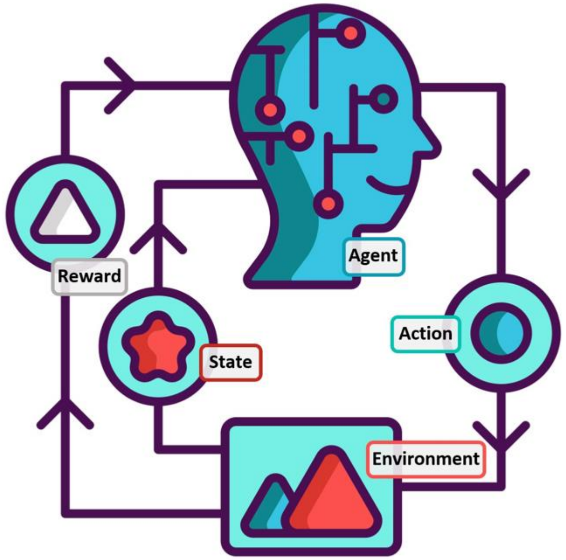


Reinforcement Learning in the Brain (Overview)

Peggy Seriès, IML
Informatics, University of Edinburgh, UK

pseries@inf.ed.ac.uk

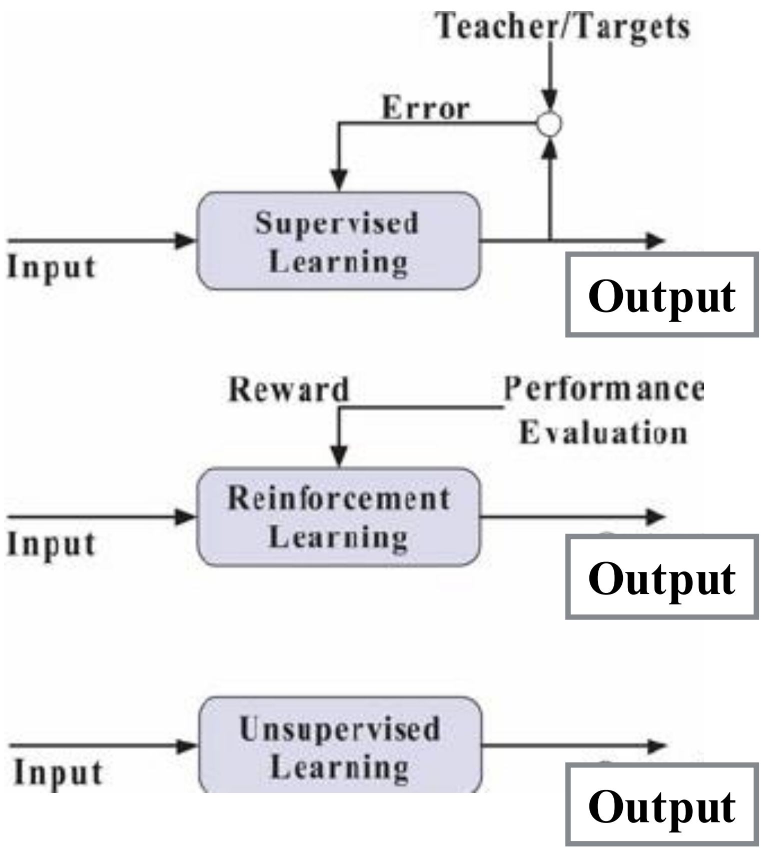
CCN Lecture 9

Reinforcement learning (RL):

- an area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

- Contrasted with supervised learning and unsupervised learning

- Thought to be a good model of how learning is occurring in the brain



Maximizing Reward as a guide to decision-making

- Key to (learning for) decision making at all levels

- Reinforcement learning : **maximize reward and minimize punishments**;

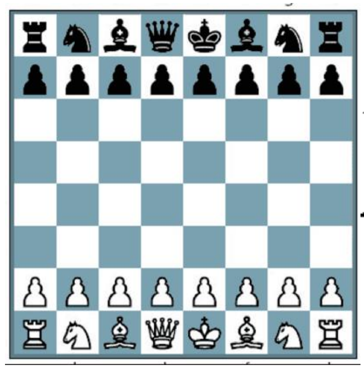
Sutton 1978; Sutton & Barto, 1990, 1998.

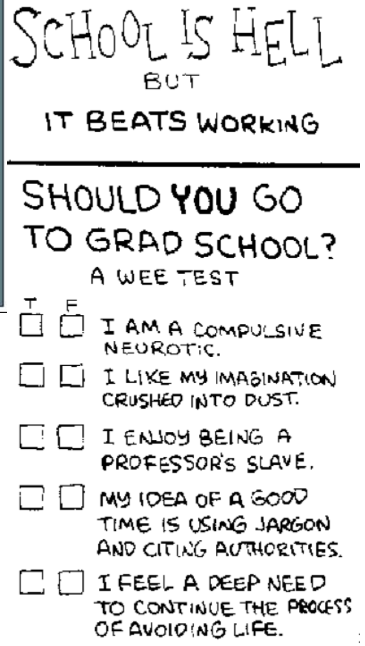
- Why is this hard?

(1) rewards/ punishment may be delayed;

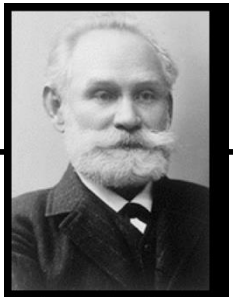
(2) outcome may depend on series of actions (credit assignment problem)

- Needs learning of predictions of events and actions





Animals learn predictions -- Pavlovian conditioning



Ivan Pavlov
(Nobel prize portrait)

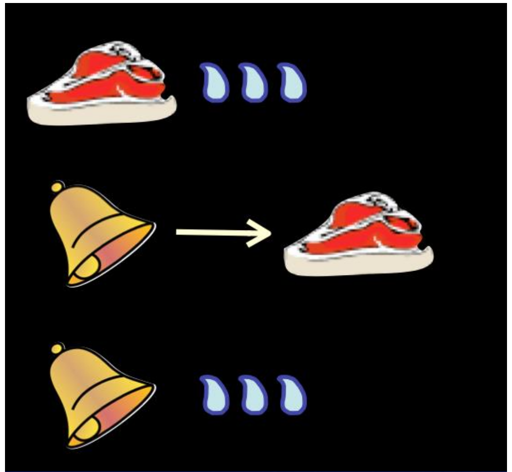
1849-1936

- Classical (aka "Pavlovian") conditioning: pairing of a conditioned stimulus (bell, CS) with an unconditioned stimulus (food, US)

- e.g Conditioned suppression, freezing to tone paired with punishment
http://www.youtube.com/watch?v=ZlZekx1P1g4

- Autoshaping, bird pecking on light that has been paired with food
http://www.youtube.com/watch?v=cacwAvgg8E

- Behaviorism: John Watson (1913) proposed that the process of classical conditioning (based on Pavlov's observations) was able to explain all aspects of human psychology.



Rescorla & Wagner Model of Classical Conditioning (1972)

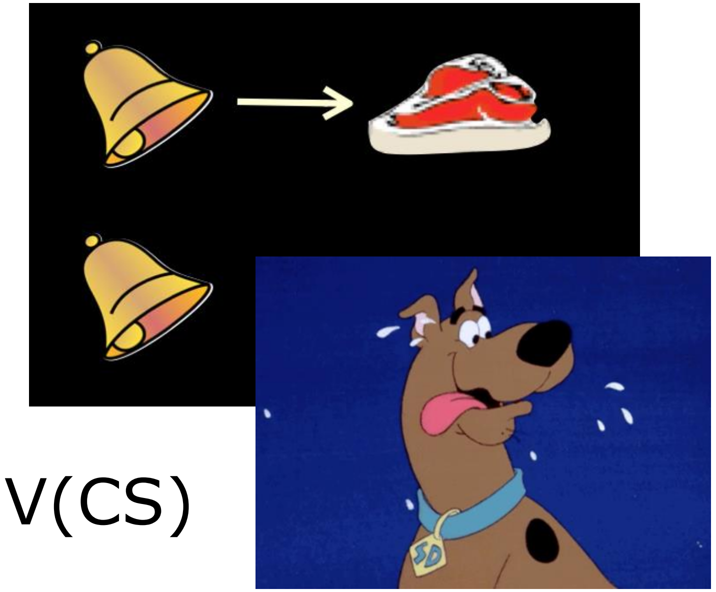
In 1972, Rescorla & Wagner proposed mathematical model to explain amount of learning that occurs on each trial of Pavlovian learning, when a signal (conditioned stimulus: CS) is paired with a subsequent stimulus (unconditioned stimulus: US).

Describes development of associative strength V(CS) between objects or events and reward or punishment, recognizing that:

1. Learning will occur if what happens on the trial does not match the expectation of the organism (surprise !),

2. The expectation on any given trial is based on the predictive value of all the stimuli present.

CS US



V(CS)

Rescorla & Wagner model of classical conditioning (1972)

- Change in value of associative strength V(CS) is proportional to the difference between actual outcome λ_{US} - and predicted outcome $\sum_i V_{old}(CS_i)$.

- The idea: error-driven learning: Learning occurs only when events violate expectations.

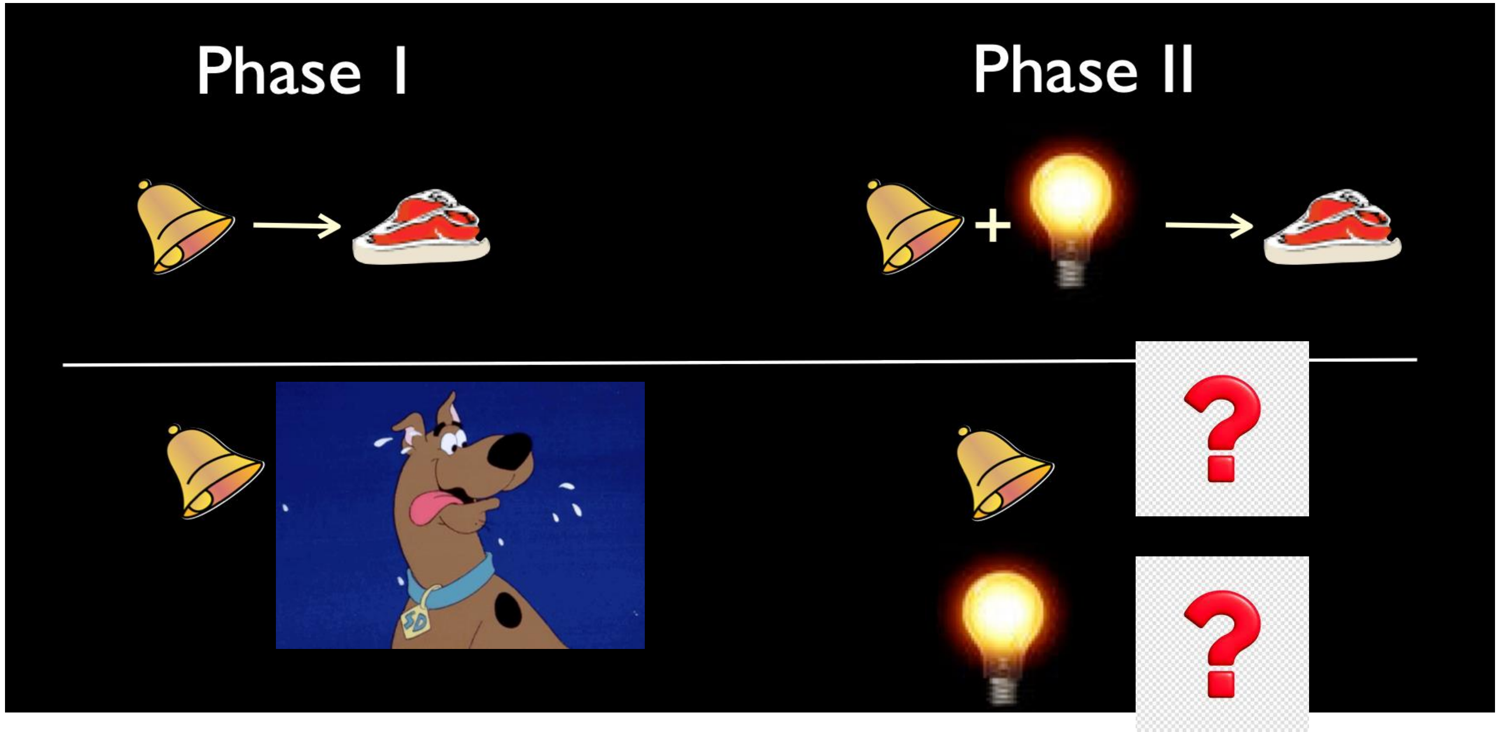
$$V_{new}(CS_i) = V_{old}(CS_i) + \underbrace{\eta}_{\text{learning rate}} \underbrace{\left[\overbrace{\lambda_{US}}^{\text{actual reward}} - \overbrace{\sum_i V_{old}(CS_i)}^{\text{prediction}} \right]}_{\textbf{reward prediction error}} .$$

- Most influential model of animal learning, explains puzzling behavioral phenomena such as blocking, overshadowing and conditioned inhibition.

How do we know that animals use an error-correcting rule ?

Leon Kamin
(1917-2017)

- (Kamin) Blocking: Adding a second stimulus

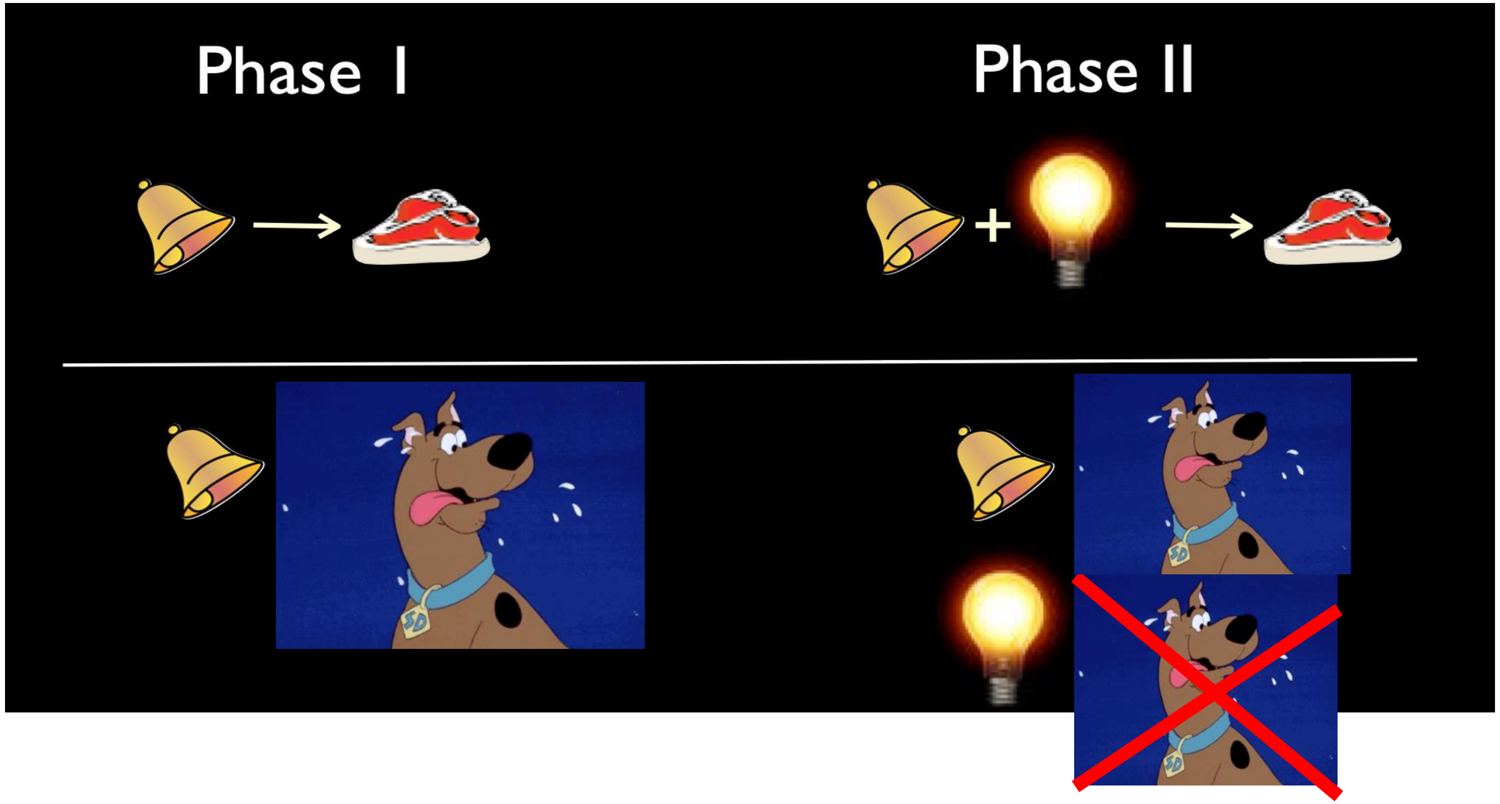


Will the dog learn to salivate to the light?

How do we know that animals use an error-correcting rule ?

Leon Kamin
(1917-2017)

- (Kamin) Blocking: Why does the light not make the animal salivate?



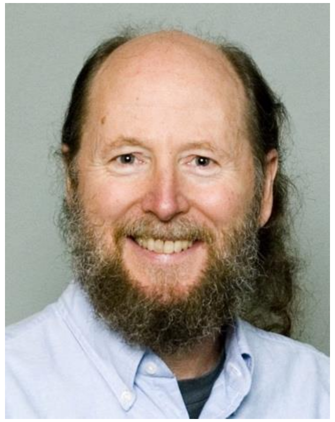
- <u>Interpretation:</u> the bell fully predicts the food, and the presence of the light adds no new predictive information -- therefore no association develops to the light.

Limitations of the Rescorla & Wagner Model

- Does not extend to 2d order conditioning, i.e. A->B->reward; where A gains reward predictive value

- Basic unit of learning = conditioning trial as discrete temporal object

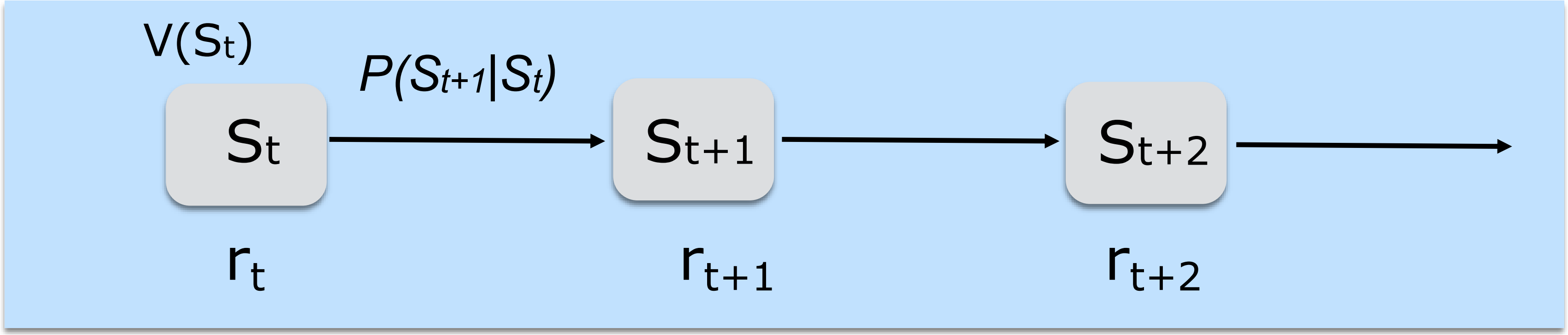
This fails to account for the temporal relations between CS and US stimuli within a trial



Richard Sutton

→ **Temporal Difference (TD) learning**, first described by Sutton (1988)
- a means to overcome these limitations
- extension of Rescorla-Wagner to take timing of events into account and solve the credit assignment problem.

Temporal Difference (TD) learning (1)



- Consider a succession of states S, following each other with $P(S_{t+1}|S_t)$
- Rewards observed in each state with probability $P(r|S_t)$

(This is a *Markov Decision Process*)

- Useful quantity to predict is the expected sum of all future rewards, given current state S_t,

= value of state S, $V(S_t)$

$$V(S_t) = E\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \middle| S_t\right] = E\left[\sum_{i=t}^{\infty} \gamma^{i-t} r_i \middle| S_t\right]$$

where E denotes expected value (or mean) and gamma the discount factor

Temporal Difference (TD) learning (1)

$$V(S_t) = E\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots \middle| S_t\right] = E\left[\sum_{i=t}^{\infty} \gamma^{i-t} r_i \middle| S_t\right]$$

- Discount factor introduced to make sure that the sum is finite, but also humans and animals prefer earlier rewards to later ones

- Incorporating probabilities $P(S_{t+1}|S_t)$ and $P(r|S_t)$, we get recursive form

$$\begin{aligned} V(S_t) &= E\left[r_t|S_t\right] + \gamma E\left[r_{t+1}|S_t\right] + \gamma^2 E\left[r_{t+2}|S_t\right] + \ldots = \\ &= E\left[r_t|S_t\right] + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t)\left(E\left[r_{t+1}|S_{t+1}\right] + \gamma E\left[r_{t+2}|S_{t+1}\right] + \ldots\right) \\ &= P(r|S_t) + \gamma \sum P(S_{t+1}|S_t) V(S_{t+1}) \end{aligned}$$

- Goal of TD learning = learn the values $V(S_t)$.

Temporal Difference (TD) learning (2)

- When estimated values are incorrect, there is a discrepancy between 2 sides of equation:

prediction error:

$$\delta_t = P(r|S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t) V(S_{t+1}) - V(S_t).$$

- Assumes knowledge of $P(S_{t+1}|S_t)$ and $P(r|S_t)$, which we usually don't know.
- Model-free Approximation which can be formally justified (sampling):

$$\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t)$$

- prediction error is a natural signal for improving estimates $V(S_t)$, giving:

$$V(S_t)_{new} = V(S_t)_{old} + \eta \cdot \delta_t,$$

~ current reward + next prediction - current prediction

Temporal Difference (TD) learning (3)

- Resulting learning rule: $V(S_t)_{new} = V(S_t)_{old} + \eta \cdot \delta_t,$

$$\boxed{V_{new}(S_t) = V_{old}(S_t) + \eta(r_t + \gamma V_{old}(S_{t+1}) - V_{old}(S_t)).}$$

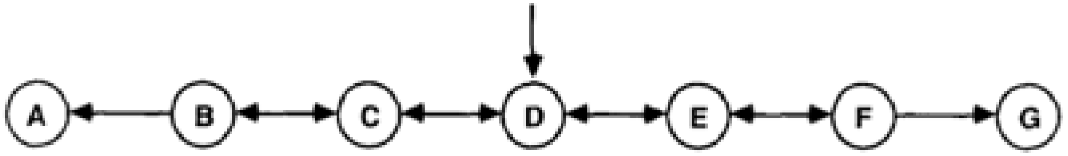
current reward+next prediction - current prediction

- This is TD(0) learning rule as proposed by Sutton & Barton (1990).

- reduces to Rescorla-Wagner model if only one step i.e. $V(S_{t+1})=0$.

$$\boxed{V_{new}(S_t) = V_{old}(S_t) + \eta(r_t - V_{old}(S_t)).}$$

TD in practice



R=0

R=1

e.g. π= random walk, at each state go left or right with 50% chance

Input: the policy π to be evaluated
Initialize $V(s)$ arbitrarily (e.g., $V(s) = 0, \forall s \in \mathcal{S}^+$)
Repeat (for each episode):
    Initialize S
    Repeat (for each step of episode):
        $A \leftarrow$ action given by π for S
        Take action A; observe reward, R, and next state, S'
        $V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$
        $S \leftarrow S'$
    until S is terminal

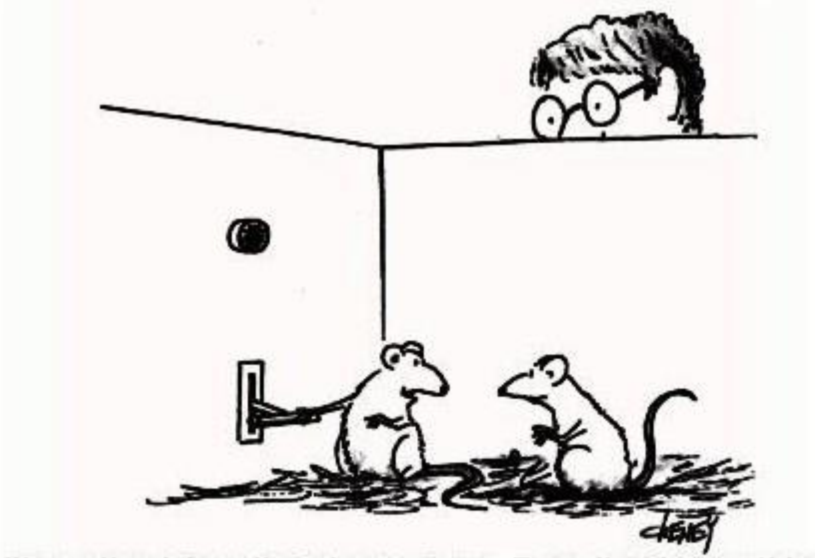
Figure 6.1: Tabular TD(0) for estimating v_π.

Sutton & Barton (1990).

Instrumental conditioning: adding control

- Animals not only learn associations between stimuli and reward but also between actions and reward
- Learning to select actions that will increase the probability of rewarding events and decrease the probability of aversive events.
- Rat lever pressing in boxes -- operant conditioning (Skinner)

Skinner
1904-1990



It's a rather interesting phenomenon. Every time I press this lever, that post-graduate student breathes a sigh of relief.

Actor/Critic Methods

- How can such action selection be learned?

- Barto (1983): credit assignment problem can be solved by a learning system comprised of 2 neurons-like elements:

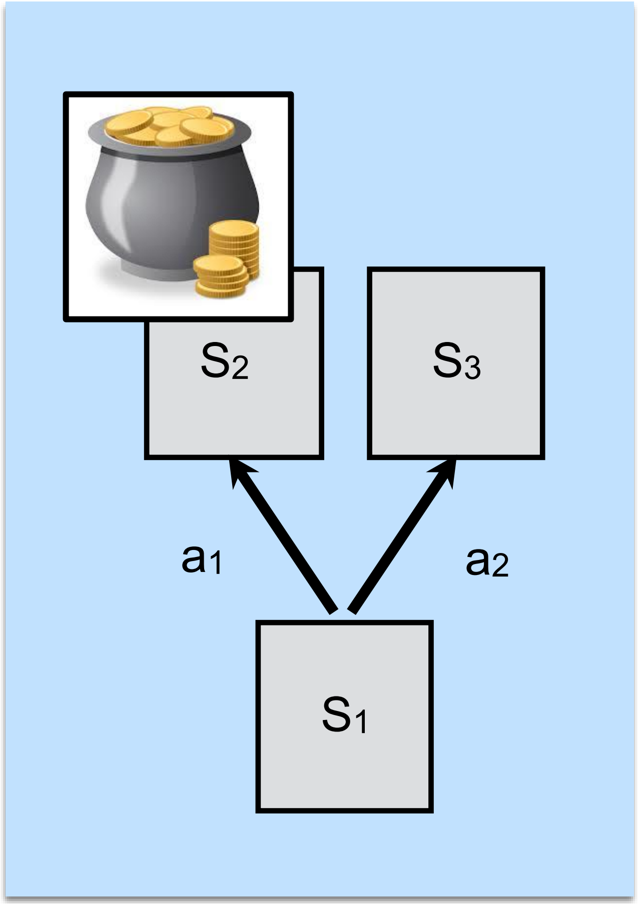
- the critic, uses TD learning to construct values of states

- the actor, learn to select actions at each state using prediction error.

<u>Idea</u>: if positive prediction error is encountered, current action should be repeated.

Learning of policies $\pi(S,a) = p(a|S)$

$$\boxed{\pi(S,a)_{new} = \pi(S,a)_{old} + \eta_{\pi}\delta_t}$$



Q-learning

- Watkins (1989)
- Alternative: explicitly learn the predictive value (future expected rewards) of taking an action at each state = learn the value of state-action pairs Q(S,a)
- Learning rule:

$$\boxed{Q(S_t, a_t)_{new} = Q(S_t, a_t)_{old} + \eta\delta_t}$$

- Q prediction error:

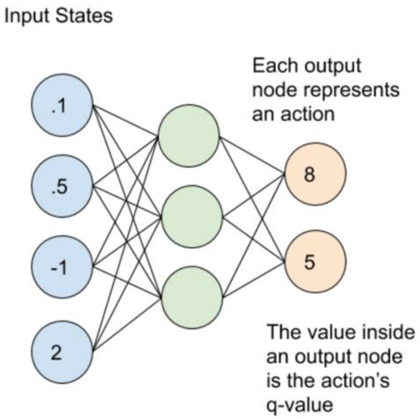
$$\boxed{\delta_t = r_t + \max_a \gamma Q(S_{t+1}, a) - Q(S_t, a_t)}$$

~ current reward+ prediction of next best action- current prediction

- SARSA algorithm a slightly different version, uses next action actually taken, instead of best action

Machine learning applications of Q learning (deep Q learning)

Google DeepMind



LETTER

doi:10.1038/nature14236

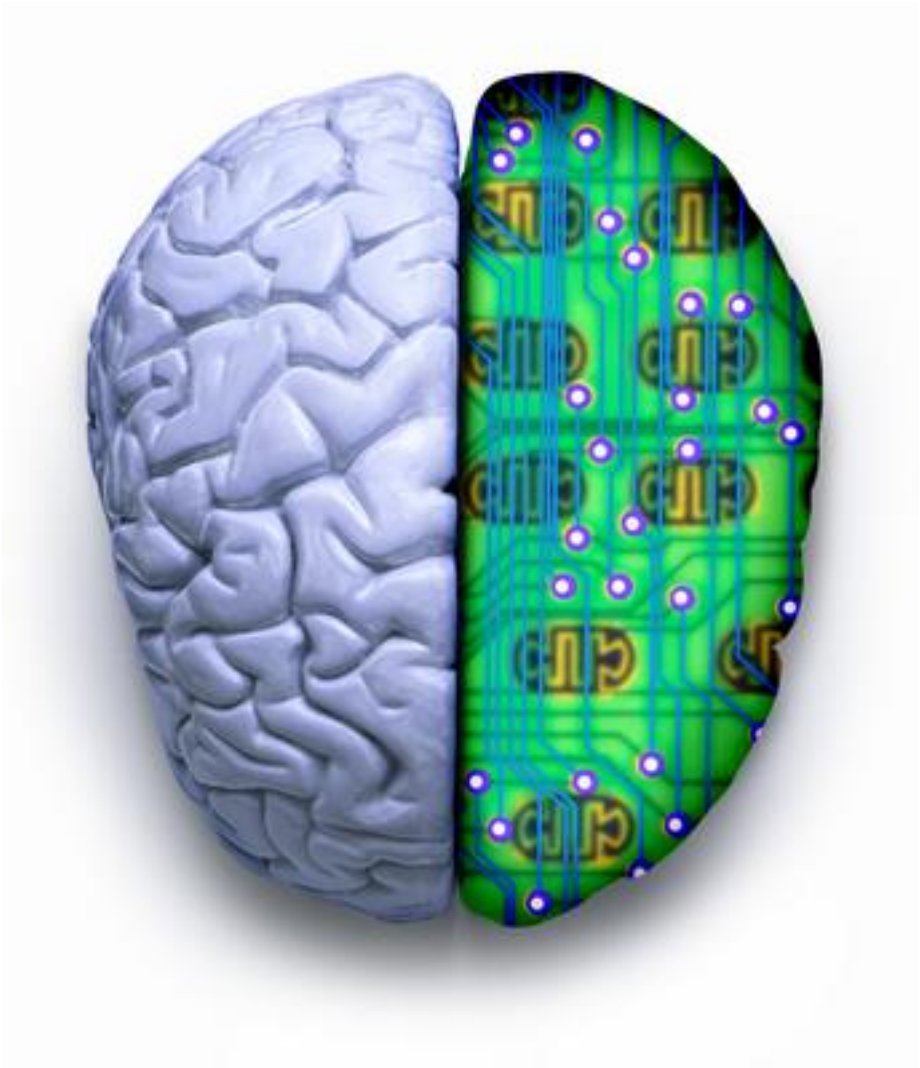
Human-level control through deep reinforcement learning

Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

2015 - now cited >15,000 times https://www.youtube.com/watch?v=V1eYniJ0Rnk

A recent application of Q-learning to deep learning, by Google DeepMind has been successful at playing some Atari 2600 games at expert human levels. Applications in Robotics & autonomous navigation, Financial decision-making, healthcare, traffic management..

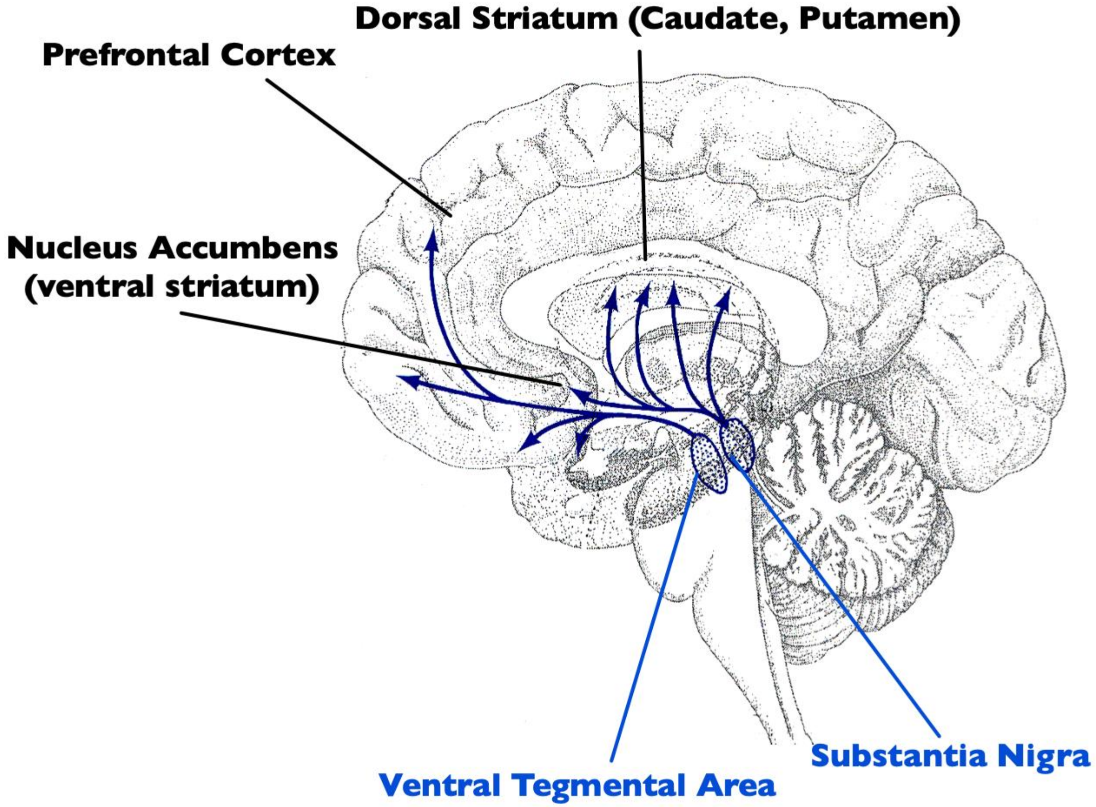
https://medium.freecodecamp.org/an-introduction-to-deep-q-learning-lets-play-doom-54d02d8017d8

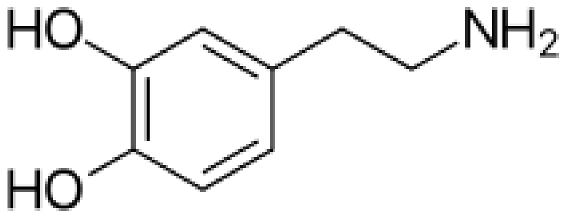


Does the brain do anything like that ?

- “the largest success of computational neuroscience”, dopamine and prediction error

What is Dopamine ?

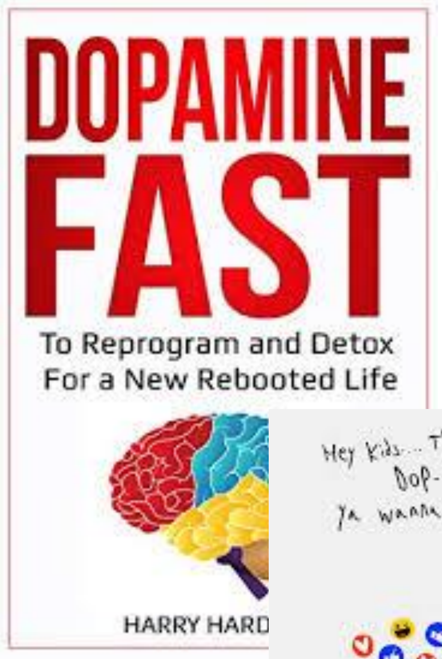


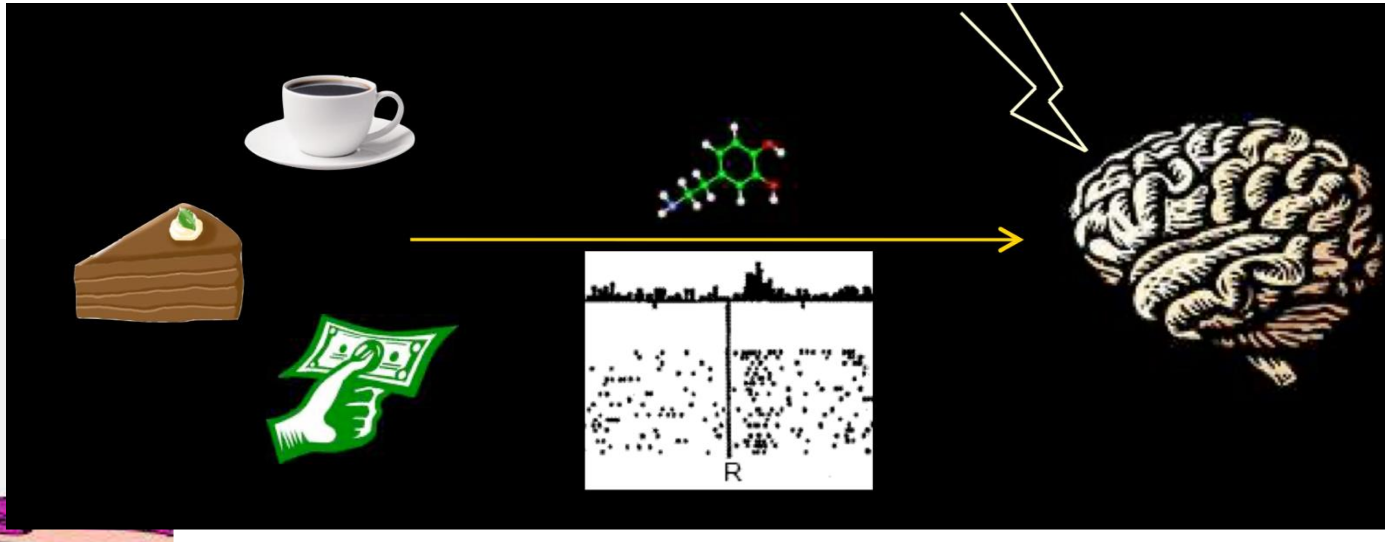


- A neurotransmitter
- Dopaminergic neurons in Ventral Tegmental Area (VTA) and Substantia Nigra (SN), both in the midbrain
- Parkinson's Disease : motor control/ initiation
- Addiction, gambling, natural rewards
- also involved in working memory, novel situations, ADHD, schizophrenia, Tourette.

Former idea: Dopamine signals Reward (Wise, ‘80s)

- Initial idea: dopamine represent reward signals
- brain self stimulation by rats http://www.youtube.com/watch?v=7HbAFYiejvo
- antipsychotic drugs (dopamine antagonists) cause anhedonia
- ‘wanting’ more than ‘liking’
- dopamine important for reward-mediated conditioning





New idea: Phasic Dopamine signals Prediction Error

Schultz et al 90s

- Monkeys underwent simple instrumental or Pavlovian conditioning
- Disappearance of dopaminergic response at reward delivery after learning
- If reward is not presented, response depression below basal firing at expected time of reward.

Dopamine Response

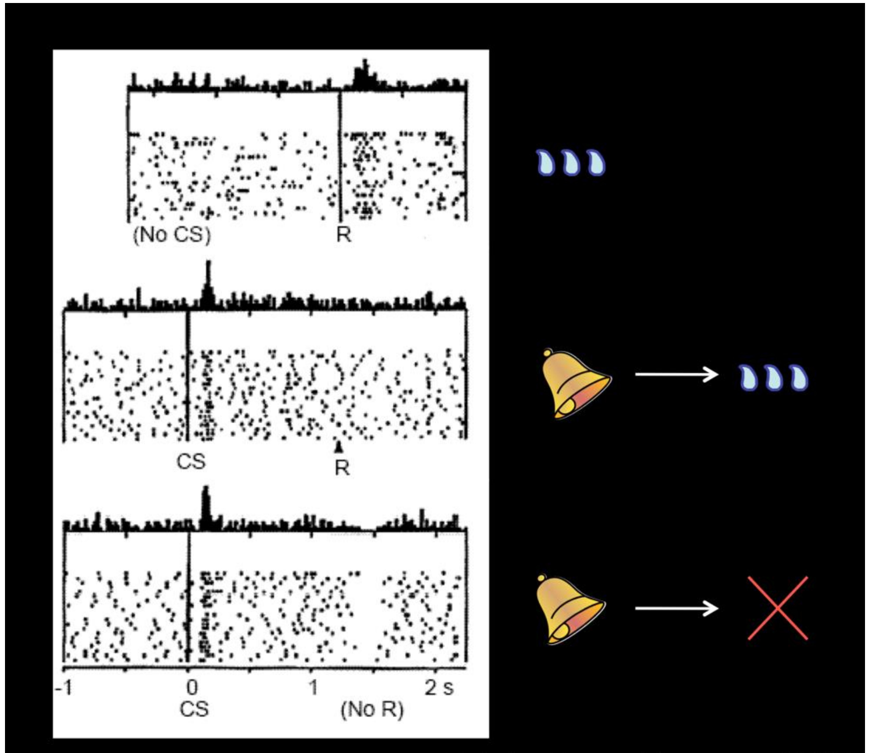
= Reward Occurred – Reward Predicted

= prediction error

A Neural Substrate of Prediction and Reward

Wolfram Schultz, Peter Dayan, P. Read Montague*

The capacity to predict future events permits a creature to detect, model, and manipulate the causal structure of its interactions with its environment. Behavioral experiments suggest that learning is driven by changes in the expectations about future salient events such as rewards and punishments. Physiological work has recently complemented these studies by identifying dopaminergic neurons in the primate whose fluctuating output apparently signals changes or errors in the predictions of future salient and rewarding events. Taken together, these findings can be understood through quantitative theories of adaptive optimizing control.



Schultz, Dayan, Montague, 1997

THE BRAIN PRIZE 2017

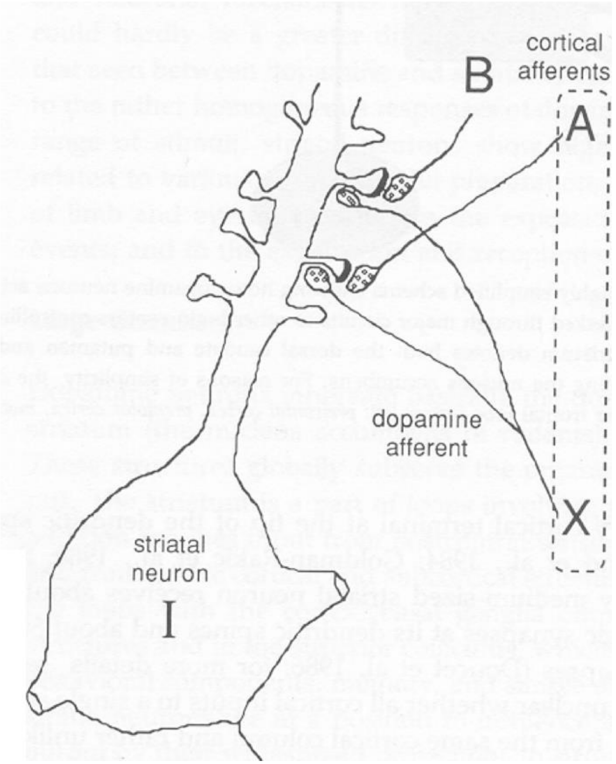
https://speakingofresearch.com/2017/03/06/winners-of-2017-brain-prize-announced-peter-dayan-ray-dolan-and-wolfram-schultz/

Dopamine and Prediction

• The idea: dopamine encodes prediction error (Montague, Dayan, Barto, 1996)
Teaching signal, crucial for learning

• Provided normative basis for understanding not only when dopamine neurons fire when they do, but also why, and what the function of these firing might be.

• Evidence for dopamine-dependent, or dopamine-gated plasticity in synapses between cortex and striatum.



Using fMRI to visualise prediction errors in humans

- Model-driven analysis -- search the brain for predicted hidden variables that should control learning:

- 1) collect behavioural data in fMRI scanner
- 2) fit a model, e.g. TD or Rescorla Wagner, to subjects'performance;
- 3) Once best-fitting model parameters have been found, then the different model components (time series, e.g. values and prediction error) can be regressed against the fMRI data.

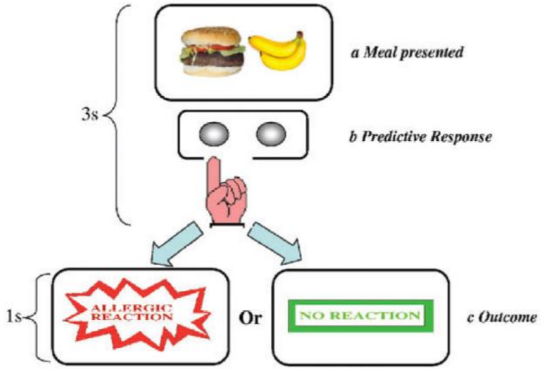
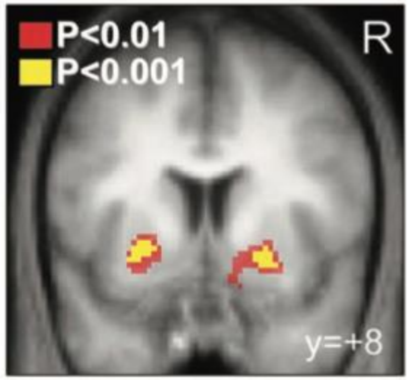
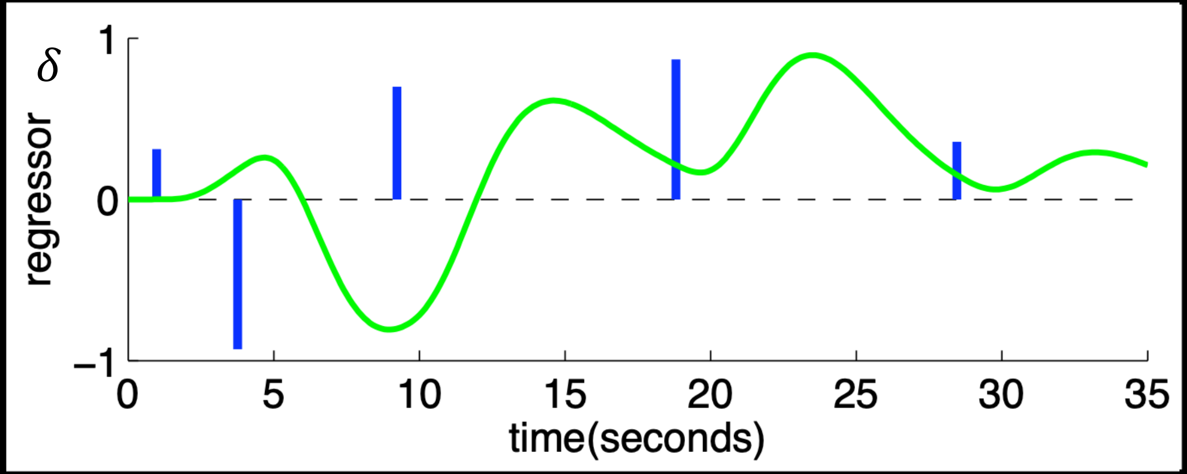


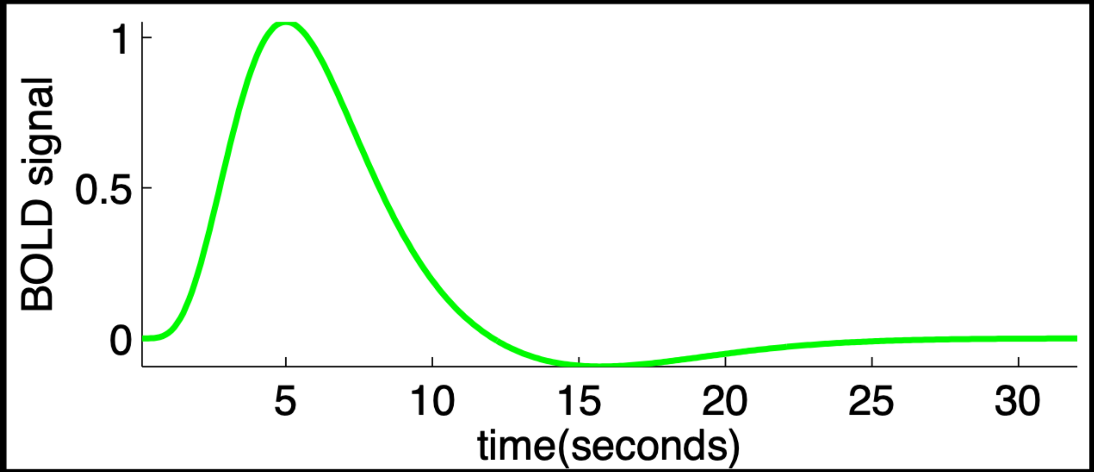
Fig. 1. Trial structure.
On each trial, subjects were presented with a meal that their patient had eaten, and then they made a predictive response. Finally they were informed of the effect of that meal on their patient.

$$V_{new}(CS_i) = V_{old}(CS_i) + \eta\left[\lambda_{US} - \sum_i V_{old}(CS_i)\right].$$



short aside: functional magnetic resonance imaging (fMRI)





Using fMRI to visualise prediction errors

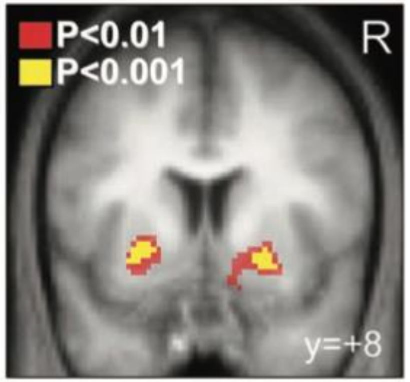
• Prediction errors signals found

in nucleus accumbens (part of striatum) and orbito-frontal cortex, both major dopaminergic targets.

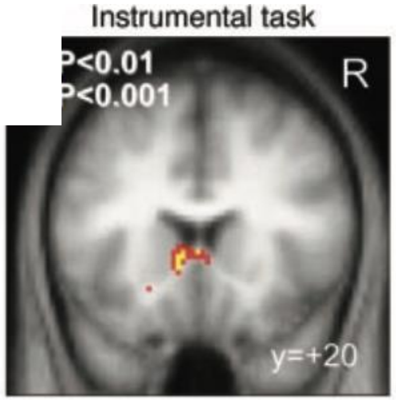
• O'Doherty et al (2004): fMRI correlates of prediction error signals can be dissociated in dorsal and ventral striatum, according to whether instrumental vs Pavlovian conditioning,

-- supporting an Actor/Critic architecture.

ventral striatum activity found in both Pavlovian and instrumental task



dorsal striatum activity found only in instrumental task



New Promising Applications to Psychiatry

- Model-based fMRI opens the door to investigating decision-making and reward signals differences in mental illness, e.g.

doi:10.1093/brain/awm173 Brain (2007), 130, 2387–2400

Disrupted prediction-error signal in psychosis: evidence for an associative account of delusions

P. R. Corlett,[1] G. K. Murray,[1,2] G. D. Honey,[1] M. R. F. Aitken,[3] D. R. Shanks,[4] T. W. Robbins,[3] E. T. Bullmore,[1,2] A. Dickinson[3] and P. C. Fletcher[1]

- Frontal cortex responses in patient group suggestive of disrupted prediction-error processing (i.e. no longer cleanly distinguishes unexpected from expected outcomes).
- Across subjects, extent of disruption was significantly related to an individual’s propensity to delusion formation.
- Delusions as a consequence of abnormal learning.

Summary

- Optimal learning depends on prediction and control

- The problem: prediction of future reward (or punishment)

- The algorithm: TD learning (or variants)

  Update values so as to minimise prediction error.

- Neural implementation: phasic dopamine as prediction error signal.

  dopamine-dependent learning in cortico-striatal synapses in basal ganglia

  (but note that this simple model is now thought to be more nuanced).

- RL has revolutionised how we think of learning in the brain.

  Implications for the understanding of disorders, such as Parkinson’s and schizophrenia, as well as addiction, depression and more..