On neural correlates of reinforcement learning

the role of dopamine in planning and action

Genela Morris
Dept. of Neurobiology and Ethology
Haifa University
Overview

- Learning through reward – reinforcement learning
- Midbrain dopamine neurons and temporal difference (TD) learning
- ACh in the striatum
- Dopamine neurons and their impact on decision behaviour
- Implications for computational models of action selection
Reinforcement learning

the basics

Supervised learning –
all knowing teacher, detailed feedback

Reinforcement learning –
scalar (correct/incorrect) feedback

Unsupervised learning –
self organization
Reinforcement learning: The law of effect

“The Law of Effect is that: Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur”

Edward Lee Thorndike (1911)
Early attempts at modeling

By associative rules

Classical conditioning
Classical conditioning

The Elements:
**US**: Unconditioned stimulus
**UR**: Unconditioned response
**NS**: Neutral stimulus
**CS**: Conditioned stimulus
  - **CS1**: Conditioned stimulus 1
  - **CS2**: Conditioned stimulus 2
**CR**: Conditioned response
Properties of classical conditioning

(Pavlov 1927)

- Acquisition.
- Partial Reinforcement
- Generalization.
- Interstimulus Interval (ISI) effects.
- Intertrial Interval (ITI) effects.
A simple association (coincidence, Hebbian) model can explain the phenomenon.
Properties of classical conditioning

(Cnt’d)

CS must RELIABLY predict US

- Conditioned Inhibition
- Relative validity (Wagner 1968).
- Blocking (Kamin 1968)

…
Which simple association can’t explain

Learning occurs not because two events co-occur, but because that co-occurrence is UNPREDICTED
Rescorla-Wagner rule (1972)

Learning to predict reward R given stimulus U=1

Goal: Form a prediction of the reward V of the form:

\[ V = \omega U \]

And learn to change \( \omega \) :

\[ \Delta \omega = \varepsilon (R-V)U \]

After learning of consistent pairing: \( \omega = R \)

Where:
- \( U = \text{CS availability (0,1)} \)
- \( V = \text{reward prediction} \)
- \( R = \text{reward availability (0,1)} \)
- \( \omega = \text{weight of the connection between U and V} \)
- \( \varepsilon = \text{learning rate} \)
- \( R-V = \text{prediction error} \)
Blocking with Rescorla Wagner

Given U1, U2 and R, after U1 has been learnt:

\[ \omega_1 = R \]

\[ V = \omega_1 U_1 + \omega_2 U_2 \]

Prediction error: \( R - V = 0 \)

And no learning occurs for \( \omega_2 \)
Critical problems in reinforcement learning (and in Rescorla-Wagner)

1. Temporal credit assignment
Critical problems, for control

2. Exploration/exploitation
TD learning - solution for temporal credit assignment

1. Estimate value of current state \((V_t = r_t + \gamma r_{t+1} + \ldots)\): (discounted) sum of expected rewards

2. Measure ‘truer’ value of current state: reward at present state + estimated value of next state \((r_t + \gamma V_{t+1})\)

3. TD error \(\delta_t = r_t + \gamma V_{t+1} - V_t\)

4. Use TD error to improve 1 \((V_t^{k+1} = V_t^k + \eta \delta_t)\)

where: \(V_t\) = value of the state reached at time \(t\) in iteration \(k\)
\(r_t\) = reward given at time \(t\); \(\eta\) = learning rate, \(\delta\) = prediction error
TD error: \( \delta_t = r_t + \gamma V_{t+1} - V_t \)
TD error: \( \delta_t = \gamma V_{t+1} - V_t + r_t \)
Basal ganglia - anatomy
Dopamine and acetylcholine meet in the striatum

Mouse

Monkey

19

Berlin 2004
Facts to remember (1)

- Basal ganglia receive cortical input
- Basal ganglia project to frontal cortex
- Dopamine and acetylcholine localization
The midbrain dopamine system

Schultz et al, 
JNS 13: 900-913 ,1993
... and it can cause long term plasticity of cortico-striatal synapses

Probabilistic instrumental conditioning task

\[ \delta_t = \gamma V_{t+1} - V_t + r_t \]
DA response
Dopamine population response to cue

- Start
- Cue
- Hold
- Go
- Delay
- Reward

Spike/S

-200 0 400

time (ms)

Spikes/s

0 5

n=114

Conditional reward probability

P=0.25

P=0.5

P=0.75

P=1
Dopamine population response-reward
Dopamine population response – reward omission
Instrumental conditioning - results

- Responses to visual cue are correlated with future reward probability.
- Responses to reward are inversely correlated with reward probability.
- Responses to reward omission are indifferent to reward probability.
- Dopamine neurons provide an accurate TD signal (but only in the positive domain).
But how is this related to behaviour?

- The law of effect is all about actions
- AI agents act to maximize a goal
- … and so do people and animals

- The basal ganglia are involved in action

*O’Doherty et al., Science 2004*
Control - Adding action

The agent has to:

- Learn to predict reinforcement
- Know the state-action-state transitions
Solution 1: actor/critic networks
How can the dopamine signal contribute to decision behaviour?

- Long term policy-shaping effect through synaptic plasticity

- Immediate effect on action

\[ P_{\text{action}} = \frac{1}{1 + e^{-m\delta(t)+b}} \]
The two armed bandit task
Decision behaviour, theory and practice

maximizing
probability-matching

monkeys?

\[
\frac{C_{\text{right}}}{C_{\text{right}} + C_{\text{left}}} \approx \frac{R_{\text{right}}}{R_{\text{right}} + R_{\text{left}}} \cdot \frac{\theta_{\text{left}}}{\theta_{\text{right}}}
\]

\( R = \text{reward} \)

\( C = \text{choice} \)

\( \theta \)
Monkeys’ decisions: probability matching
Lost in translation?

reward \rightarrow \text{behaviour}

\times

\text{dopamine response} \rightarrow \text{plasticity in action circuits}
Monkeys’ decisions: shaping by dopamine

![Graph showing relationship between dopamine levels and decision-making.](image)

- **Correlation Coefficient:** $R^2 = 0.930$

---

[Image 39x61 to 252x146]

[Image 90x163 to 553x715]