COMS30017
Computational Neuroscience

Week 7 / Video 4 / Temporal difference learning and dopamine

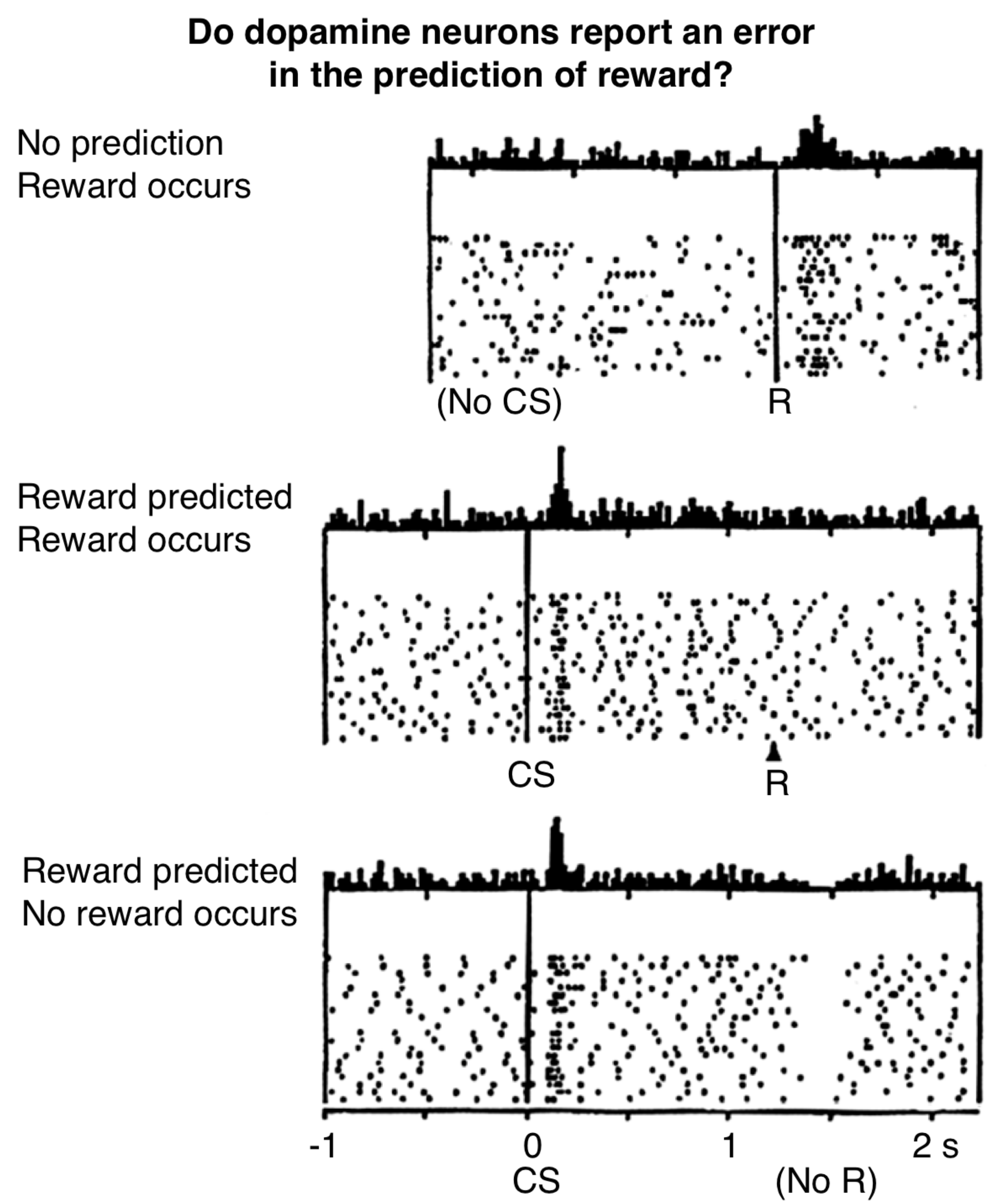
Dr. Laurence Aitchison
[*laurence.aitchison@bristol.ac.uk*](mailto:laurence.aitchison@bristol.ac.uk)

Intended Learning Outcomes

- Dopamine represents temporal reward prediction errors
- Modelling dopamine:
  - Define expected discounted value
  - Temporal difference learning

Dopamine neurons code temporal reward prediction error

- Initially, just reward presented. Dopamine neurons fire in response to reward.
- After many CS-reward pairings, dopamine neurons fire in response to CS (expected), not reward (expected).
- If reward is then omitted, a pause in firing



Schultz Dayan and Montague (1997)

Rewards through time

- In Rescorla Wagner, stimuli and rewards were not time-dependent (e.g. you couldn't have two rewards, delivered one after the other).
- Here, the animal gets different rewards at different times, $r(t)$. The animal's goal is to estimate the return, i.e. the sum of future reward,

$$R(t) = \sum_{\tau=0}^{T-t} r(t+\tau)$$

- where t is the current time and T is the length of the episode

Temporal difference predictor

- Use $V(t)$ to estimate the return,

$$V(t) = \sum_{\tau=0}^{t} w(\tau)x(t-\tau)$$

- backwards looking, so a stimulus τ ago always predicts the same amount of reward

Temporal difference learning

By definition of the return, $R(t) = \sum_{\tau=0}^{T-t} r(t+\tau)$

$$0 = (r(t) + R(t+1)) - R(t)$$

Replacing (not a formal derivation) the true returns, $R(t)$, with our estimates, $V(t)$, we obtain the "temporal-difference" (TD) error,

$$\delta(t) = \underbrace{\big(r(t) + V(t+1)\big)}_{\text{better estimate of } R(t)} - \underbrace{V(t)}_{\text{estimate of } R(t)}$$

This error can be used to update the weights,

$$\Delta w(\tau) = \eta \delta(t) x(t-\tau)$$

(proving why this works is hard – it isn't gradient descent)

Temporal difference learning: meaning and function

The TD error,

$$\delta(t) = \left(r(t) + V(t+1)\right) - V(t)$$

Represents "are things better or worse than I expected at the last timestep?"

Can be used to update probabilities of actions:

- If you take an action and expected reward increases ($0 < \delta(t)$), then maybe the action was good, and you should do it more.
- If you take an action and expected reward decreases ($\delta(t) < 0$), then maybe the action was bad, and you should do it less.

End