University of BRISTOL

Reinforcement learning

Seán Froudist-Walsh

Lecturer in Computational Neuroscience

1

1

- ”The greatest teacher, failure is”
- - Yoda

2

2

Intended Learning Outcomes

By the end of this lecture, you will understand

- The goal of reinforcement learning (RL)
- The fundamental elements of reinforcement learning
- Algorithms for learning the values of states and actions
- The basics of reinforcement learning in the brain.
- Brain-inspired advances in deep reinforcement learning
- How the brain builds cognitive maps & its connection to model-based RL
- How reinforcement learning is beginning to tackle met—cognition (learning-to-learn) and curiosity

3

3

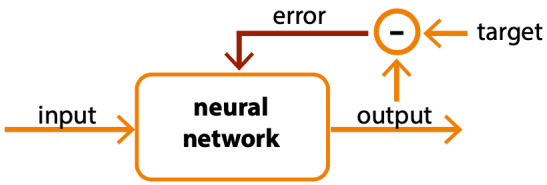
The goal of reinforcement learning

- is to maximise the total amount of (discounted) reward in the future

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \ldots = \sum_{k=0}^{T} \gamma^k R_{t+k+1}$$

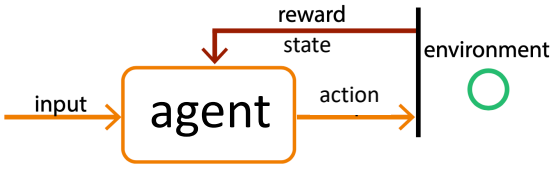
The value of future rewards is discounted (often γ = 0.9) to reflect a preference for more immediate rewards.

Supervised learning



The correct output is given by a mysterious teacher

Reinforcement learning



Rewards (including negative rewards = punishments), and information about the current state are given by the environment

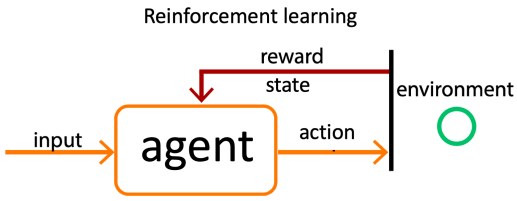
4

images adapted from Rui Ponte Costa

4

The elements of a reinforcement learning system (parts 1-3/5)

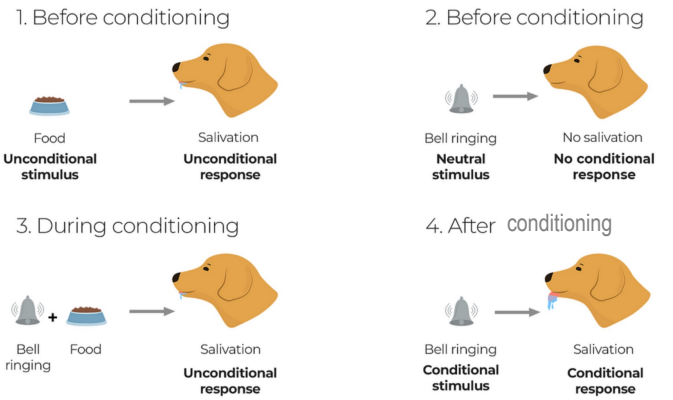
1. **an agent** - the thing that performs the task, e.g. a robot, a neural network etc.)
2. **the environment** - the external system or setting with which the learning agent interacts. It encompasses everything that the agent observes and responds to, but which is outside the agent's control.
3. **a reward signal –** a single number (which can be zero, positive or negative) sent from the environment to the agent on every timestep. The agent's only goal is to maximize the total reward it receives over the long run



5

5

Pavlovian (classical) conditioning - Learning to predict stimuli & rewards



conditioning
=
learning

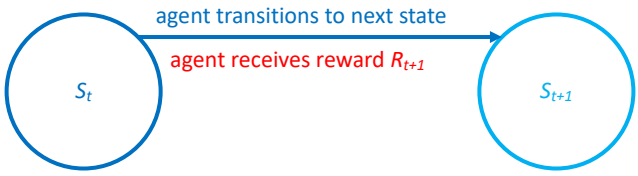
Pavlov, 1897

6

6

The elements of a reinforcement learning system (part 4/5)

1. **an agent**
2. **the environment**
3. **a reward signal**
4. **a value function** – specifies what is good in the long run.



state-value function

S_t = s $\quad v(s) = \mathbb{E}[\, R_{t+1} + \gamma\ R_{t+2} + \gamma^2\ R_{t+3} \ldots + \gamma^2\ R_{t+3} \ldots \mid S_t = s] = \mathbb{E}[\sum_{k=0}^{T} \gamma^k R_{t+k+1} \mid S_t = s]\,]$

The value of a state is the total amount of (discounted) reward
an agent can expect to accumulate over the future, starting from that state.

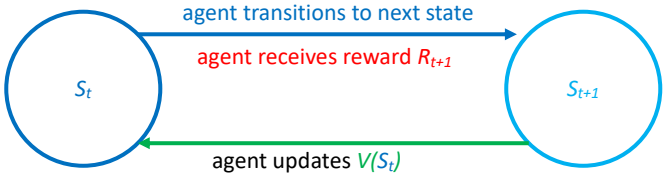
$V(S_t) = \mathbb{E}[R_{t+1}] + \gamma\,[V(S_{t+1})]$

The estimated value of the current state is just the discounted estimated value of the next state
plus the expected reward received during the transition to the next state.

Sutton & Barto, 2018

7

The Temporal Difference (TD) learning model of classical conditioning



$V(S_t) \leftarrow V(S_t) + \alpha \quad [\, R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\,]$

The temporal difference error (in neuroscience called the reward prediction error) captures how much higher/lower a reward is than expected and/or if the following state's value is higher/lower than expected.

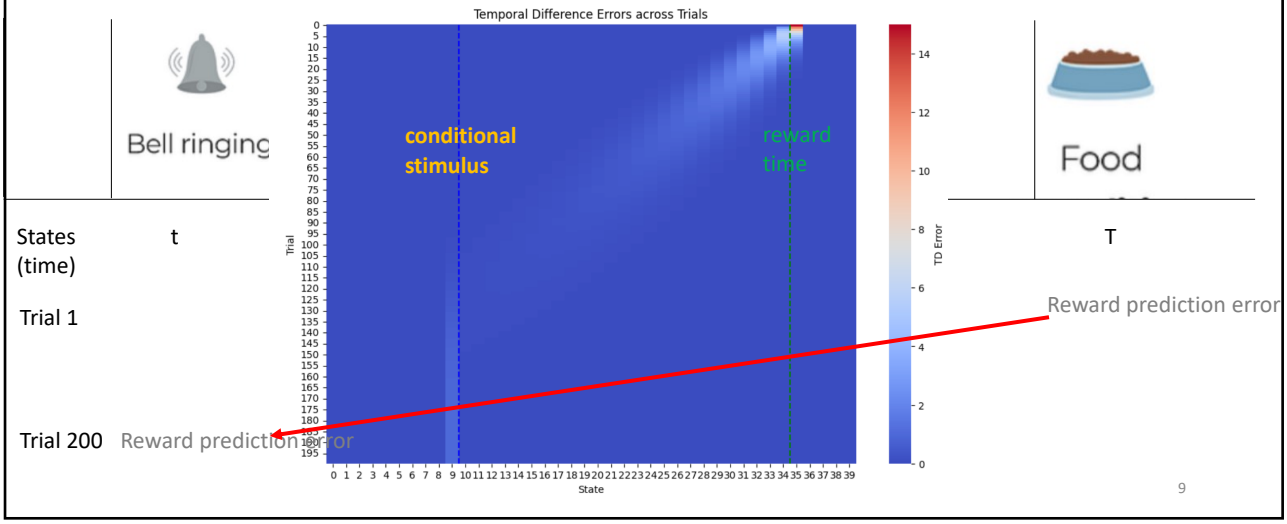
The TD model updates the estimated value of the previous state based on surprising rewards and the value of the following state.

The estimated value of the previous state is updated using the reward prediction error (RPE, aka TD error).
The size of the update is determined by the learning rate.

Sutton & Barto, 2018

8

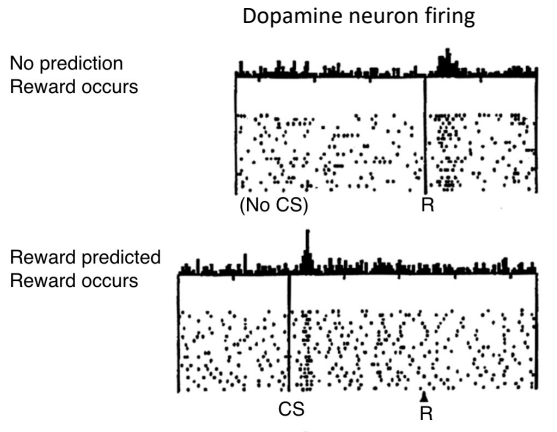
The Reward Prediction Error shifts backwards in time from the reward to the stimulus that predicts it.

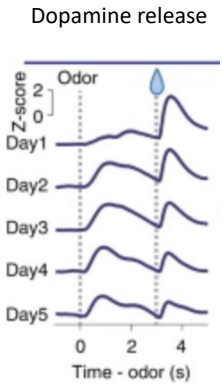


9

9

Dopamine strongly resembles the reward prediction error





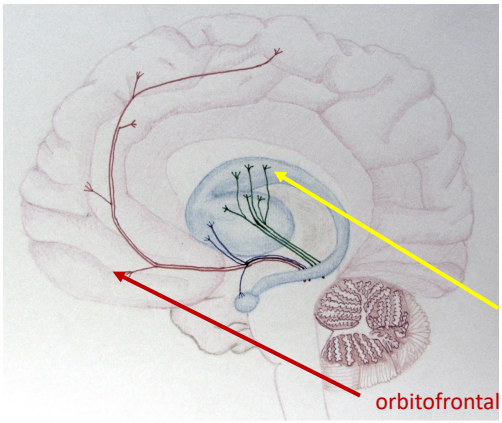
Schultz et al., *Science*, 1997

Amo et al., *Nature Neuroscience*, 2022

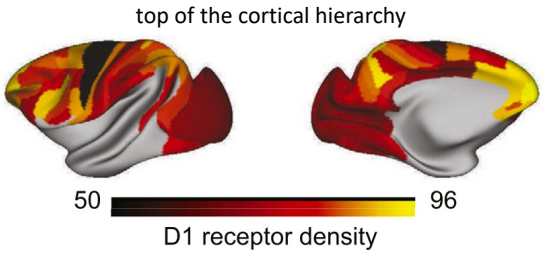
10

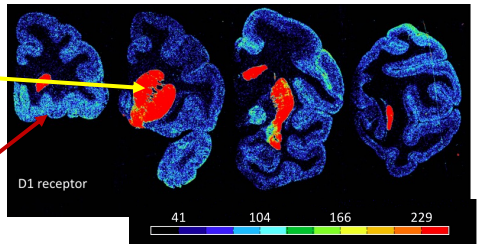
10

Where does dopamine target in the brain?



Froudist-Walsh, PhD Thesis, 2015
Artist: Sofía Minguillón Hernández





Brain activity in dopamine-responsive regions resembles the reward-prediction error
O'Doherty et al., *Neuron*, 2003

Froudist-Walsh et al., *Neuron*, 2021

11

11

Blackboard quiz

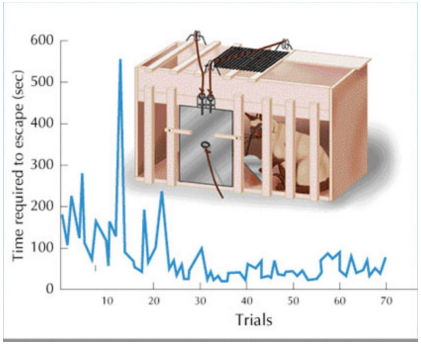
Using the TD-learning, given the following reward sequence *R* and state space sequence *S* compute the value function for *V*(*St*=1), *V*(*St*=2) and *V*(*St*=3), for the *first*, *second* and *third* update steps, corresponding to three trials of the same sequences of states and rewards. Assume that the value function starts at 0 for all states, learning rate = 1 and future discounting (λ)=0.5. *S* = {*St*=1, *St*=2, *St*=3} *R* = {*Rt*=1 = 0,*Rt*=2 = 0,*Rt*=3 = 0,*Rt*=4 = 0.5}

12

12

Instrumental conditioning - Learning which actions to take

- Thorndike's law of effect
- "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;
- those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur.
- The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond."
- Learning from trial and error.



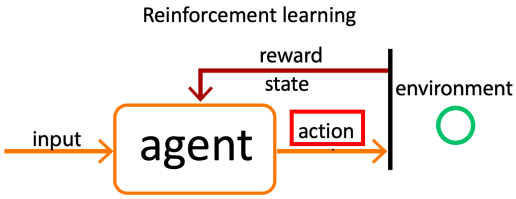
Thorndike, 1898, 1911

13

13

The elements of a reinforcement learning system (part 5/5)

1. **an agent**
2. **the environment**
3. **a reward signal**
4. **a value function**
5. **a policy** is a mapping from states to probabilities of selecting each possible action



$$\pi(a|s) = p(A_t = a|S_t = s)$$

The policy gives the probability of taking any action in any state

In any system in which the agent takes actions, the value functions are defined with respect to the policy being followed.

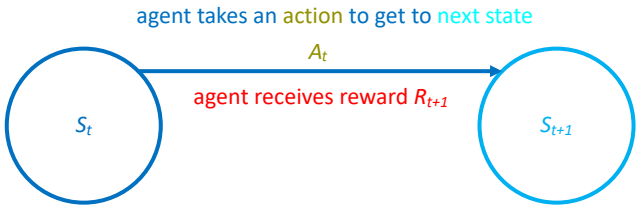
$$v_\pi(s) = \mathbb{E}_\pi[\sum_{k=0}^{T} \gamma^k R_{t+k+1} | S_t = s]]$$

Sutton & Barto, 2018

14

14

Action-value function



action-value function for policy π

$$q_\pi(s,a) = \mathbb{E}_\pi[\sum_{k=0}^{T} \gamma^k R_{t+k+1} | S_t = s, A_t = a]$$

The value of a taking an action a in a state s under the policy π is the total amount of (discounted) reward an agent can expect to accumulate over the future, starting from taking action a in state s and following the policy π afterwards.

Sutton & Barto, 2018

15

ε-greedy action selection

- At each state,
  - With probability 1- ε select the action a with the highest value Q(s,a) – (exploit)
  - With probability ε select randomly among all actions (explore)
- ε-greedy combined with
  - an estimate Q(s,a) of the action-value function
  - defines a policy.

16

16

Q-learning

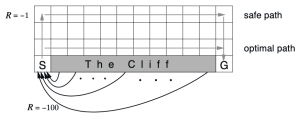
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right]$$

The estimated value of an action taken in the previous state is updated using the reward prediction error (RPE).

The size of the update is determined by the learning rate.

To calculate the RPE, we need the old estimate of the value of the state-action pair (before seeing the reward)

And we have to subtract this from the reward received plus the value of the following state-action pair.



But should we use the value of the next action chosen? — on-policy methods e.g. SARSA

Or the value of the best available action? — off-policy methods e.g. Q-learning

What if the next action is exploratory/random?

Q-values (value of any action in any state) are stored in a look-up table

	A1	A2	A3	A4
S1	+1	+2	-1	0
S2	+2	0	+1	-2
S3	-1	+1	0	-2
S4	-2	0	+1	+1

Watkins, 1989
Sutton & Barto, 2018

17

17

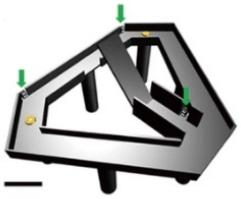
Blackboard Quiz

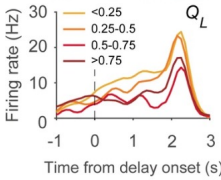
- What makes Q learning different from TD learning?

18

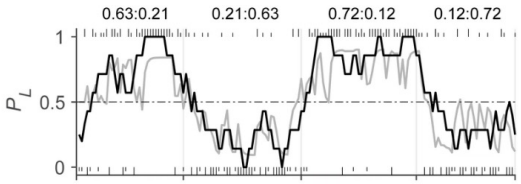
18

Action values in the brain



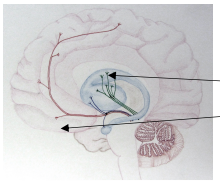


Activity in a neuron in the striatum correlates with the estimated value of choosing left (Q_L)



rat's choice (probability of choosing left averaged over 10 trials)

Q-learning prediction (probability of choosing left)



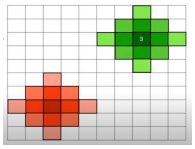
Action-value signals found in neurons in the striatum, orbitofrontal cortex and hippocampus (not shown)

Shin et al., *eLife*, 2021

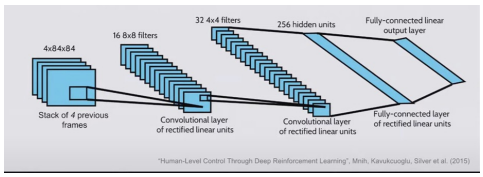
What if the state—space is so large that you can't explore it all? – Deep Q learning

- In real life, the sensory input at any moment is slightly different from all other moments. Is each moment a state?
- Having values for each state & action in a look-up table is infeasible
- Solution?
- Use deep network as a Q(S,A; ϑ) function approximator when in the presence of high-dimensional state space
- Make reinforcement learning work a big more like supervised learning

	A1	A2	A3	A4
S1	+1	+2	-1	0
S2	+2	0	+1	-2
S3	-1	+1	0	-2
S4	-2	0	+1	+1



Each state is a sequence of frames from a video game



$$L_i(\theta_i) = \mathbb{E}_{(s,a,r,s')\sim U(D)}\left[\left(\boxed{r+\gamma \max_{a'} Q(s',a';\theta_i^-)} - Q(s,a;\theta_i)\right)^2\right]$$

target

Output the estimated value for each possible action from that state Q(s,a)

Lex Fridman, Volodymyr Mnih, Serrano Academy

Mnih et al., *Nature*, 2015

Replay memory – a brain-inspired trick to make deep Q learning work

The problem:

1. Consecutive states are highly correlated.
    1. Learning 'on-line' would therefore be inefficient.
    2. Can get stuck in local minima

The solution:

1. Replay past experiences
    1. Store experiences (actions, state transitions & rewards) in a memory bank
    2. Replay these experiences randomly, and use this for training

21

21

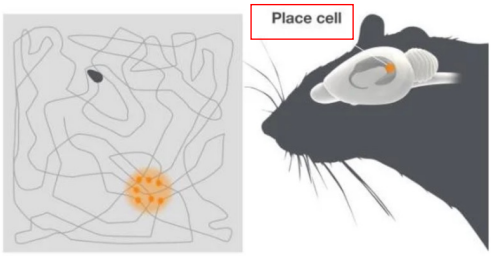
Blackboard Quiz

Briefly explain what is the problem in standard reinforcement learning methods that deep reinforcement learning methods address. How does it address it?

22

22

Hippocampal replay & memory consolidation

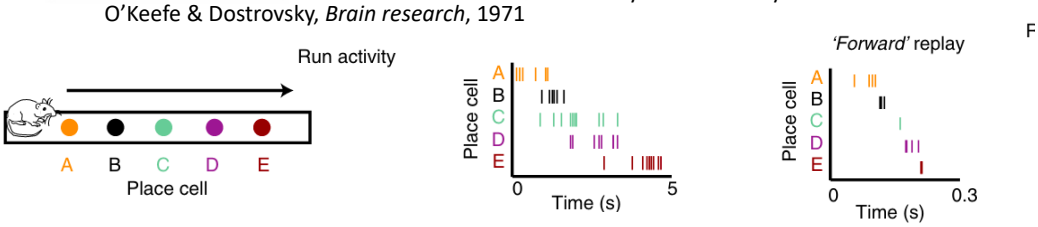


O'Keefe & Dostrovsky, *Brain research*, 1971

The hippocampus contains place cells – cells that fire preferentially in particular locations.

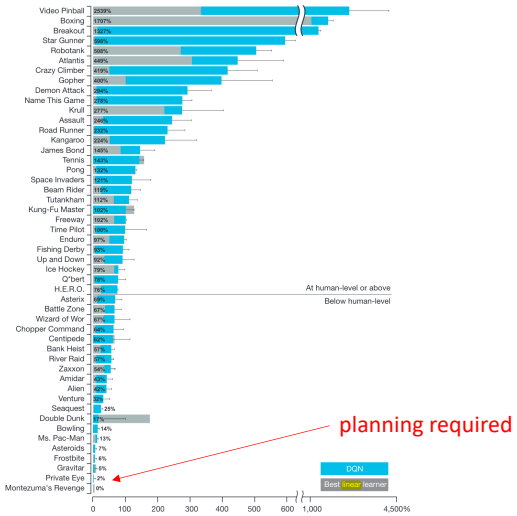
During sleep, and rest, the neurons that fired during the day's behavior reactivated in a similar sequence in the hippocampus (and to some degree the cortex)

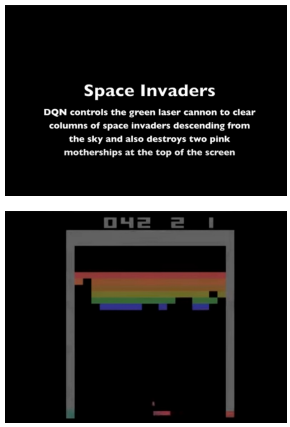
This may drive memory consolidation



WIlson & McNaughton, *Science*, 1994; Ólafsdóttir et al., *Current Biology*, 2018

Above human-performance on video games ….. that don't involve planning



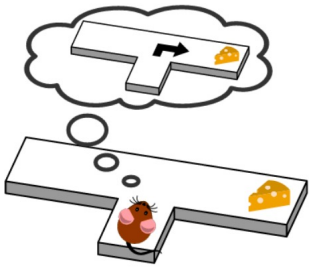


Google Deepmind

Mnih et al., *Nature*, 2015

The elements of a reinforcement learning system (part 6/5)

1. **an agent**
2. **the environment**
3. **a reward signal**
4. **a value function**
5. **a policy**
6. **a model of the environment** is anything the agent can use to predict how the environment will respond to its actions (optional, but awesome).



- An RL system with a model (i.e. a *model-based* system) can
  - *simulate* the next state, an entire trial, or all possible trials
  - *learn value functions* from simulated experience and
  - use this to improve its policy for really interacting with that environment.
  - This entire process is known as *planning.*
- An RL system without a model is known as a *model-free* system

Mattar & Daw, *Nature Neuroscience*, 2018

25

25

Blackboard quiz

- Is Q-learning model-free or model-based?

26

26

Model-free vs model-based Reinforcement learning

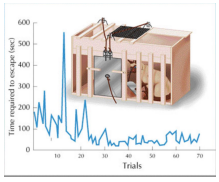
- Model-free RL
- Habits
- Trial & error
- Behaviourism

- Model-based RL
- Goal-directed behavior
- Cognitive maps
- Cognitive Psychology

27

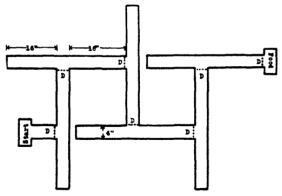
27

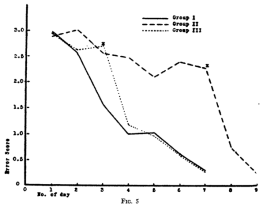
Learning without reinforcement



Thorndike, 1898
- trial & error learning

- Behaviourism
  - all behaviour are learned reactions to stimuli
  - ignore unobservable mental processes (outside the realm of science)





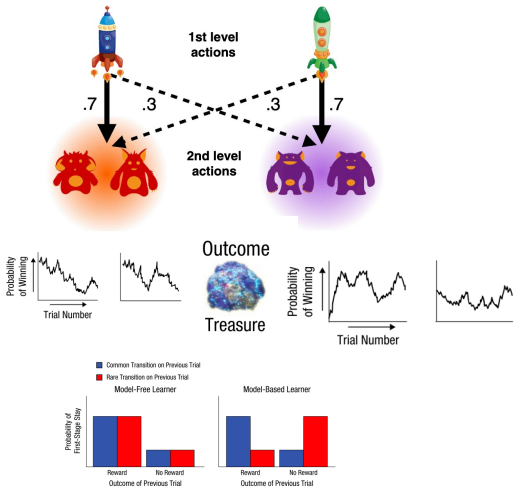
- The environment can be learned without explicit rewards or penalties
- This is not possible with model-free reinforcement learning
- Animals can learn associations between stimuli (states) by experiencing sequences of stimuli as they explore an environment.
- This was a forerunner to Cognitive Psychology, which broke off from behaviourism

Blodgett, 1929

28

28

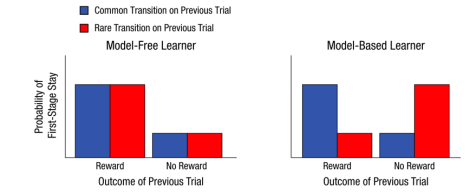
Assessing if people use model-based or model-free reinforcement learning

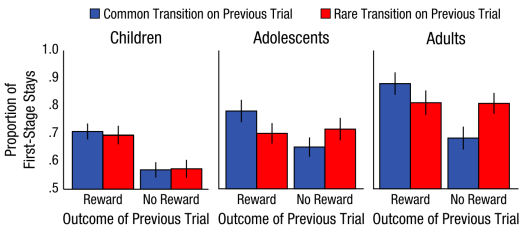


- Do people use model-free or model-based RL?
- This is commonly assessed with this two-step task
  - On each trial, participants choose between two spaceships (first-stage choice), which was followed by a probabilistic transition to a red planet or a purple planet.
  - Then participants choose between two aliens (second-stage choice)
  - and were rewarded with space treasure or not.
  - The probability of winning space treasure changes over time for each alien
- Key analysis – if you are rewarded following a rare transition (e.g. blue spaceship to purple planet), on the next trial should you:
  - be more likely to choose the blue spaceship (model-free RL – trial & error learning)
  - be less likely to choose the blue spaceship (model-based RL – choose the option more likely to get you back to the purple planet?)

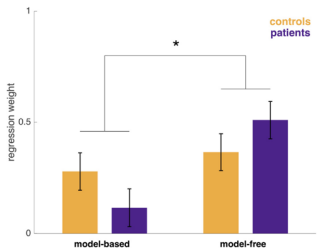
Daw et al., *Neuron*, 2011; Decker et al., *Psych. Science*, 2016

Model-based learning across brains



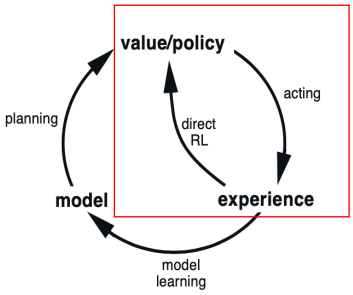


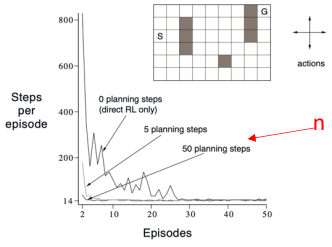
- Adults use a mixture of model-based and model-free RL
- Children use more model-free RL
- Patients with lesions to the hippocampus use more of a model-free RL strategy
- Model-based RL develops over childhood/adolescence, and probably depends (in part) on the hippocampus



Decker et al., *Psych. Science*, 2016; Vikhbladh et al., *Neuron*, 2019

Simulating model-based learning





In the Dyna-Q model-based RL algorithm, learning and planning are accomplished by exactly the same algorithm, operating on real experience for learning and on simulated experience for planning.

Tabular Dyna-Q

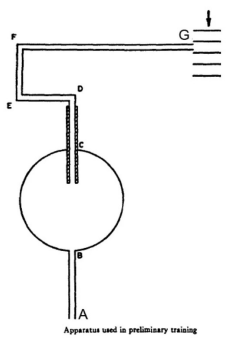
Initialize $Q(s,a)$ and $Model(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$
Do forever:
- (a) $S \leftarrow$ current (nonterminal) state
- (b) $A \leftarrow \epsilon$-greedy(S,Q)
- (c) Execute action A; observe resultant reward, R, and state, S'
- (d) $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$
- (e) $Model(S,A) \leftarrow R, S'$ (assuming deterministic environment)
- (f) Repeat n times:
  - $S \leftarrow$ random previously observed state
  - $A \leftarrow$ random action previously taken in S
  - $R, S' \leftarrow Model(S,A)$
  - $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$

(model-free) Q-learning

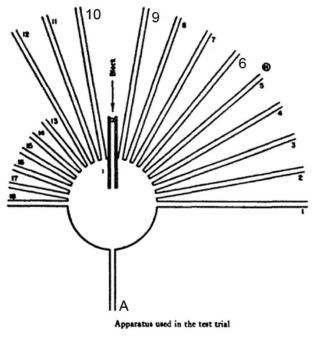
n = planning steps

Sutton & Barto, 2018

Cognitive map



Apparatus used in preliminary training



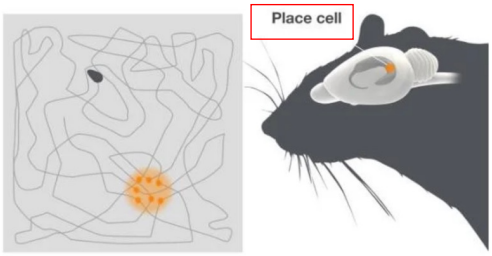
Apparatus used in the test trial

(From E. C. Tolman, B. F. Ritchie and D. Kalish, Studies in spatial learning. I. Orientation and short-cut. *J. exp. Psychol.*, 1946, **36**, p. 17.)

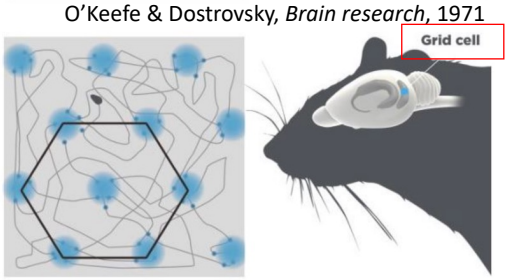
- When the usual route is blocked, and new routes are opened, rats choose the correct direction
- They can do this despite never having experienced that path before (no trial-and-error learning)
- They are likely learning a 'map' of the environment in their mind – a 'cognitive map'
- In other words, the rats must have some mental 'model' of the environment.

Tolman, *Psychological Review*, 1948

The hippocampus and entorhinal cortex contain a cognitive map



O'Keefe & Dostrovsky, *Brain research*, 1971



Hafting et al., Nature, 2008

The hippocampus contains place cells – cells that fire preferentially in particular locations.

the entorhinal cortex (main input to hippocampus) contains grid cells - cells that fire at multiple evenly-spaced locations on a hexagonal (or triangular) grid.

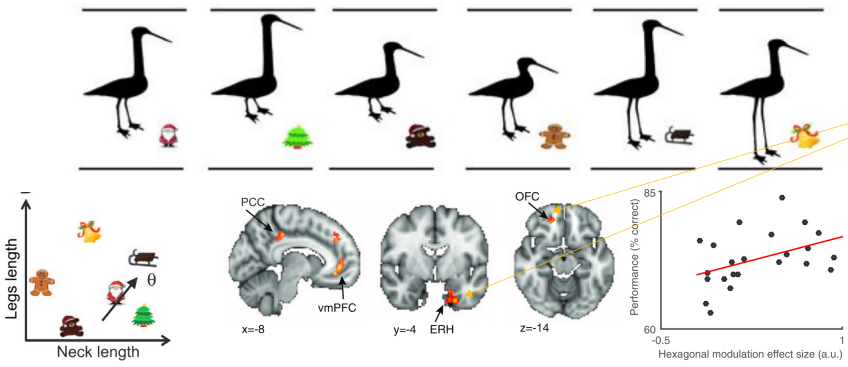
Each location could be a state in a reinforcement learning problem

Nobel prize in physiology or medicine, 2014
for "discovering the brain's GPS system"

33

33

What about problems that don't involve navigating space?



- Despite not being related to physical space, the prefrontal cortex (vmPFC/OFC) and entorhinal cortex (ERH) encoded this abstract space in a hexagonal grid (see slide on 'grid cells')
- This suggests that the same grid-systems used to navigate physical space may be used to plan and navigate abstract problems.

- Participants had to navigate in 'bird space'. Each bird differed in neck and leg length, and was associated with a particular Christmas decoration
- Given one Christmas decoration, they had navigate, by changing neck and leg lengths, to the target Christmas decoration.

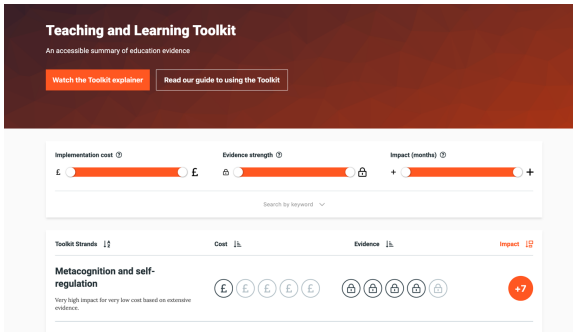
Constantinescu et al., *Science*, 2016

34

34

Recap - Week 1 - The most effective strategy to improve learning?

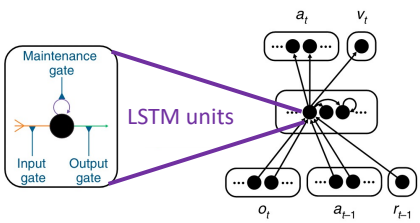
- Learning-to-learn (meta-cognition)
- Identifying specific strategies for planning, monitoring, and evaluating your own learning.
- Develop a repertoire of strategies to choose from and the skills to select the most suitable strategy for a given learning task.
- Self-regulated learning can be broken into three essential components:
  - cognition – the mental process involved in knowing, understanding, and learning
  - metacognition – often defined as 'learning to learn';
  - motivation – willingness to engage our metacognitive and cognitive skills.



https://educationendowmentfoundation.org.uk/education-evidence/teaching-learning-toolkit

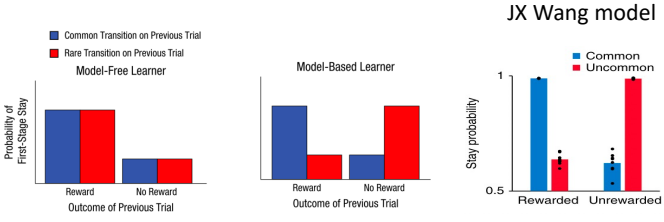
35

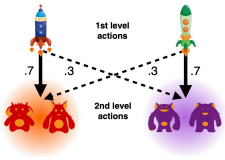
Learning to reinforcement learn



trained with model-free RL

JX Wang model



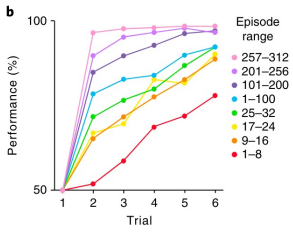


- Weights frozen during test time – any 'learning' occurs due to dynamics of activity, not changes to weights
- Deep & recurrent neural networks can approximate any function – including model-based RL
- Model-free RL can be used to learn model-based RL (learning to reinforcement learn)

J.X. Wang et al., *Nature Neuroscience*, 2018

36

Learning to learn- monkeys



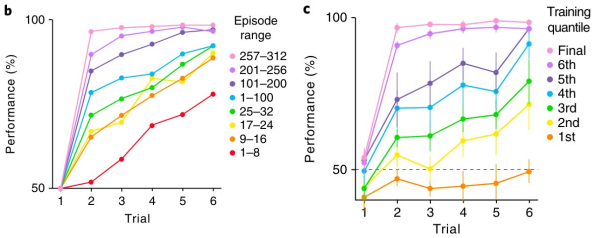
- The rule. One of the two stimuli (picture cards) is rewarded. The other is not.
- The monkey first learns by trial-and-error which stimulus is rewarded.
- Then the stimuli are changed.
- The monkey eventually learns the rule, enabling it to 'learn' how to respond to new stimuli after just a single trial.

Harlow, *Psychol. Rev.*, 1949

37

37

Learning to learn with ConvNets, RNNs & reinforcement learning



- A ConvNet is used to present different stimuli
- The RNN (with LSTM units) is trained as before with model-free RL
- The network 'learns to learn' more efficiently – seemingly learning the rule, which is not possible under model-free RL.

Harlow, *Psychol. Rev.*, 1949 ; J.X. Wang et al., *Nature Neuroscience*, 2018

38

38

Recap - Week 1 - Active engagement – curiosity

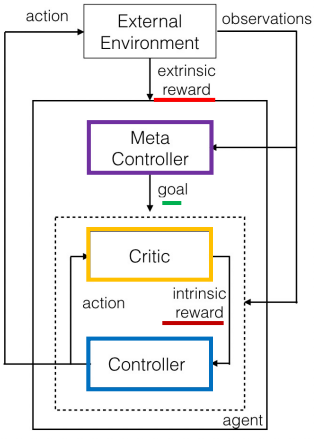
- Curiosity occurs whenever we detect a gap between what we want to know and what we already know. (Loewenstein, 1994)
- Curiosity is driven by the value of acquiring information
- Greater curiosity is associated with greater academic achievement (Shah et al., 2018)

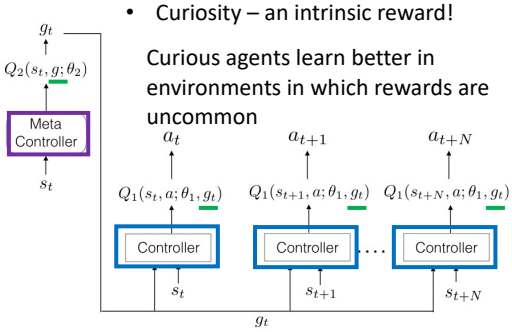
Neil deGrasse Tyson
(astrophysicist & science communicator)

39

39

Curiosity – an intrinsic reward signal





- How to deal with environments that give feedback rarely?
- Curiosity – an intrinsic reward!

Curious agents learn better in environments in which rewards are uncommon

Performance on 'Montezuma's revenge' - (compare with Mnih et al., deep Q network)

- The *Meta-controller* chooses a goal to maximise future extrinsic reward
- The *Controller* chooses among actions according to the current goal
- The *Critic* gives 'intrinsic reward' to controller if the current goal is complete

Kulkarni et al., *NeurIPS*, 2016; Gottlieb & Oudeyer, *Nature Rev. Neurosci.,* 2018

40

40

Blackboard Quiz

- Provide a definition of curiosity from a reinforcement learning point of view.

41

41

Recap

- The goal of reinforcement learning is to maximise the total amount of (discounted) reward in the future.
- The temporal difference (TD) learning model successfully explains many aspects of classical conditioning (state-value prediction)
- The reward prediction error from TD-learning strongly correlates with dopamine neuron activity in the brain
- Q-learning is a successful off-policy algorithm for learning action-values
- Deep RL approximates the value function – enabling RL in environments with many states
- Replay of past memories (like in the hippocampus) helps deep RL to learn
- A model of the environment is anything the agent can use to predict how the environment will react to its actions
- Models of the environment ('cognitive maps') are built in the hippocampus & orbitofrontal cortex (and potentially other areas)
- An intrinsic reward signal – like curiosity – can help agents learn in environments that rarely give out rewards.

42

42

Still curious? Check out these references

- The foundational book on reinforcement learning
  - Sutton, Richard S., and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018 (2nd edition).
- The main paper describing the link between the TD reward prediction error & dopamine neuron activity
  - Schultz, Wolfram, Peter Dayan, and P. Read Montague. "A neural substrate of prediction and reward." *Science* 275, no. 5306 (1997): 1593-1599.
- Deep reinforcement learning
  - Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves et al. "Human-level control through deep reinforcement learning." *nature* 518, no. 7540 (2015): 529-533.
- Hippocampal replay
  - Wilson, Matthew A., and Bruce L. McNaughton. "Reactivation of hippocampal ensemble memories during sleep." *Science* 265, no. 5172 (1994): 676-679.
- Cognitive maps
  - Tolman, Edward C. "Cognitive maps in rats and men." *Psychological review* 55, no. 4 (1948): 189.
- Learning to reinforcement learn
  - Wang, Jane X., Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. "Prefrontal cortex as a meta-reinforcement learning system." *Nature neuroscience* 21, no. 6 (2018): 860-868.
- Curiosity & intrinsic rewards
  - Kulkarni, Tejas D., Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. "Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation." *Advances in neural information processing systems* 29 (2016)

43

43