Reinforcement Learning
Informatics 1 Cognitive Science

Matthias Hennig

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

$$V^A_t = V^A_{t-1} + \alpha (R - V^A_{t-1})$$

$R$ is the reward, $\alpha$ the learning rate, and $\delta = R - V^A_{t-1}$ is called the prediction error. Also called: $\delta$-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?
Visual discrimination task with reward: Dopamine neurons of the substantia nigra appear to signal the prediction error as predicted by the Rescorla-Wagner rule.

Delayed rewards are difficult in practice

How to know which past action was responsible for an observed outcome? This is called the *temporal credit assignment problem.*
First the neurons signal a prediction error. Once learned, the same neurons now signal reward at the time of the cue.

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 $\gamma$.

Value update:

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

This iteratively learns a stable state / value map (but takes time).
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.
Dopamine neurons in the midbrain: prediction errors $\delta$

Striatum: Value

Cortex: World
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).
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 reflect the learning of goal-directed behaviour, represented through value.

RL finds the behaviours that maximise value.

This works well in simple examples, but is data-inefficient for problems with real world complexity.