Reinforcement Learning

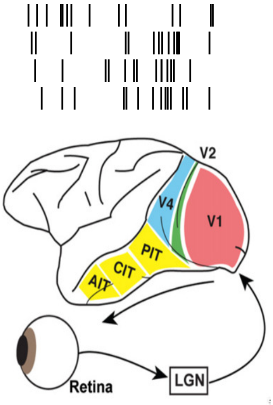
Informatics 1 Cognitive Science

Matthias Hennig

School of Informatics
University of Edinburgh
mhennig@inf.ed.ac.uk

From Stimulus to Action

Sensory input



Motor output

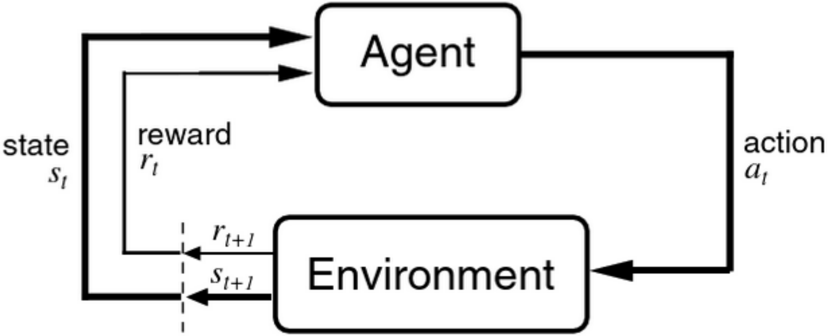
Usually there are many stimuli, and many possible actions. How to decide which action to take?

Reinforcement Learning: Aims

- To keep track of complex, high-dimensional environments (states).
- To bridge time delays between action(s) and outcomes.
- To assign value to actions/states and remember these to choose appropriate actions.

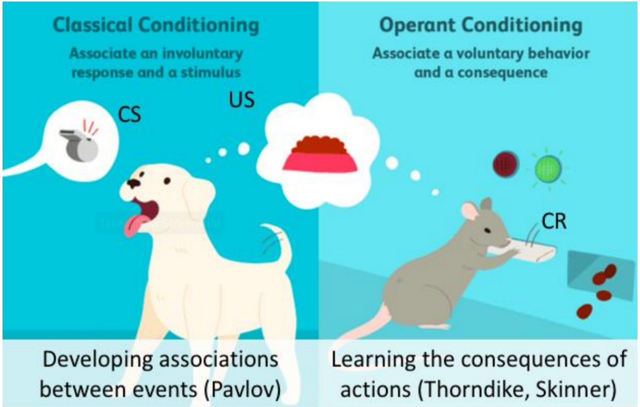
(Image is an example from Google Research, 2019)

Reinforcement Learning: Approach



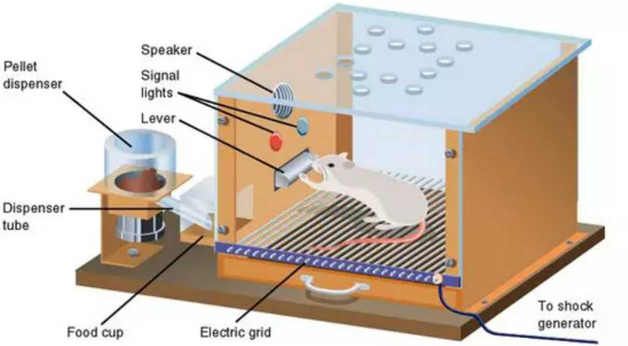
- RL agents have explicit goals, manifested through rewards (or punishments).
- RL agents act on the environment and collect information to inform the next action.

Learning associations: Classical and Operant Conditioning



Classical conditioning: unconditioned stimulus (food) - conditioned stimulus (bell)
Operant conditioning: an action / conditioned response (lever press) - reward (food)

Operant Conditioning



Law of effect (Edward Thorndike):
CR more frequent (rare) when it elicits a positive (negative) stimulus.
Reinforcers: Stimuli increasing behaviour rate
Punishers: Stimuli decreasing behaviour rate

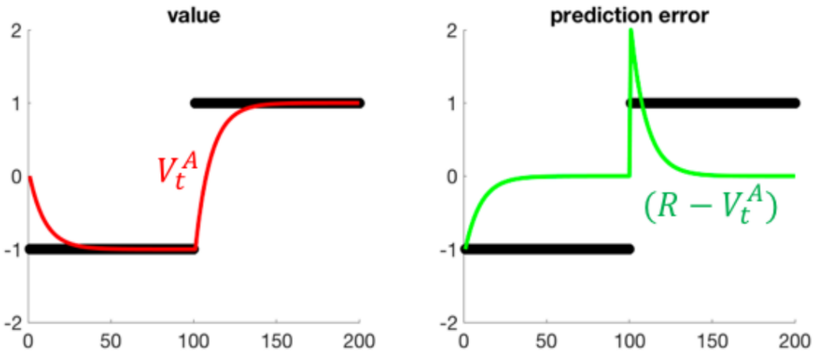
Predicting action outcomes (Rescorla-Wagner Rule)

The *value* V^A of action A over time changes according to:

$$V_t^A = V_{t-1}^A + \alpha(R - V_{t-1}^A)$$

R is the reward, α the learning rate, and $\delta = R - V_{t-1}^A$ is called the *prediction error*.
Also called: δ-*rule*.

The Rescorla-Wagner Rule for Conditioning

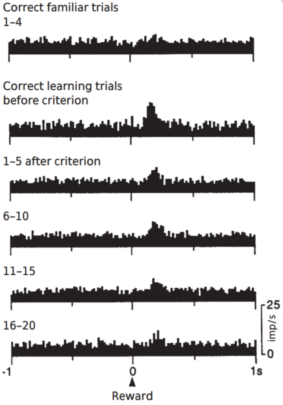


$$V_t^A = V_{t-1}^A + \alpha(R - V_{t-1}^A)$$

First R was set to -1, and then changed to $+1$ ($\alpha = 0.1$).
So the key quantity of the model is the prediction error. Is it computed in the brain?

Prediction Errors in the Brain

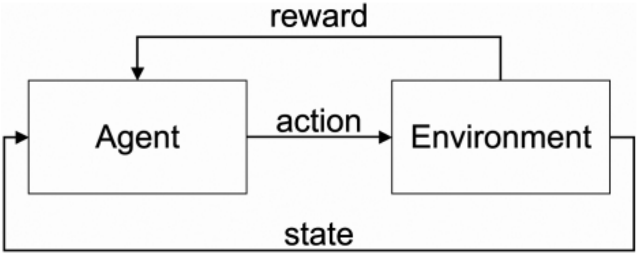
Fig. 4. Changes of average population response (54 neurons tested) to reward during learning. Trials with familiar pictures are also shown for comparison. Note the absence of response to reward with familiar pictures, strong activations during initial learning trials before reaching the criterion and progressive decrease after the criterion. Learning data are from episodes with at least 20 correct trials. Average population activity is shown for the first five trials (familiar pictures, top panel), for the total set of correct trials before criterion (second panel) or for sets of 5 consecutive correct trials at different stages (first to fifth, sixth to tenth, etc.) after criterion was reached (bottom four panels).



Visual discrimination task with reward: Dopamine neurons of the substantia nigra appear to signal the prediction error as predicted by the Rescorla-Wagner rule.

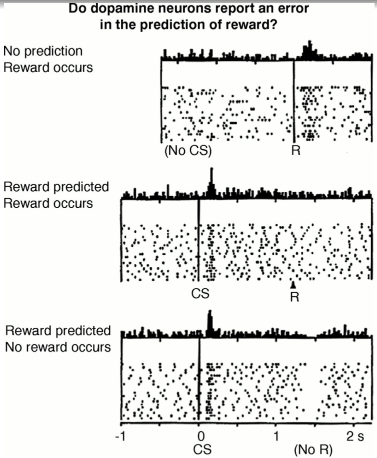
Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. Nature neuroscience, 1(4), 304-309.

Delayed rewards are difficult in practise



How to know which past action was responsible for an observed outcome? This is called the *temporal credit assignment problem*.

Reward Learning in the VTA



First the neurons signal a prediction error. Once learned, the same neurons now signal reward at the time of the cue.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. Journal of Neurophysiology.

TD learning

We have states s_i, rewards r_t and the value of the states $V(s)$.

Prediction error:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

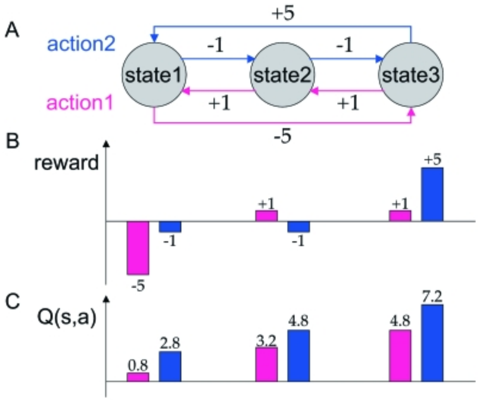
Future potential rewards are taken into account, but discounted by γ.

Value update:

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

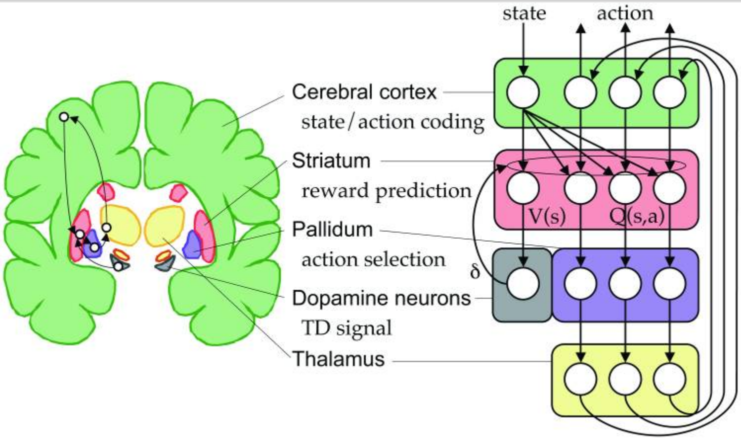
This iteratively learns a stable state / value map (but takes time).

Q-Learning: TD Learning + choosing actions



Discount factor: $\gamma = 0.8$. For state 2, it is now better to accept negative reward first, as more reward is on the horizon later.

Action Learning in the Brain

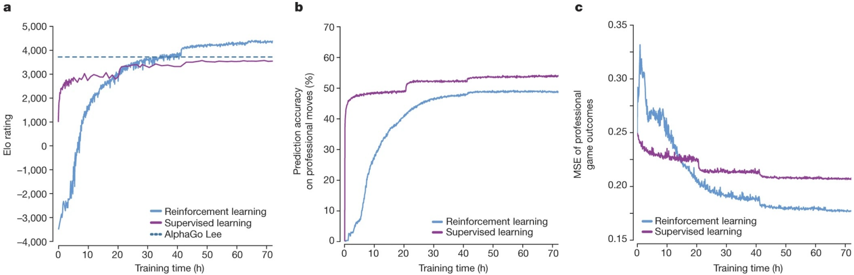


Dopamine neurons in the midbrain: prediction errors δ
Striatum: Value
Cortex: World

Reinforcement Learning Successes

- Learns to play complex board and video games (and even learns rules).
- Fine-tune large language models (ChatGPT).

AlphaGo Zero



- Knows game rules.
- Trained by playing against itself on 64 GPU workers and 19 CPU parameter servers. Best models are used as new opponents for self-play.
- 0.4s thinking time/move, 4.9 million games played; 216,000 moves/day. About 3 days training in total.
- Comparison to human-trained supervised model (move prediction).

Summary

- Evidence that the brain learns to predict the outcomes of actions and stimuli.
- Responses corresponding to prediction errors are observed in the dopaminergic system.
- RL assumes that prediction errors reflects the learning of goal-directed behaviour, represented through value.
- RL finds the behaviours that maximises value.
- This works well in simple examples, but is data-inefficient for problems with real world complexity.