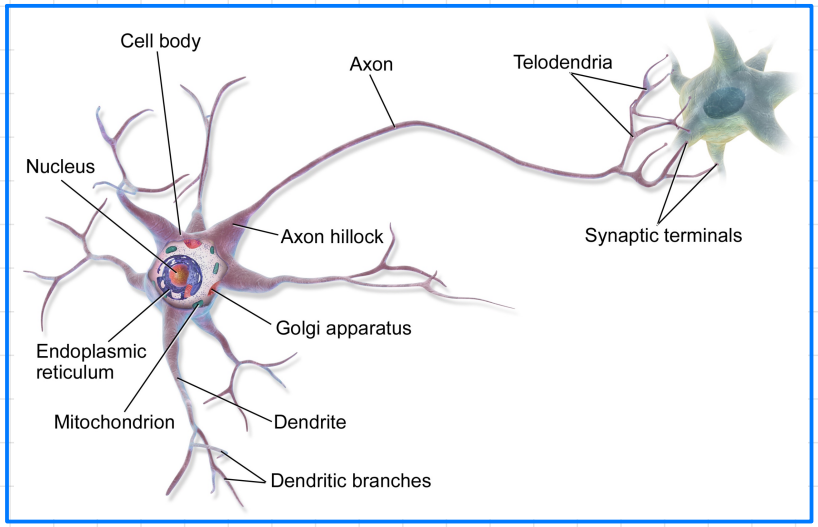
- main criticism of Rescorla-Wagner model: does not predict higher-order conditioning
- TD, however, does predict it
- first, condition CSB on US
- then omit US and condition CSA on CSB
- higher-order conditioning as bootstrapping bias !



Reinforcement Learning and Neuroscience

Neurons

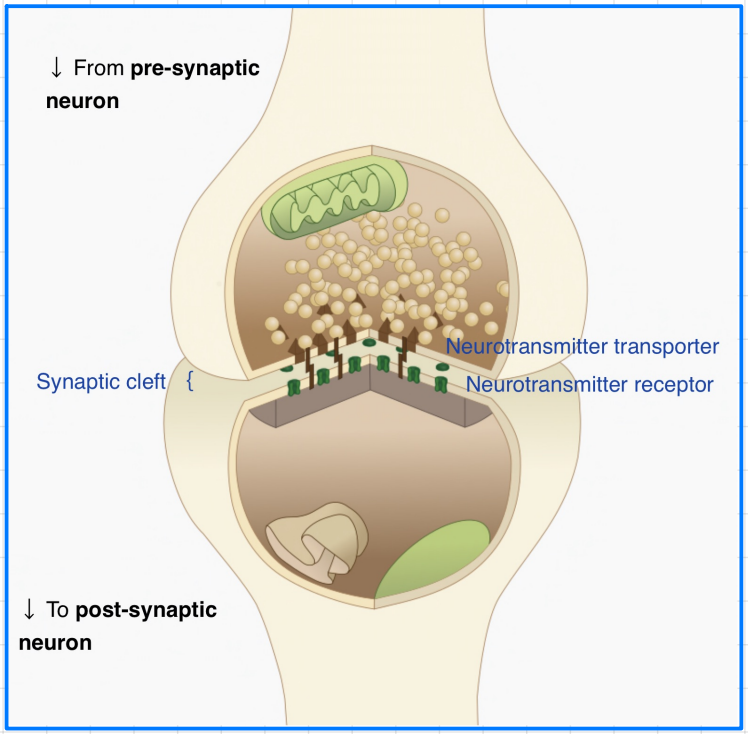
- dendrites: receive inputs from other neurons
- axon: single fiber carrying output to other neurons, the action potentials (spikes)
- axonal arbor: branching structure of axon, can reach many thousands of other neurons
- firing rate: average number of spikes per time unit
- synapse: mediates communication between neurons
- background activity: basic activity (firing) when neuron is not particularly excited
- phasic activity: bursts of spiking activity caused by synaptic input



wikipedia.org

Synapses

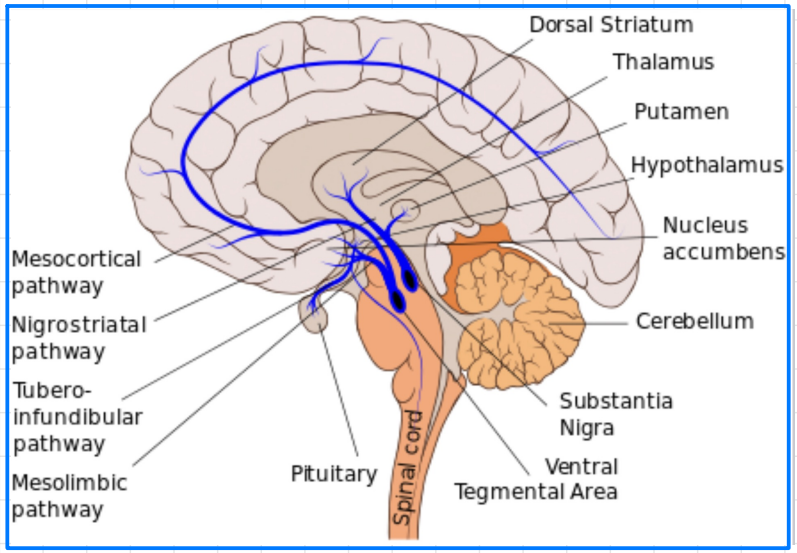
wikipedia.org

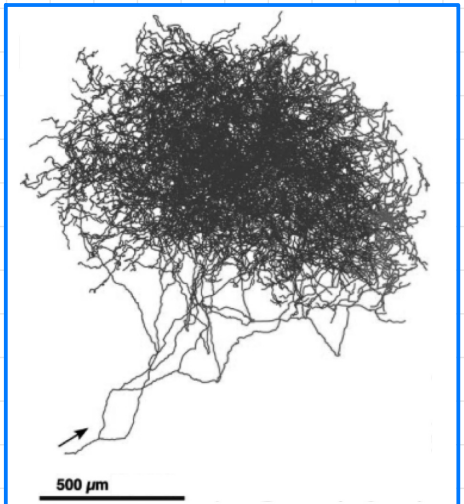


- synapses mediate communication between neurons
- pre-synaptic neuron sends spike, causing synapse to release neuro-transmitter (certain chemicals) which diffuse via synaptic cleft, which may bind to receptors on receiving neuron
- coupling strength (efficacy) of synapse can be modulated → responsible to a large part for learning
- there are many different neurotransmitters and receptors
- some neurotransmitters modulate nervous system, i.e. influence it on a large scale (e.g. for learning): neuromodulators

Dopamine

- dopamine is a neuromodulator connected with reward signaling, motivation, learning, addiction, Parkinson's disease, Schizophrenia, ...
- mainly produced by neurons in Substantia Nigra and Ventral Tegmental Area
- dopamine neurons have large axonal arbors, reaching 100-1000 times more neurons than usual (suitable for global signaling)
- most of them make synaptic contact with frontal cortex and basal ganglia, which are associated with voluntary movement, decision making, and cognitive functions





Axonal arbor of a single neuron producing dopamine as a neurotransmitter. These axons make synaptic contacts with a huge number of dendrites of neurons in targeted brain areas.
Adapted from *The Journal of Neuroscience*, Matsuda, Furuta, Nakamura, Hioki, Fujiyama, Arai, and Kaneko, volume 29, 2009, page 451.

Reward Prediction Error Hypothesis

- in RL algorithms, a single reward R_t determines central objective
- unlikely to be found in the brain...
- a reinforcement signal directs learning in an algorithm
- e.g. TD error $\delta_t = R_t + \gamma v(S_{t+1}) - v(S_t)$
- more generally, any reward prediction error (RPE)
- "Reward Prediction Hypothesis of dopamine neuron activity" proposes that one of the functions of dopamine-producing neurons is to communicate RPEs to relevant brain areas

- unpredicted reward ⟶ phasic response of dopamine-neurons
- predicted reward ⟶ background activity of dopamine-neurons
- reward predicted, but omitted ⟶ drop below background activity

Experimental Evidence

(Romo & Schultz, 1990)

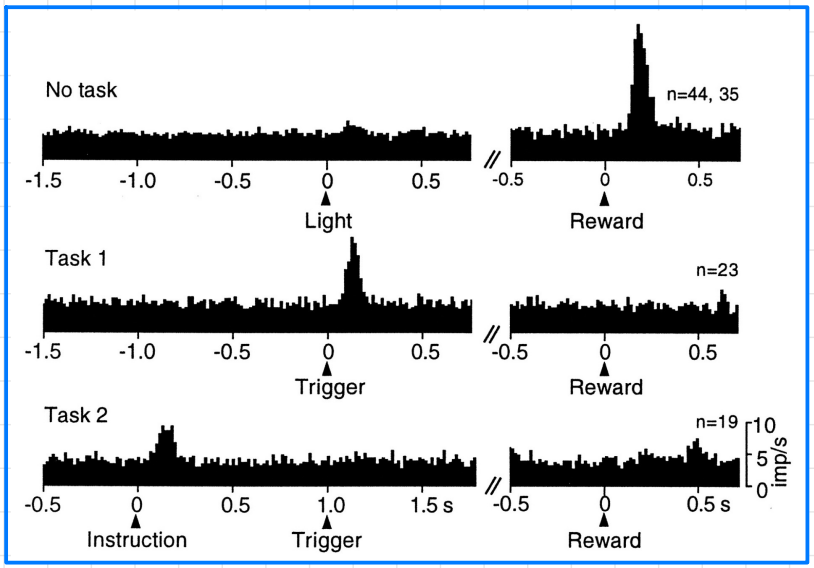
- trained two monkeys to search for food in bin, covered from above (so monkey could not see inside, but reach in)
- some dopamine-neurons were recorded
- first, neurons produced phasic responses when food was first touched
- later, they responded to sight and sound of bin opening
- this indicates conditioning

Experimental Evidence

(Schultz et al., 1995)

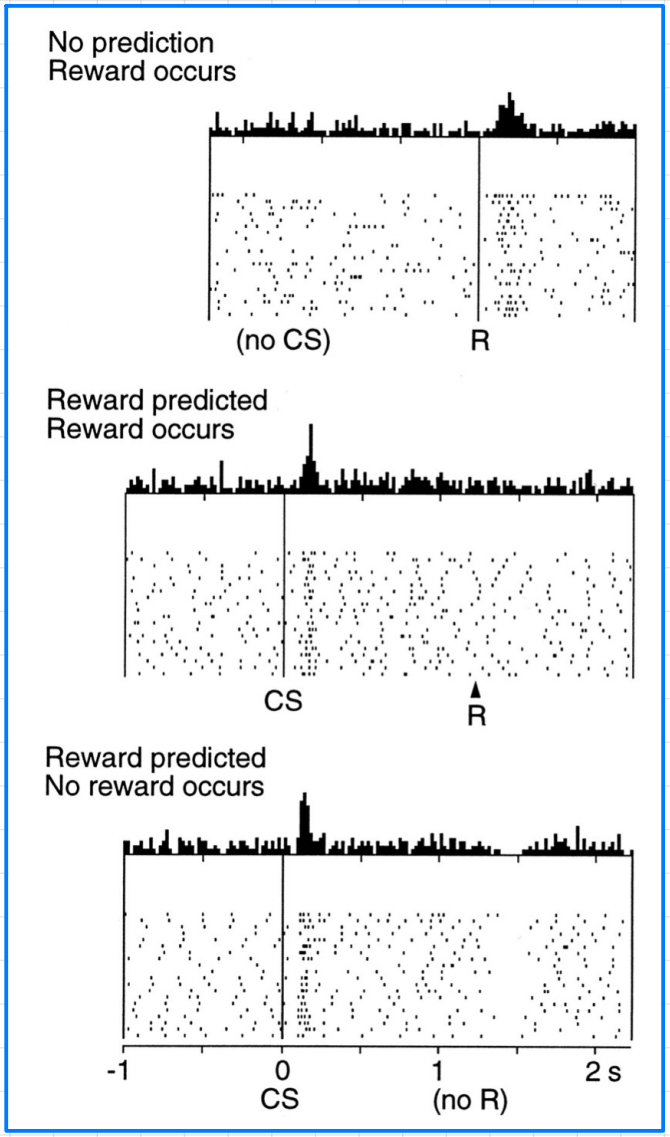
- in another experiment, monkeys were trained to depress a lever, in order to get a reward (juice)
- the task was indicated with a trigger (light) (Task 1)
- dopamine-neurons fired first upon the reward, but over time, response was shifted to trigger signal
- monkeys were subsequently trained on an "instruction cue" (light over one of two levers), indicating which lever would yield a reward (Task 2)

response shifted gradually to instruction cue



Experimental Evidence

- dopamine-neurons become phasic as soon as unexpected reward is predicted, not when it is received
- sometimes monkeys press wrong lever, thus they don't get a reward
- in these cases, fire rate falls below background activity



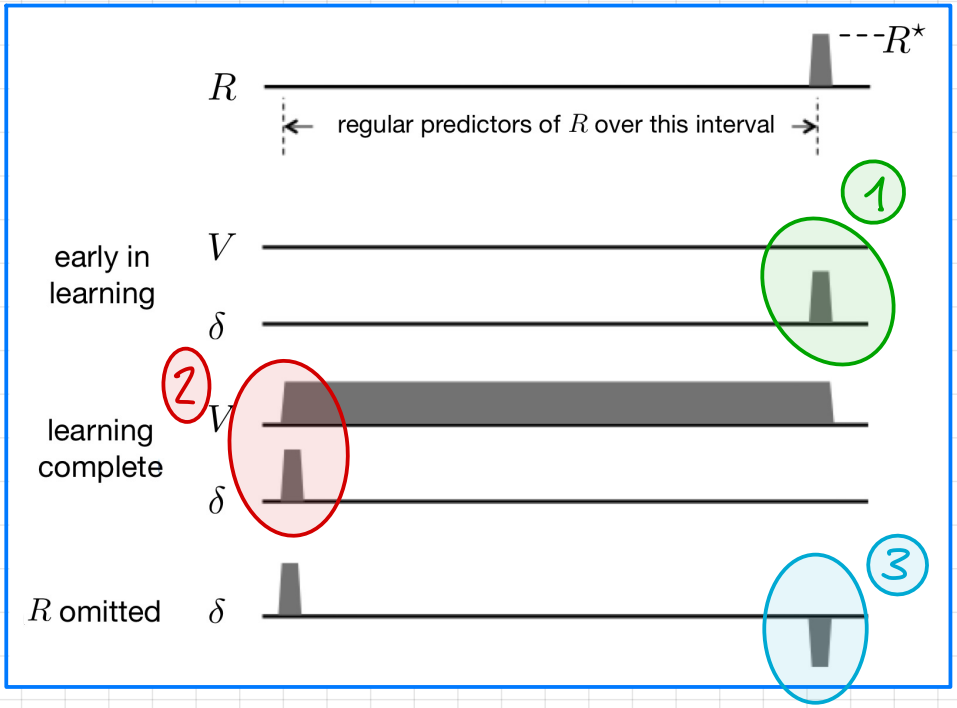
Correspondence Dopamine and TD

- the behavior of dopamine-neurons corresponds to TD error:

$$\delta_t = R_{t+1} + \hat{v}(S_{t+1}) - \hat{v}(S_t)$$

- early in learning, δ_t goes up when receiving reward, since it is not predicted:

$\delta_t = 1 + 0 - 0$ (1)



- when prediction is learned: $\delta_t = 0 + 1 - 0$ (2)
- when reward is omitted: $\delta_t = 0 + 0 - 1$ (3)

Early Reward

- main discrepancy between TD and dopamine when a reward comes earlier than expected
- both react with an increase upon the reward
- however, the TD error becomes negative at the time where the reward was originally expected, while the dopamine-neurons remain in their background activity
- thus, TD "expects second reward"
- explainable by higher cognitive model, object permanence, counting, etc., in real neural system