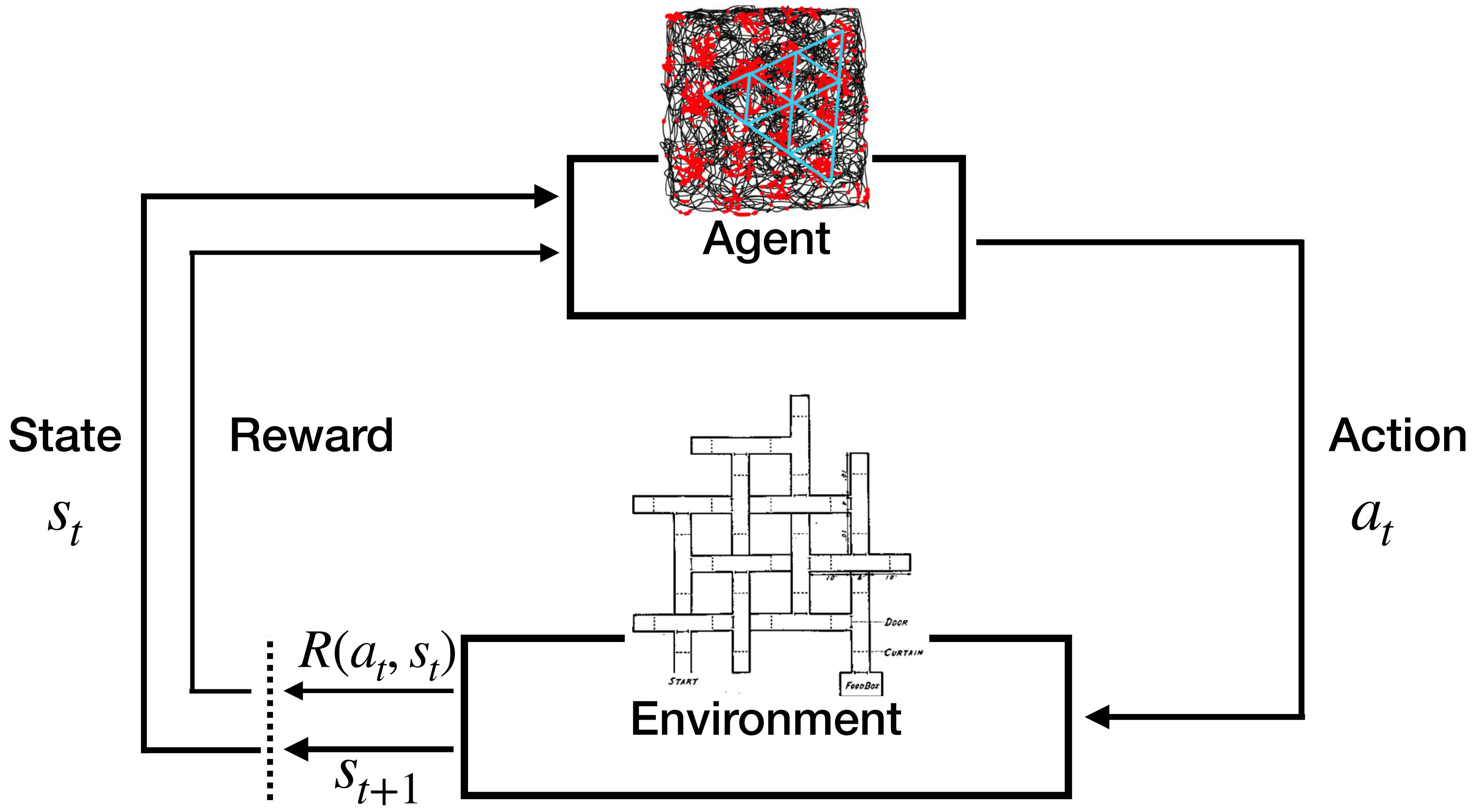


RL in the brain (+ behaviour)

Cognitive Maps Seminar
09th of November 2022

Admin Recap - form groups

- [Required] **Attendance** of at least 80% of sessions

- [30% of grade] **Submit 1 engaging discussion question** prior to every paper session
  - 16. November onwards - **next week**!
  - List: https://docs.google.com/spreadsheets

- [70% of grade] **Give one presentation** (90-minute session with discussion) on a relevant paper of your choice
  - In a group of 3-4 students
  - List: https://docs.google.com/spreadsheets

https://hmc-lab.com/Cogmaps.html

Groups - Preferences as of 08 Nov

Rules
- Groups of 4 are full (but you can swap)
- At least 3 students per group
- 7 papers in total

Your Job
- Can the unassigned students assign themselves to one of the open groups? (Or we will do so next week)
- **Group** and **paper change** is possible until next week - self-organised according to the rules on the left
- Email us if there are **dates** that totally don't work for you

Brunec, I. K., & Momennejad, I. (2022). Predictive representations in hippocampal and prefrontal hierarchies. Journal of Neuroscience, 42(2), 299-312.	**Timcenko**, Aleksejs	**Schach**, Katja	**Verde Puerto**, Paula	**Gekeler**, Franziska
Pouncy, T., Tsividis, P., & Gershman, S.J. (2021). What is the model in model-based planning? Cognitive Science, 45, e12928.	**Xiong**, Yirong	**Prasad**, Shweta	**Gholamzadeh**, Ali	**Grötzinger**, Dennis
Cruse, H., & Wehner, R. (2011). No need for a cognitive map: decentralized memory for insect navigation. PLoS computational biology, 7(3), e1002009.	**Leerssen**, Paige	**Bailey**, Mark	**John**, Dan	
Buzsáki G, Tingley D. Space and Time: The Hippocampus as a Sequence Generator. Trends Cogn Sci. 2018;22(10):853-869	**Liu**, Jiatong	**Tammaro**, Ruben	**Lin**, Yuguang	
Peer, M., Brunec, I. K., Newcombe, N. S., & Epstein, R. A. (2021). Structuring knowledge with cognitive maps and cognitive graphs. Trends in cognitive sciences, 25(1), 37-54.	**Müller**, David	**de Oliveira**, Paula	**Heuschkel**, Simon	
He, Q., Liu, J. L., Eschapasse, L., Beveridge, E. H., & Brown, T. I. (2022). A comparison of reinforcement learning models of human spatial navigation. Scientific Reports, 12(1), 1-11.	**Wolters**, Peter			
Eldar, E., Lièvre, G., Dayan, P., & Dolan, R. J. (2020). The roles of online and offline replay in planning. eLife.	**Mehnert**, Lena			

Unassigned: **Missori**, Janù **Asik**, Ayberk **Kossack**, Daniel **Höfer**, Antonia **García Manzano**, Laura **Barbashova**, Nadezhda

Recap: so why do we care about RL?

Do you remember the difference between:

$$V_\pi(s) \qquad Q(s, a) \qquad P(s', r \mid s, a)$$

RL examples

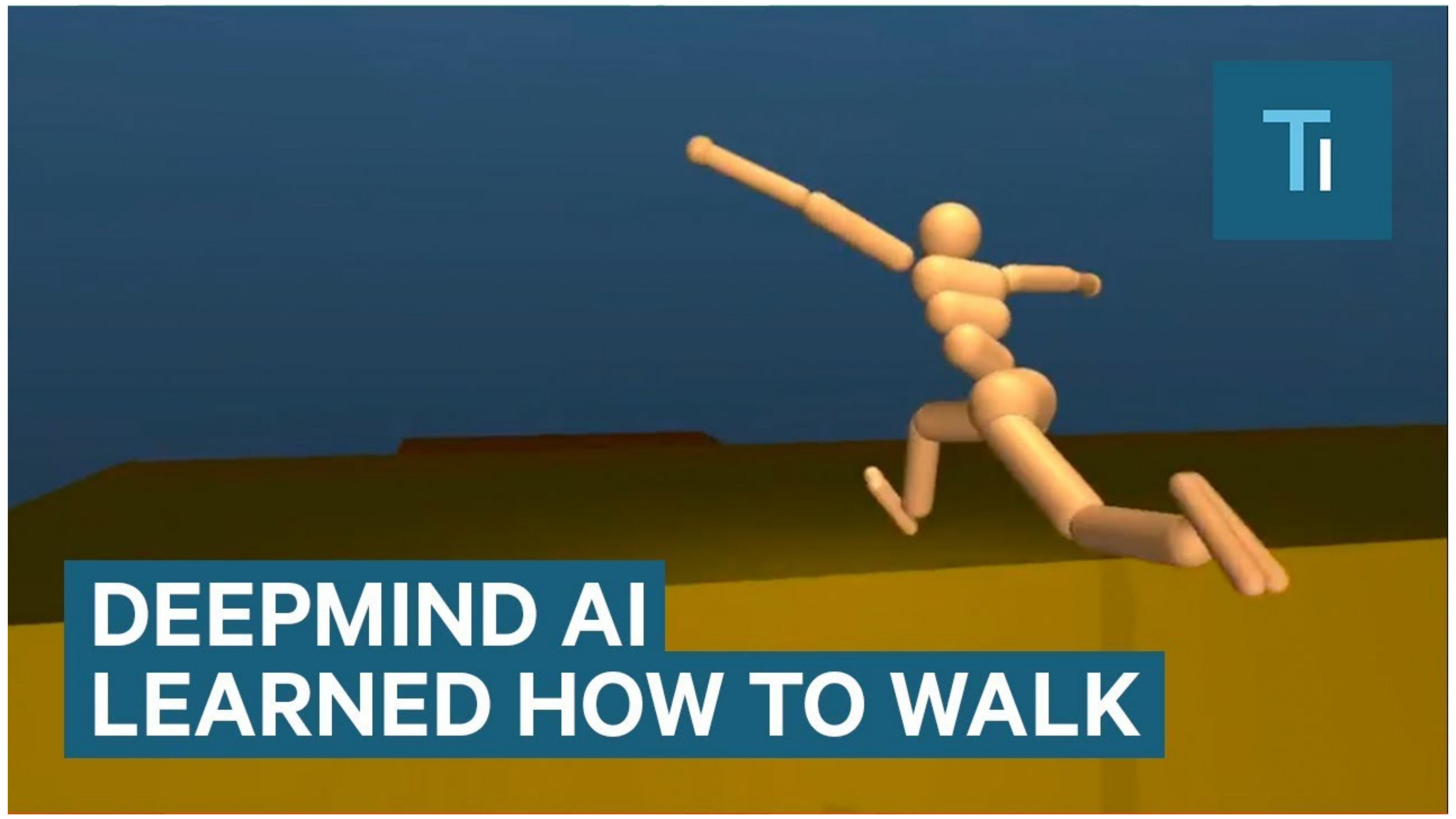
Learn useful actions:

RL examples

A few hours (+a bit of evolution) after birth:

RL examples

This process is perhaps not too
different from AI learning to walk:



What is reinforcement learning (RL)?

- RL is a **computational approach** to learning from **interactions** with the **environment**
  - Trial-and-error
  - Delayed reward
- Considers whole problem of **goal-directed** agent interacting with an **uncertain** environment
- RL agents
  - Have explicit goals
  - Sense aspects of their environments
  - Choose actions to influence their environments

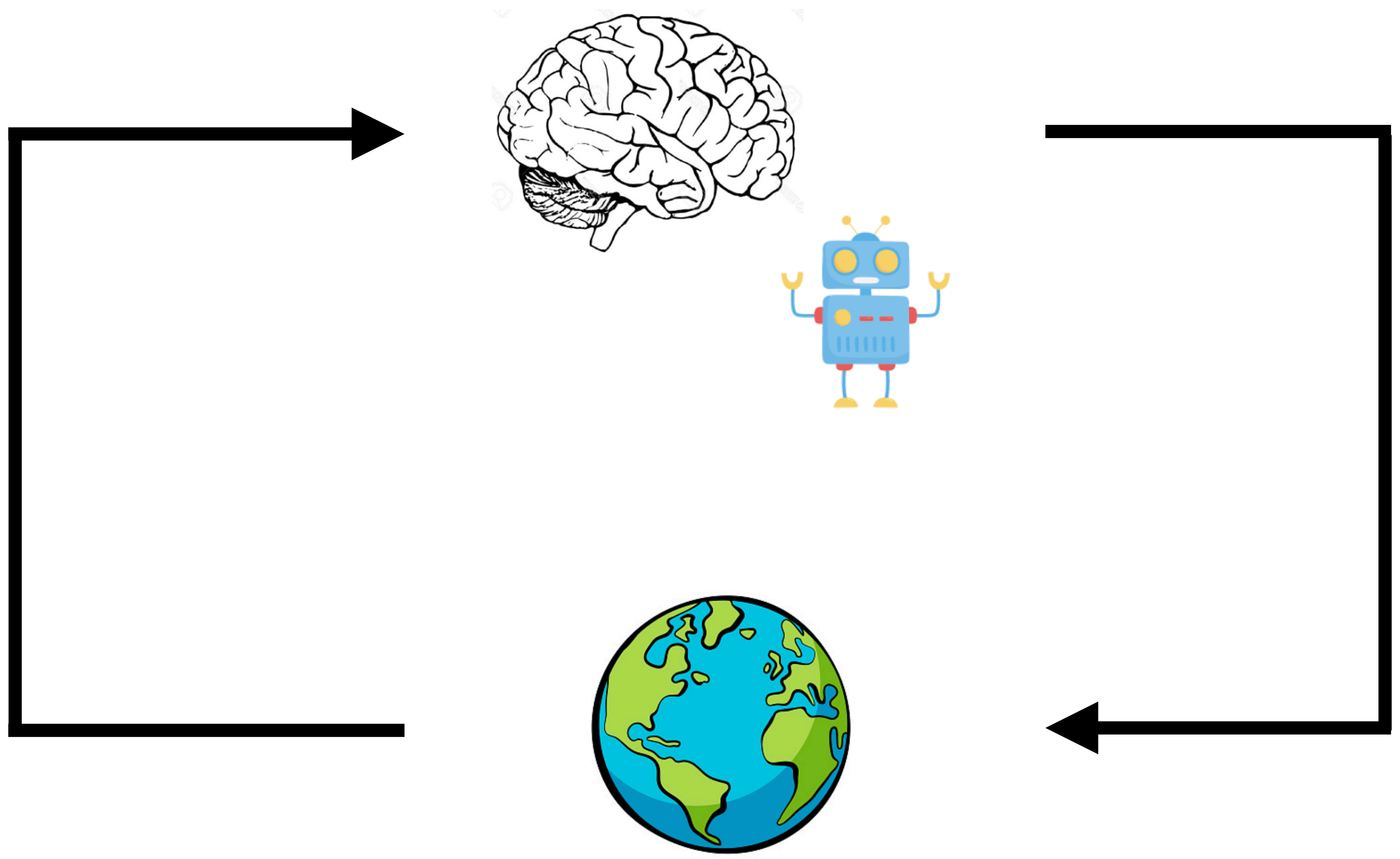
Basic setup: how do agents learn to act?

1. Based on a reward signal, agents learn **values of actions/states:**

$$V_\pi(s) = \mathbb{E}_\pi[R \,|\, s_0 = s]$$

2. Action is governed by a **policy**:

$$\pi(a, s) = P(a_t = a \,|\, s_t = s)$$



Reward r_t

Action a_t

State s_t

3. Agents can learn a **model of the environment** to make smarter decisions, e.g.:

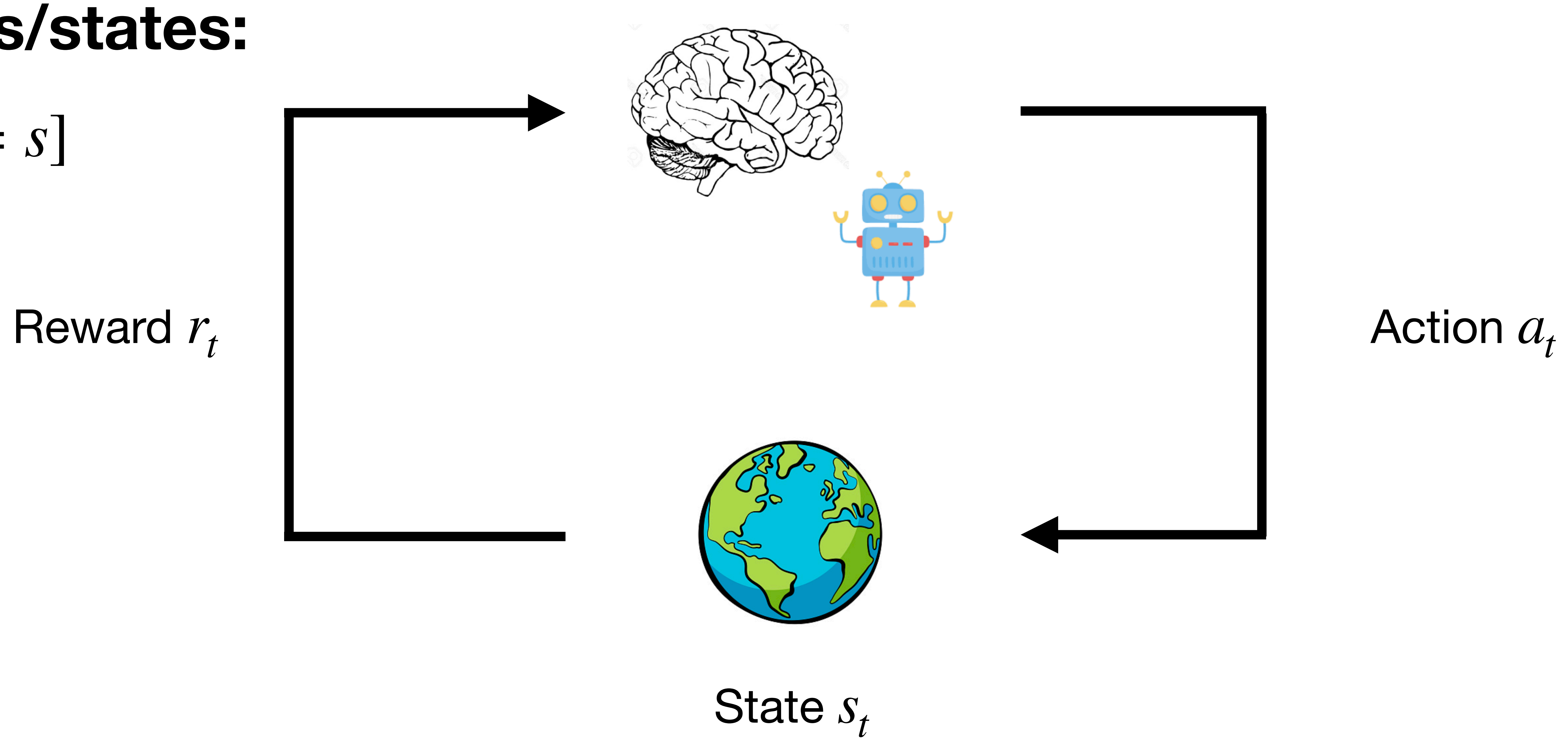
$$P(s_{t+1} = s, r_{t+1} = r \,|\, s_t = s, a_t = a)$$

1. Value and Value Learning (in the brain)

Values

Based on a reward signal, agents learn **values of actions/states:**

$$V_\pi(s) = \mathbb{E}_\pi[R \mid s_0 = s]$$



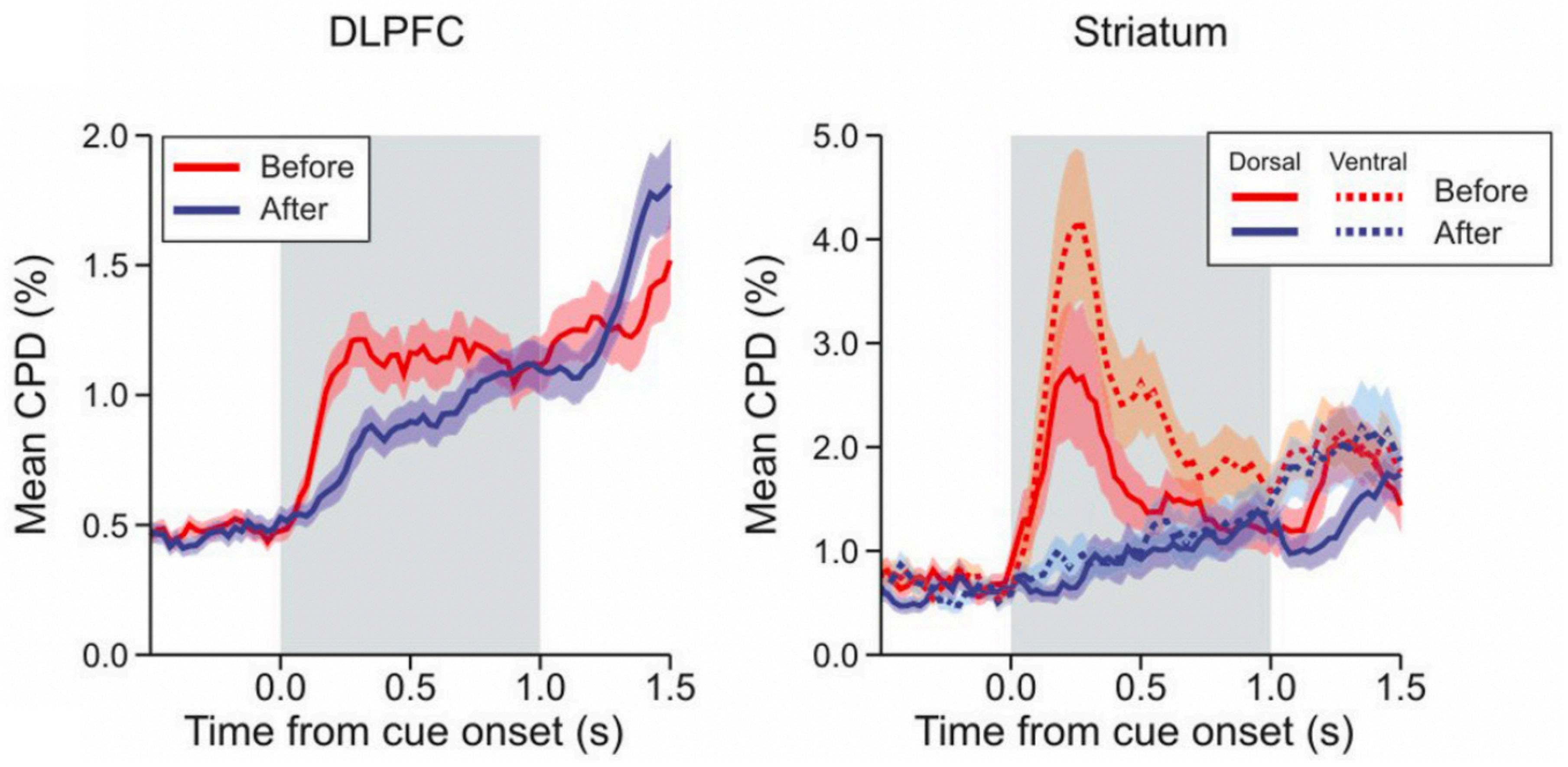
Values approximate long-term future reward

Values

Nature of value representation changes - value of states vs. chosen value

$V_\pi(s)$

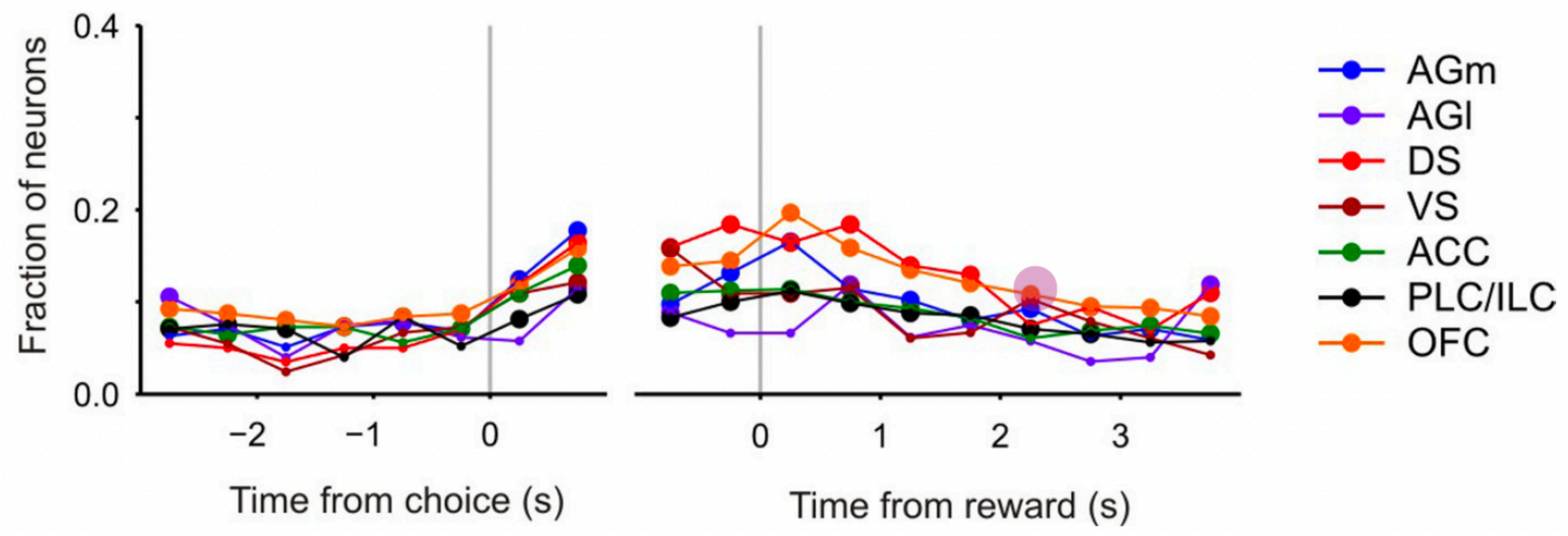
$Q(s, a_{chosen})$

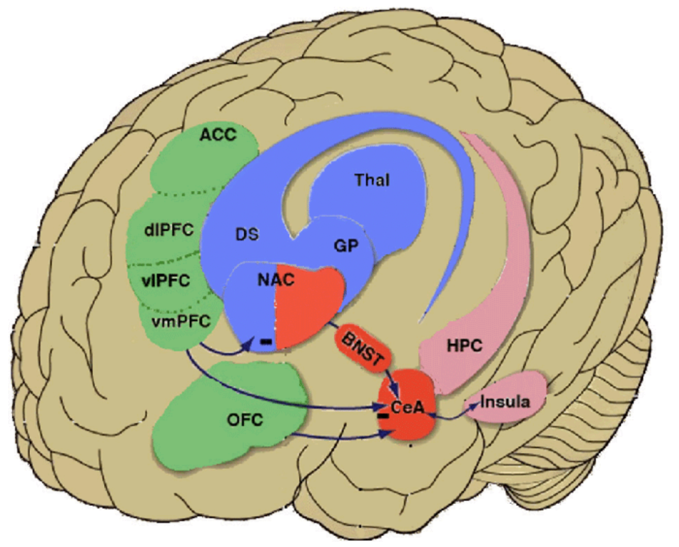


Lee et al., Annu Rev Neurosci. 2012

Value represented in a lot of brain regions

$Q(s, a_{chosen})$





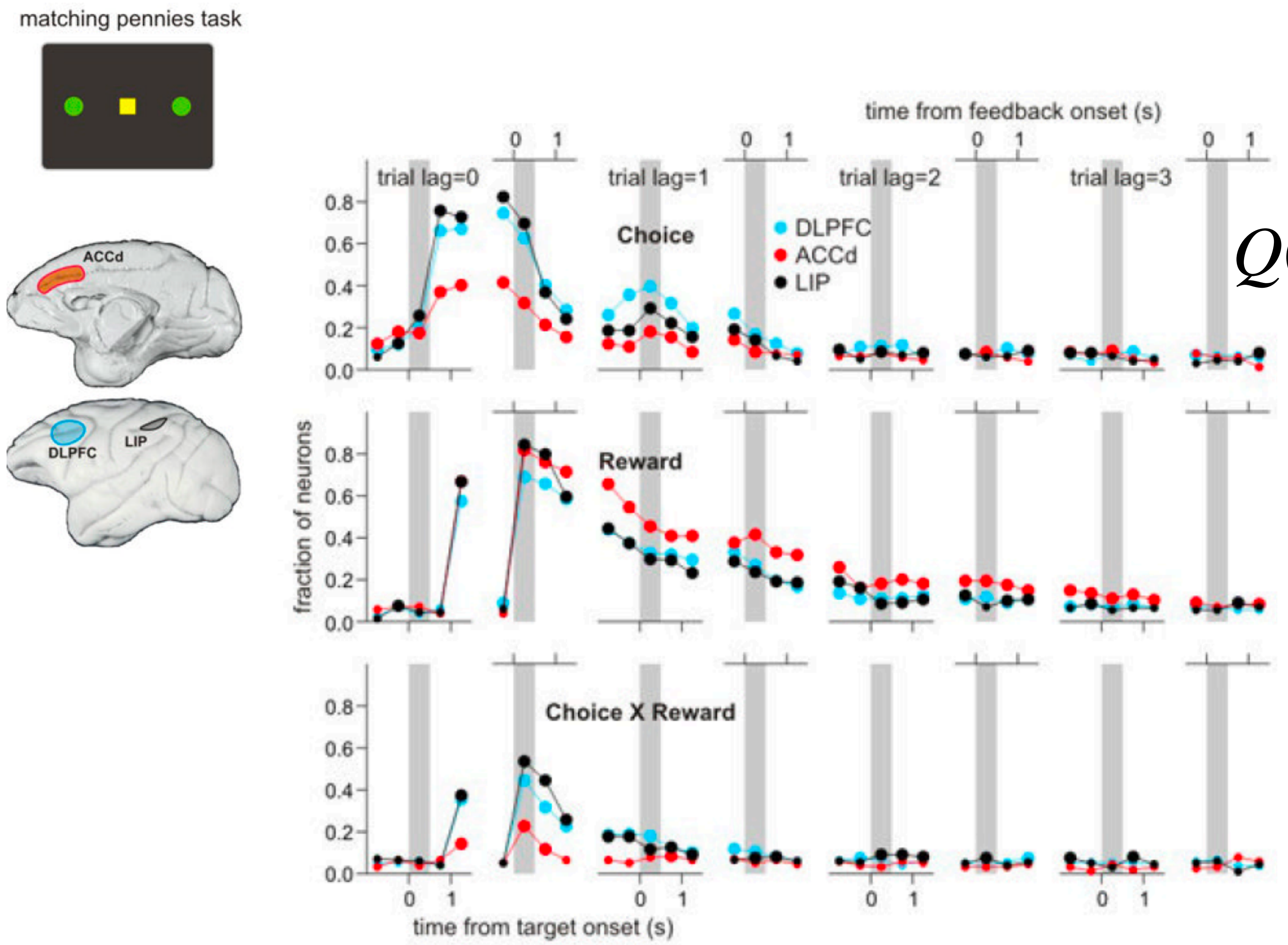
Zorrilla & Koop, 2019

Lee et al., Annu Rev Neurosci. 2012

Values at choice and outcome

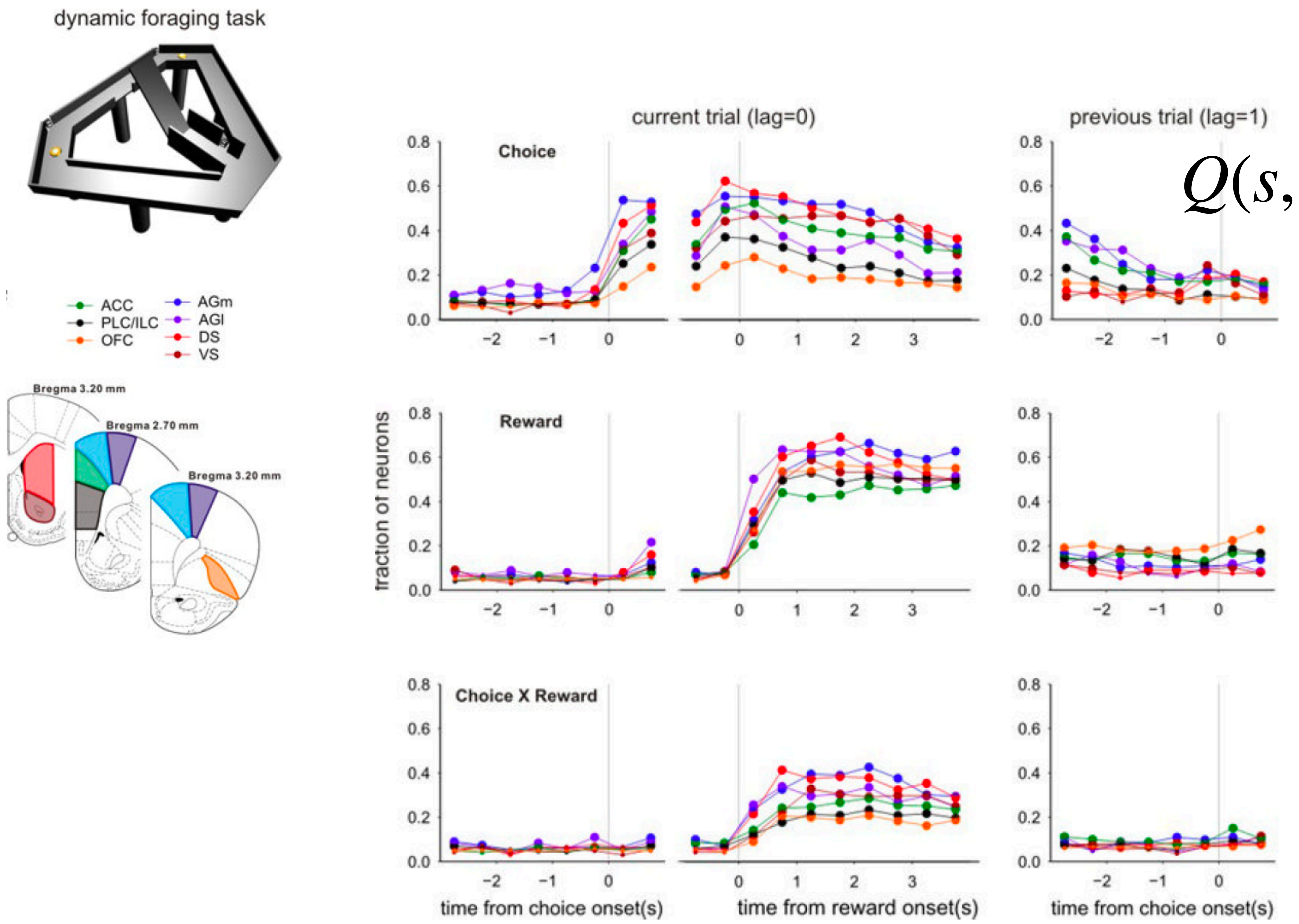
Key RL variables in different brain regions: choice, outcome, and choice x outcome

- Some of them even represent past trials



$Q(s, a_{chosen})$

Lee et al., Annu Rev Neurosci. 2012

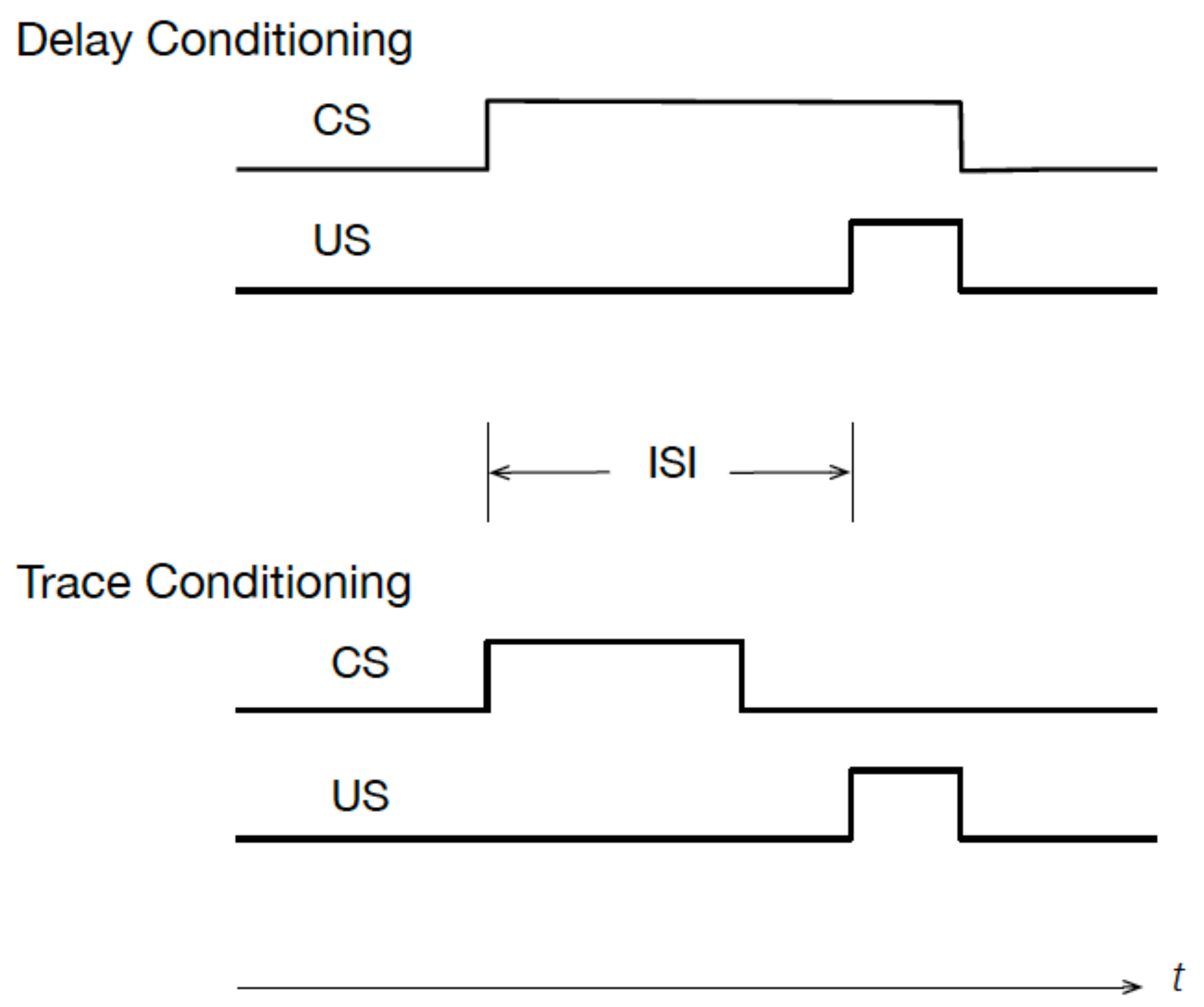


$Q(s, a_{chosen})$

Lee et al., Annu Rev Neurosci. 2012

Learning values

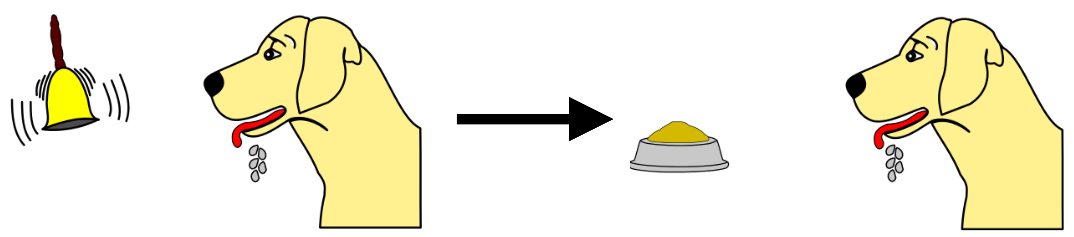
- Two learning algorithms you should know about:
  - **Rescorla-Wagner (RW-)Learning**
    - Learn stimulus-outcome associations
  - **Temporal Difference (TD-)Learning**
    - Learn stimulus-outcome associations *across* time

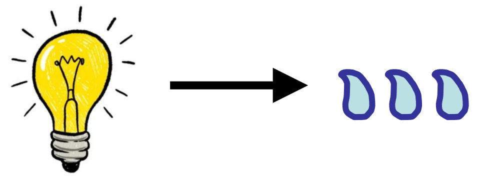


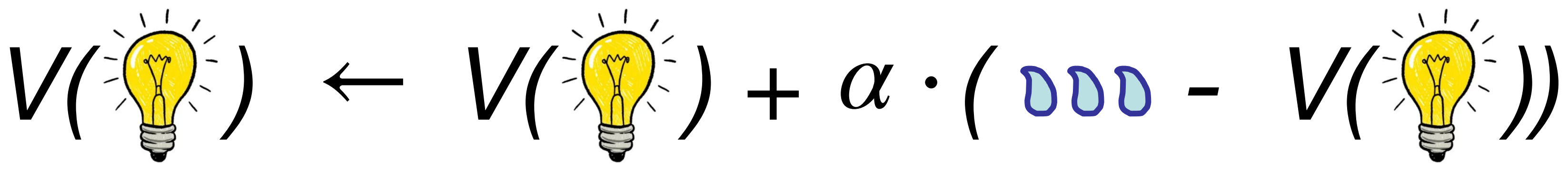
Basics of Learning: Rescorla-Wagner Learning

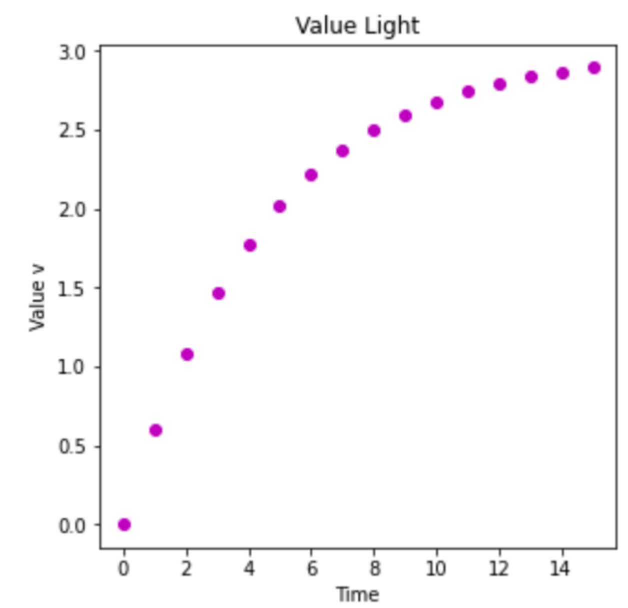
Learn associative strength between a CS and US

$$V(s) \leftarrow V(s) + \underbrace{\alpha}_{\text{Learning rate}} \cdot \overbrace{(r - V(s))}^{\text{Prediction error}}$$



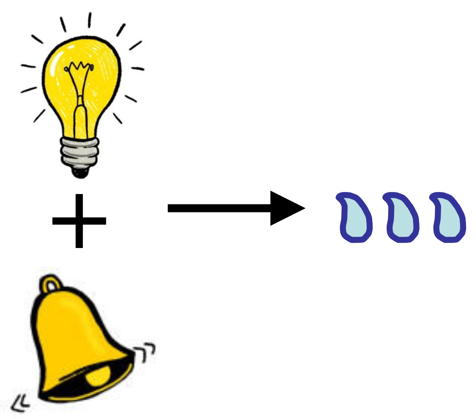


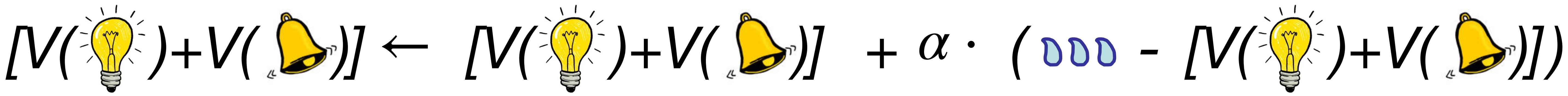




Link to code here

Introducing a second CS
can lead to **blocking**:





Temporal Difference Learning

- “If one had to identify one idea as **central** and **novel** to reinforcement learning, it would undoubtedly be temporal-difference (TD) learning.”

- Update based on other learned estimates, without waiting for final outcome (**bootstrap**)
    - Learn “a guess from a guess”

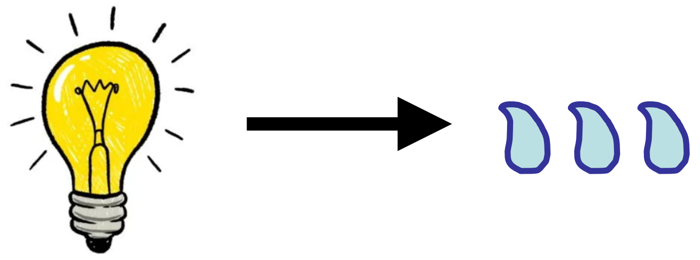
$$V(s_t) \leftarrow V(s_t) + \underbrace{\alpha}_{\text{Learning rate}} \cdot \overbrace{(r + \underbrace{\gamma}_{\text{Discount rate}} \cdot V(s_{t+1}) - V(s_t))}^{\text{Prediction error}}$$

- Operates in ‘real-time’
    - t labels time steps within trials
    - Think of time between t and $t+1$ as a small time interval (e.g. 1ms)

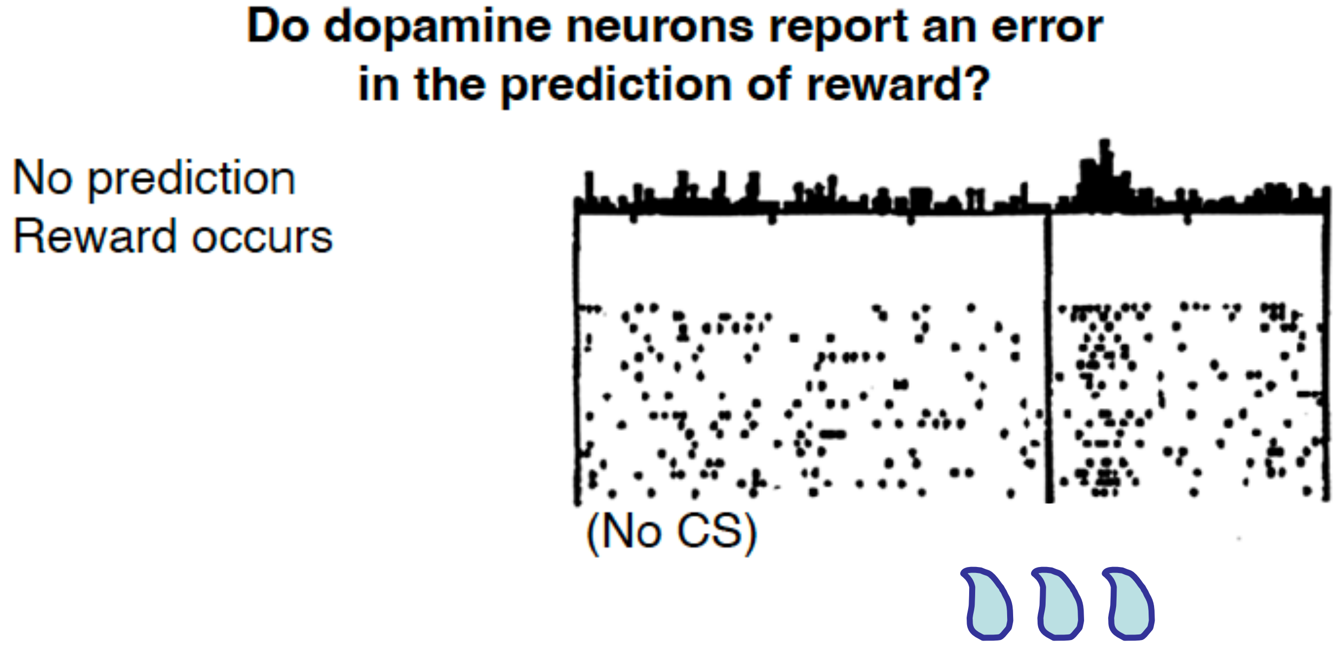
Can RL tell us anything about the brain?

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (r + \gamma \cdot V(s_{t+1}) - V(s_t))$$

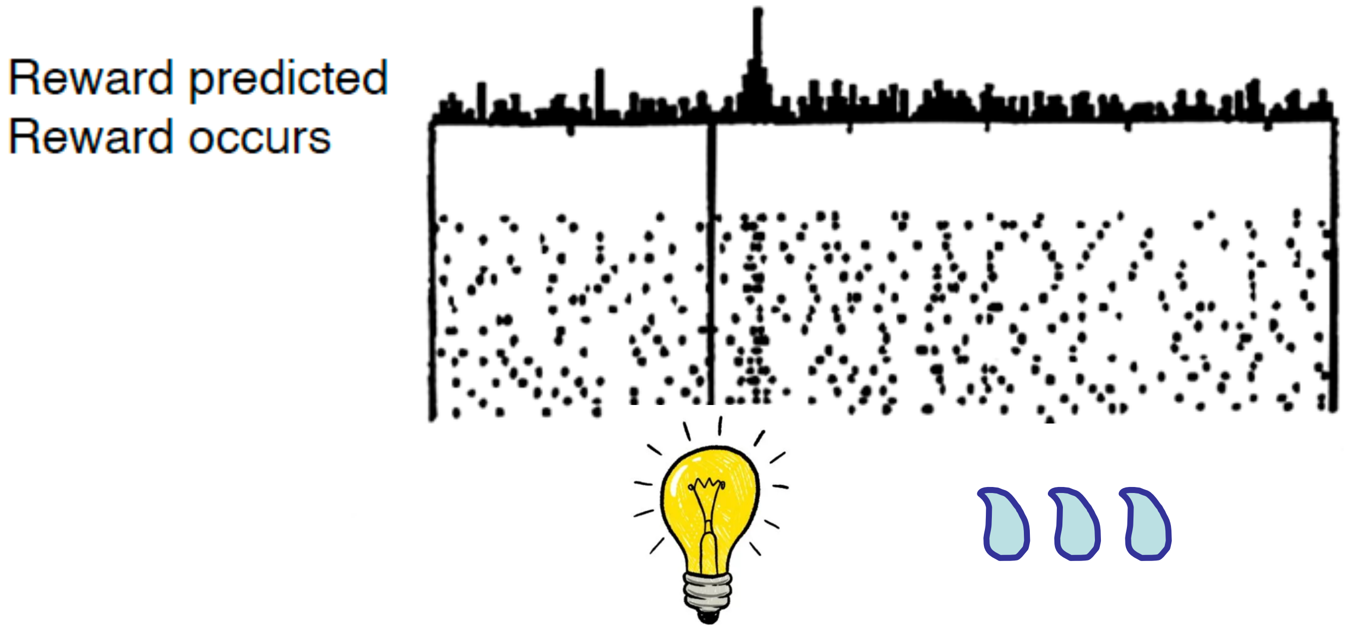
- Yes, quite a lot.
- Particularly, it looks like dopamine (DA) is a key neurotransmitter for (TD) reward learning



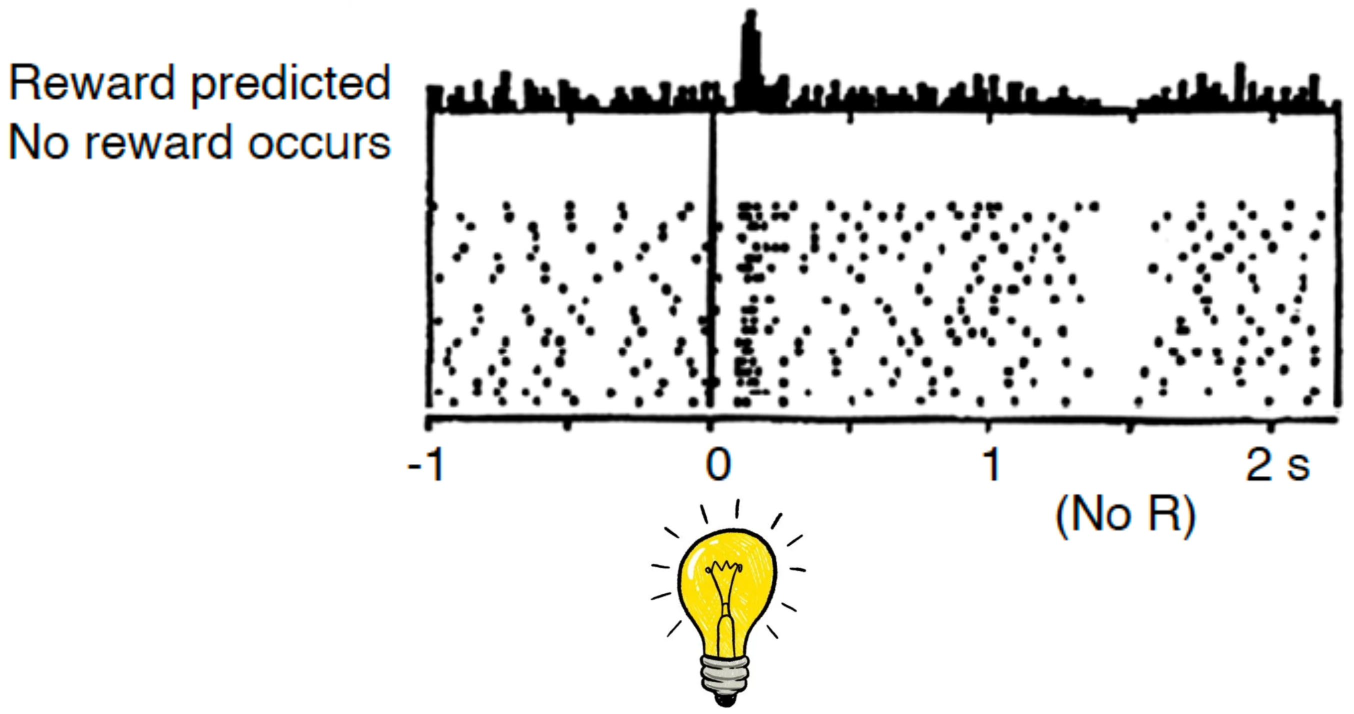
Dopamine neurons signal immediate reward



BUT: after training…
- DA signal reward prediction
- But not correctly predicted reward!



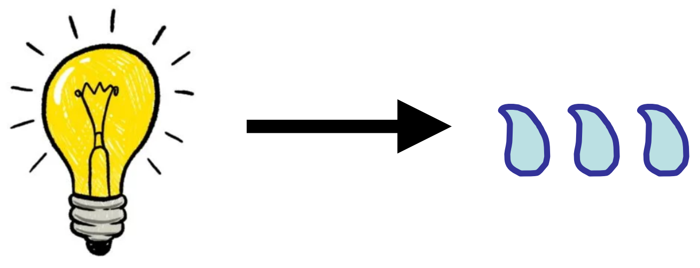
AND: it signals the unexpected omission of a reward!

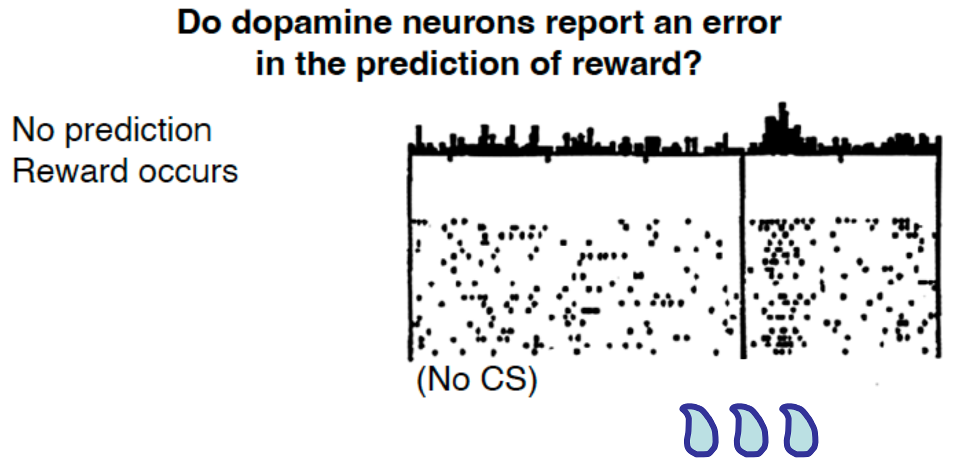


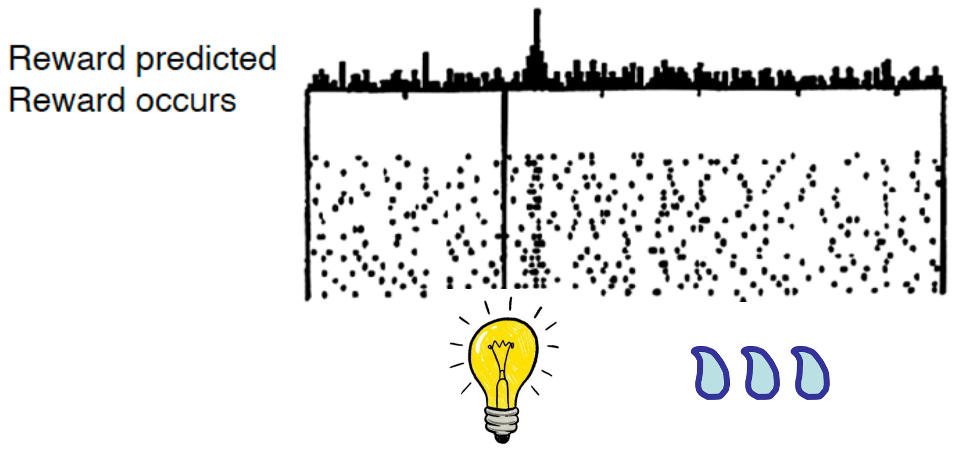
Schultz, Dayan & Montague (Science, 1997)

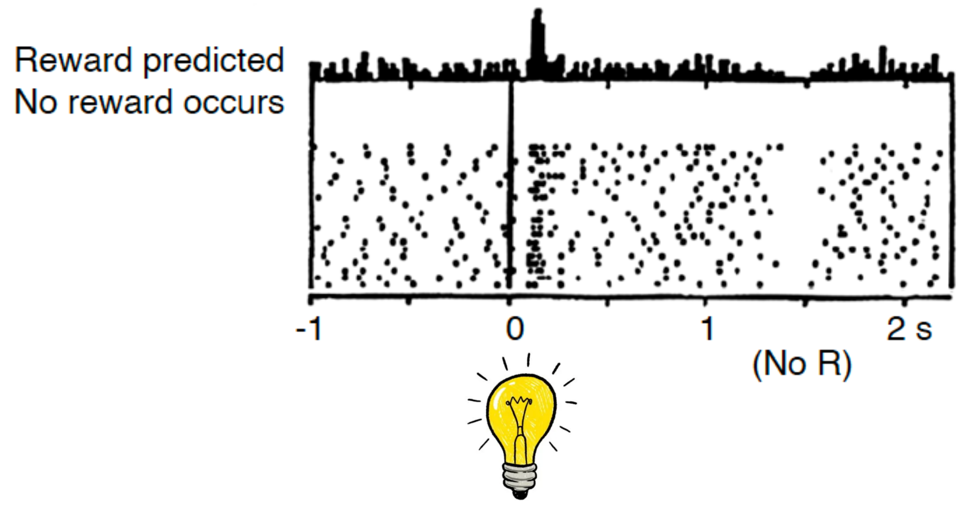
Temporal Difference Learning

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (r + \gamma \cdot V(s_{t+1}) - V(s_t))$$

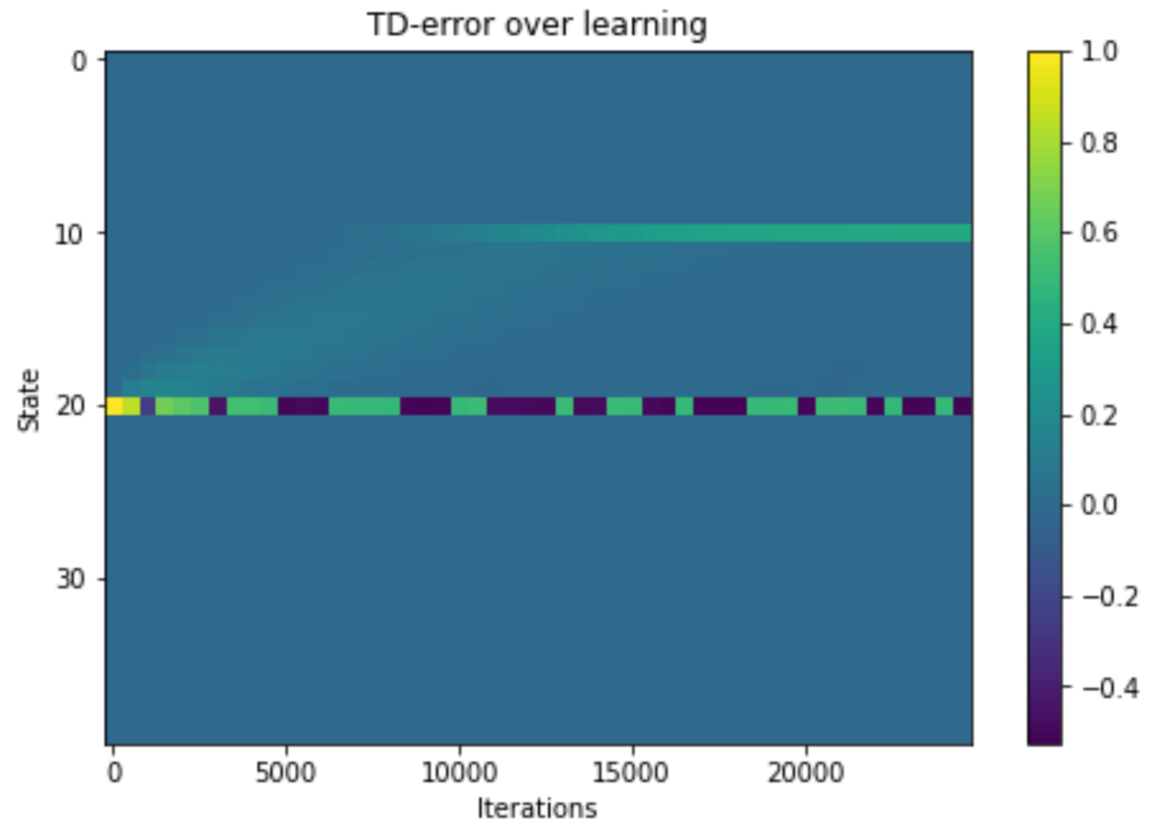




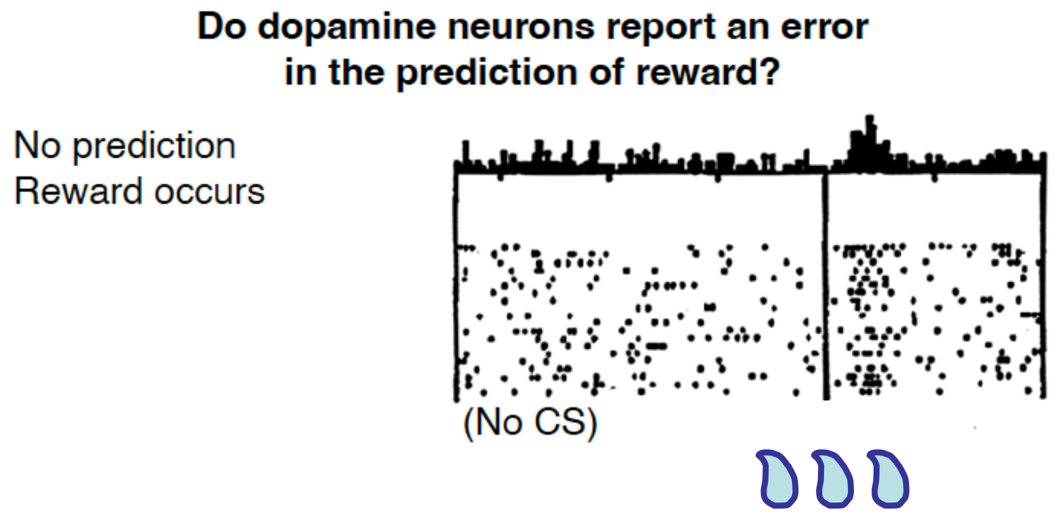


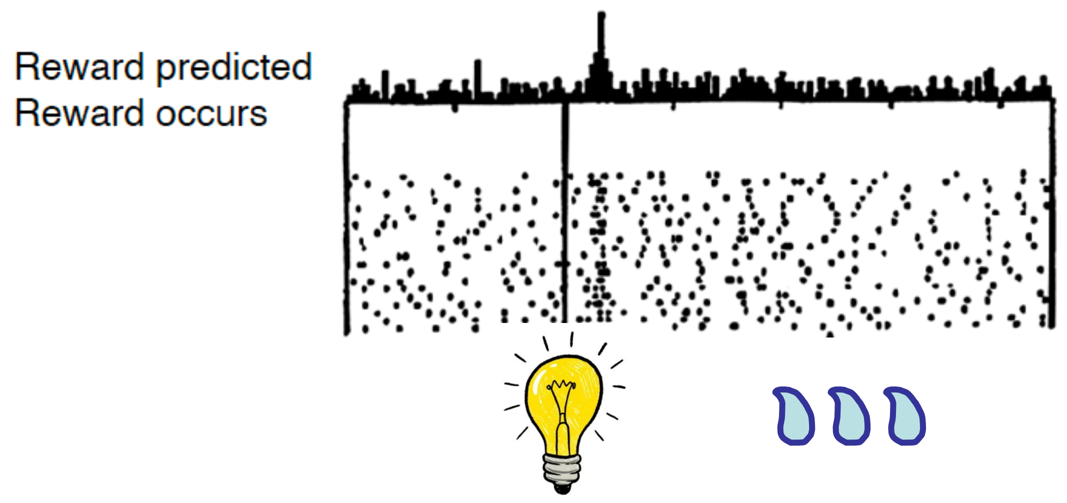


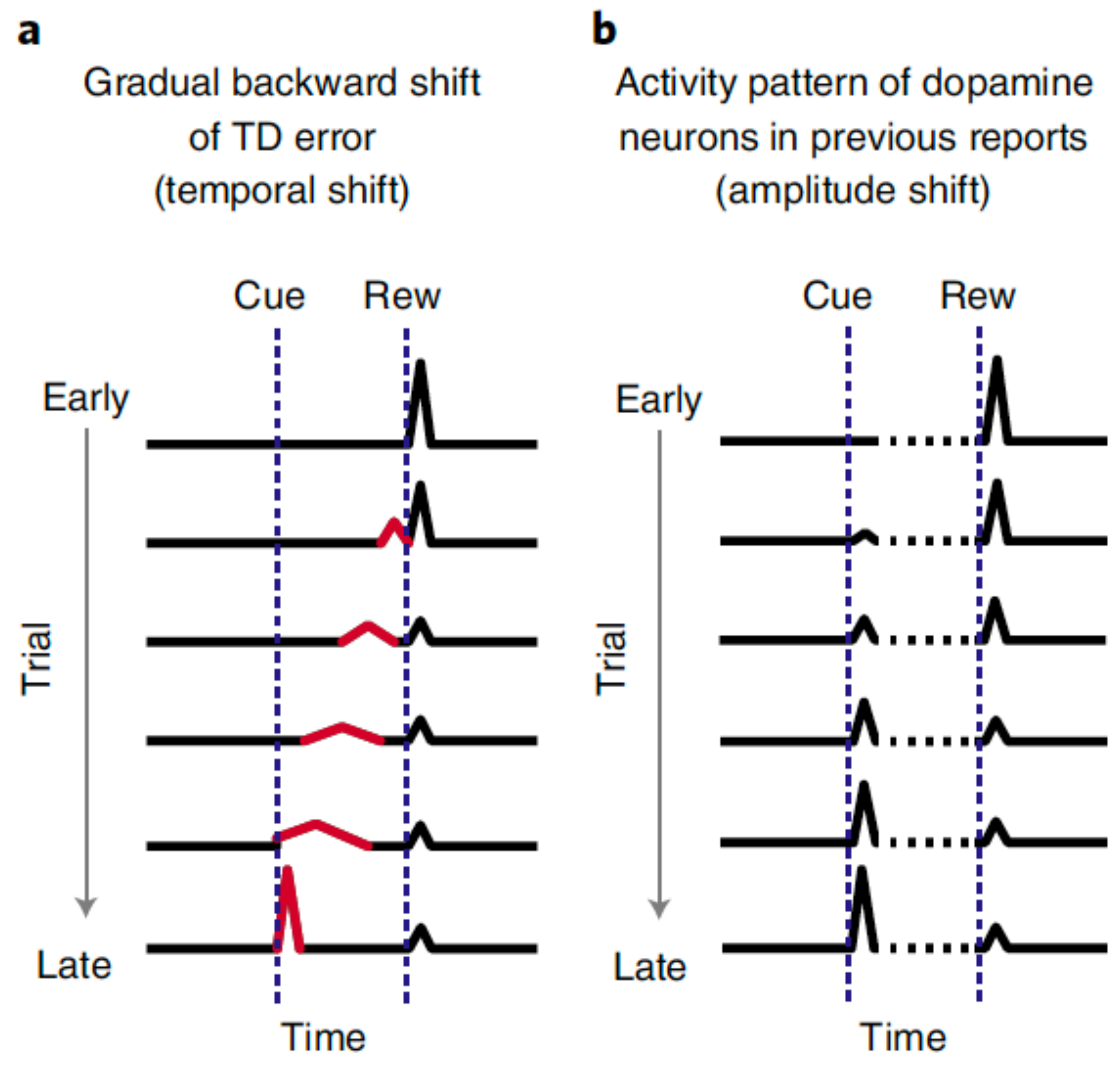
We can simulate this (link to code here):

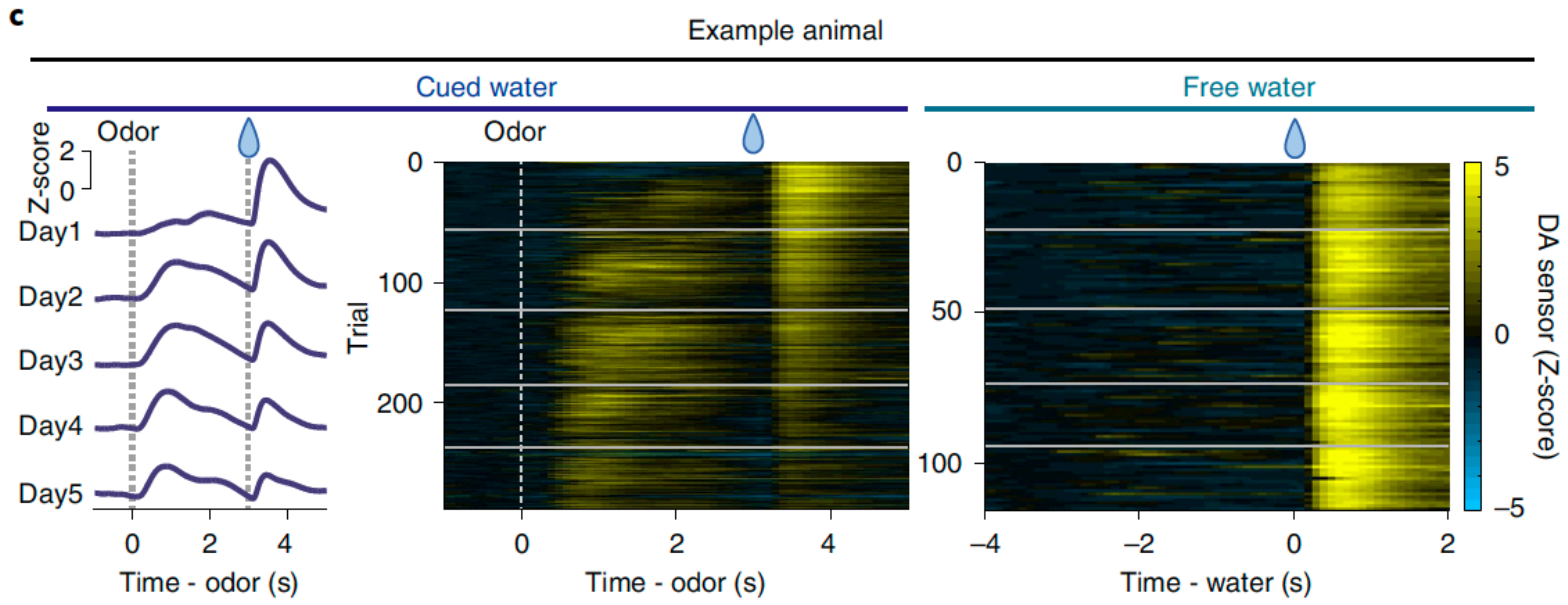


More TD-learning









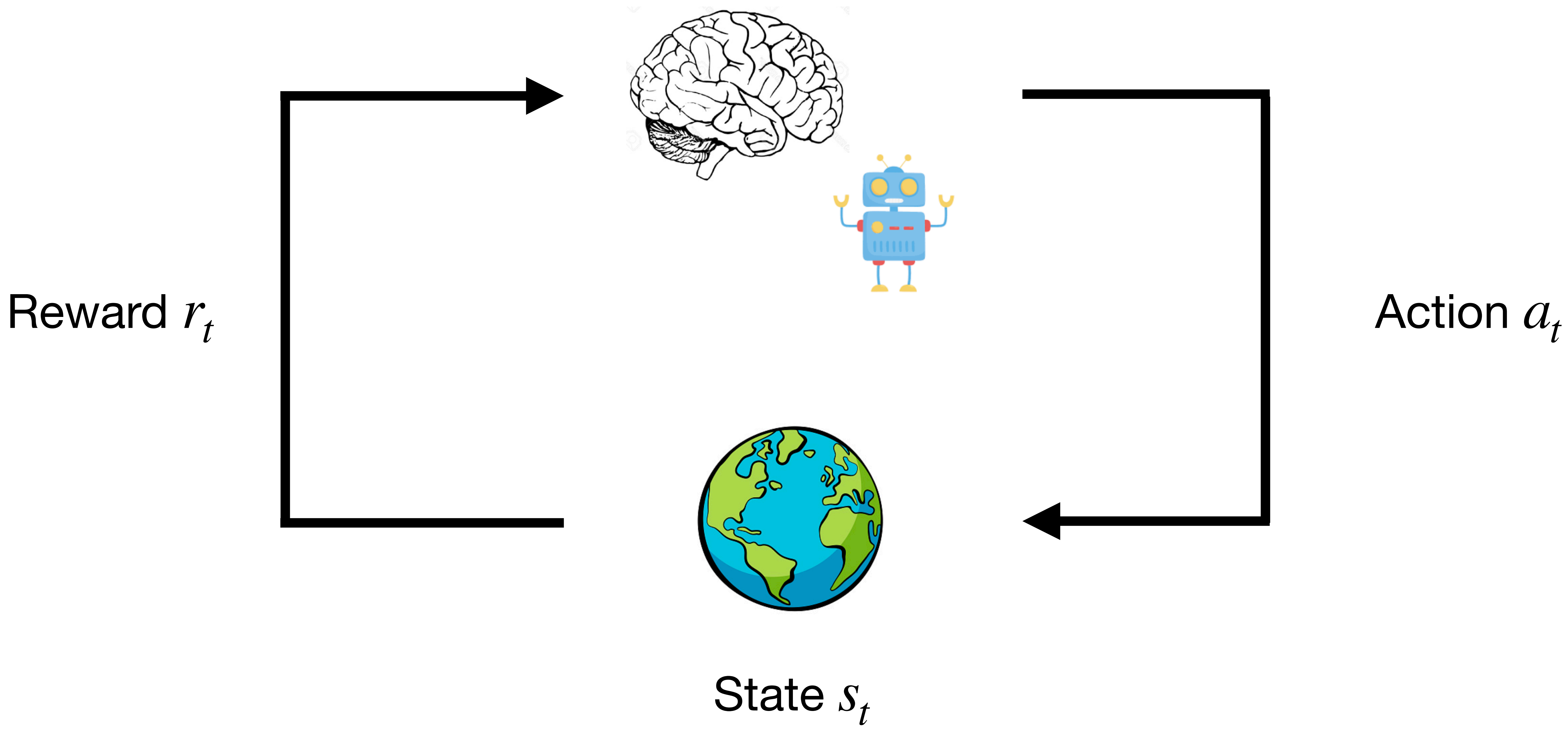
Amo, …, Watabe-Uchida, Nature Neuroscience 2022

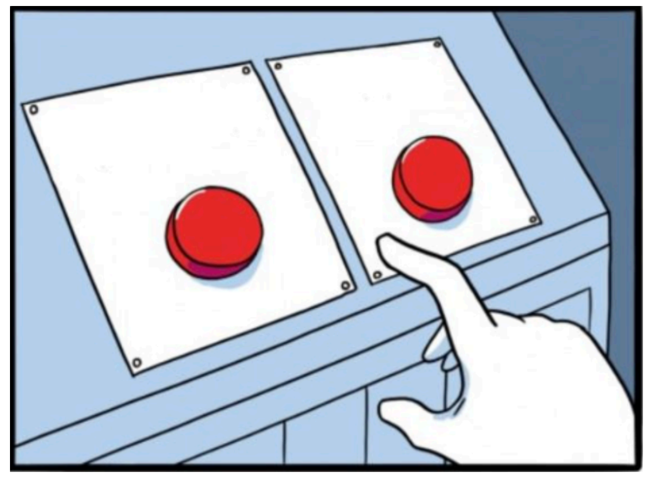
2. Models of action selection (in the brain)

Basic setup: how to agents learn to act?

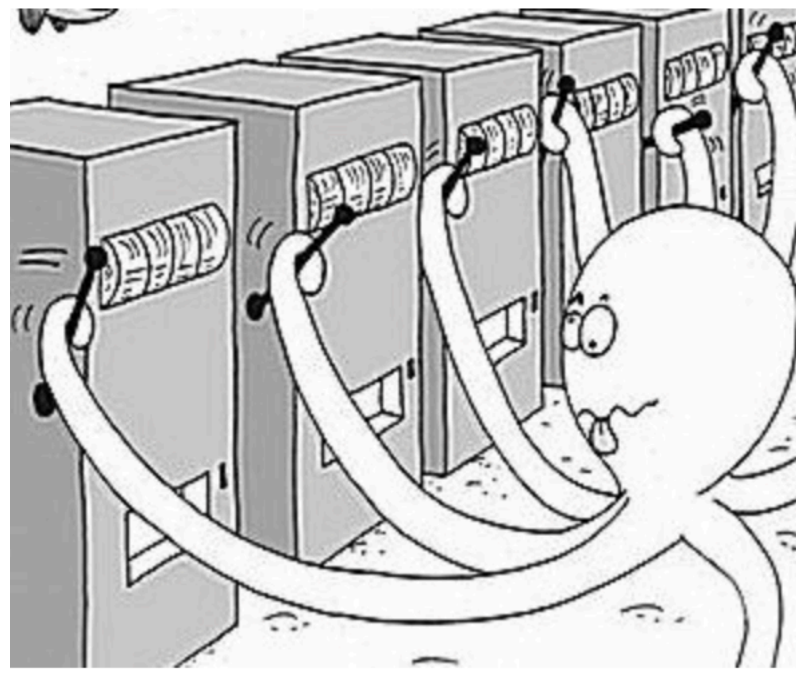
Action is governed by a **policy**:

$$\pi(a, s) = P(a_t = a \mid s_t = s)$$





How to find good actions?

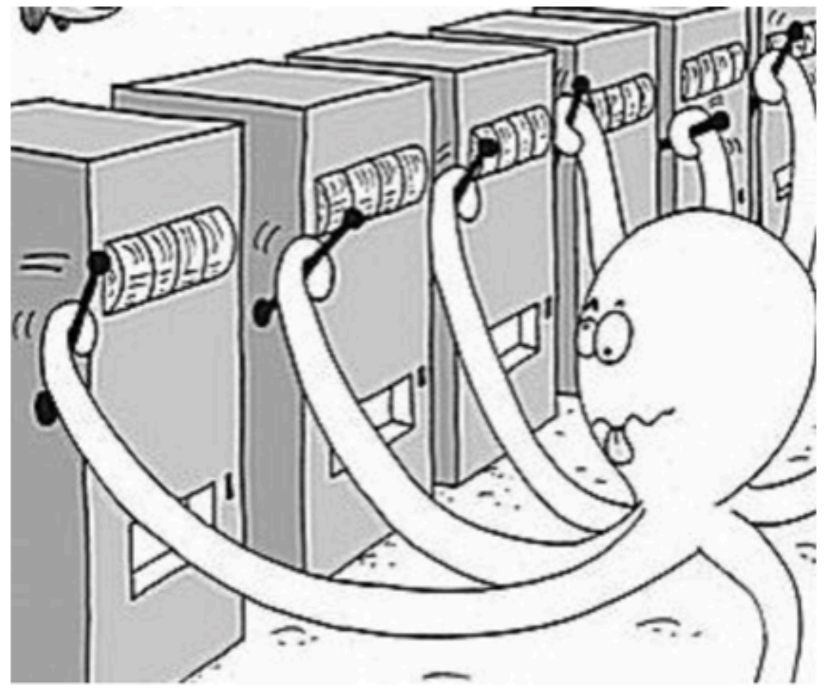


- How do values translate into actions?
- Classic testbed: multi-armed bandits
  - Several options
  - Have to find out which of these are good or bad via trial-and-error
- Key problem: **exploitation vs. exploration**

How to find good actions?

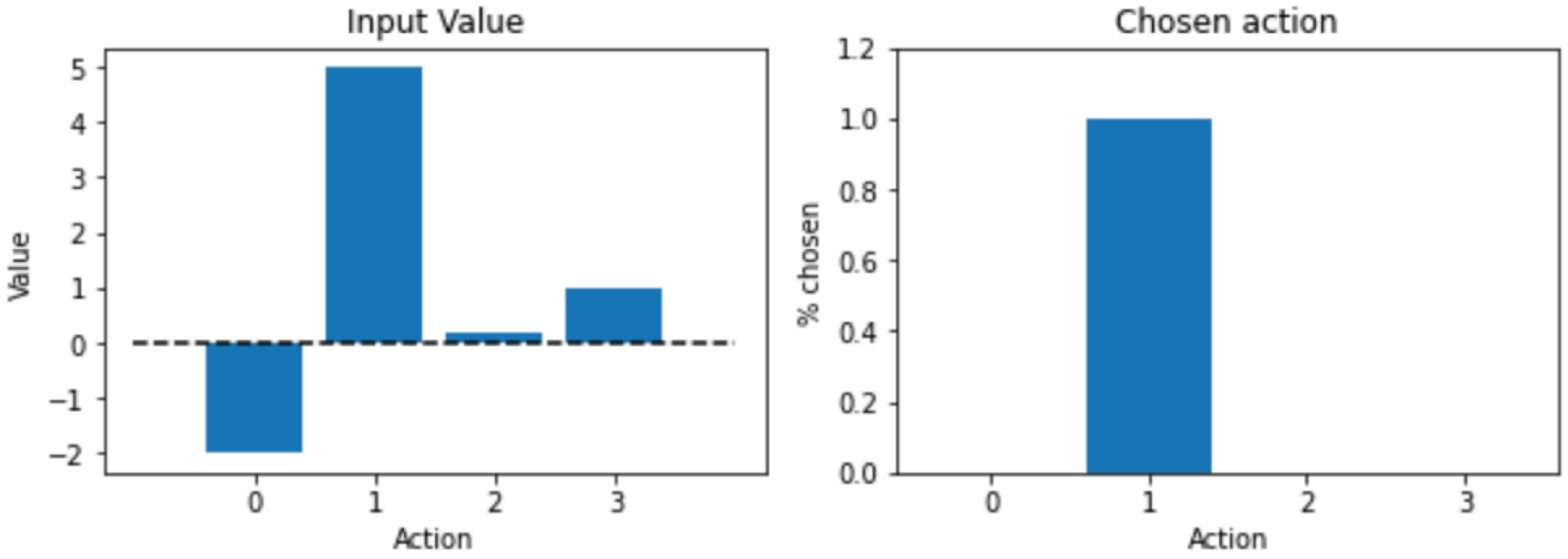
Greedy action selection:

$$P(a_t = a) = \begin{cases} 1 & \text{if } a_t = \text{argmax}_a \ V_t(a) \\ 0 & \text{otherwise} \end{cases}$$



Action is governed by a **policy**:

$$\pi(a, s) = P(a_t = a \,|\, s_t = s)$$

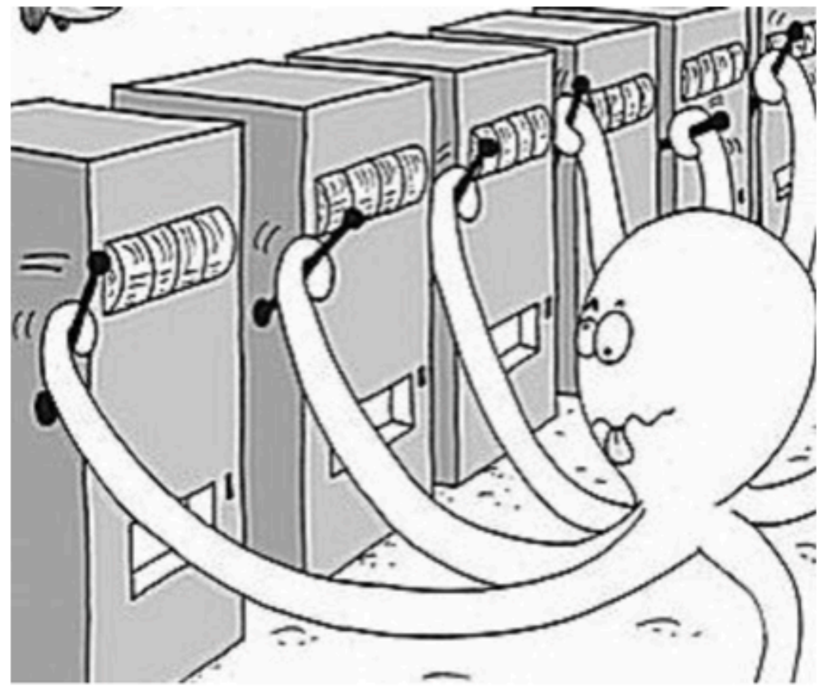


Can you see a problem with this type of behaviour?

How to find good actions?

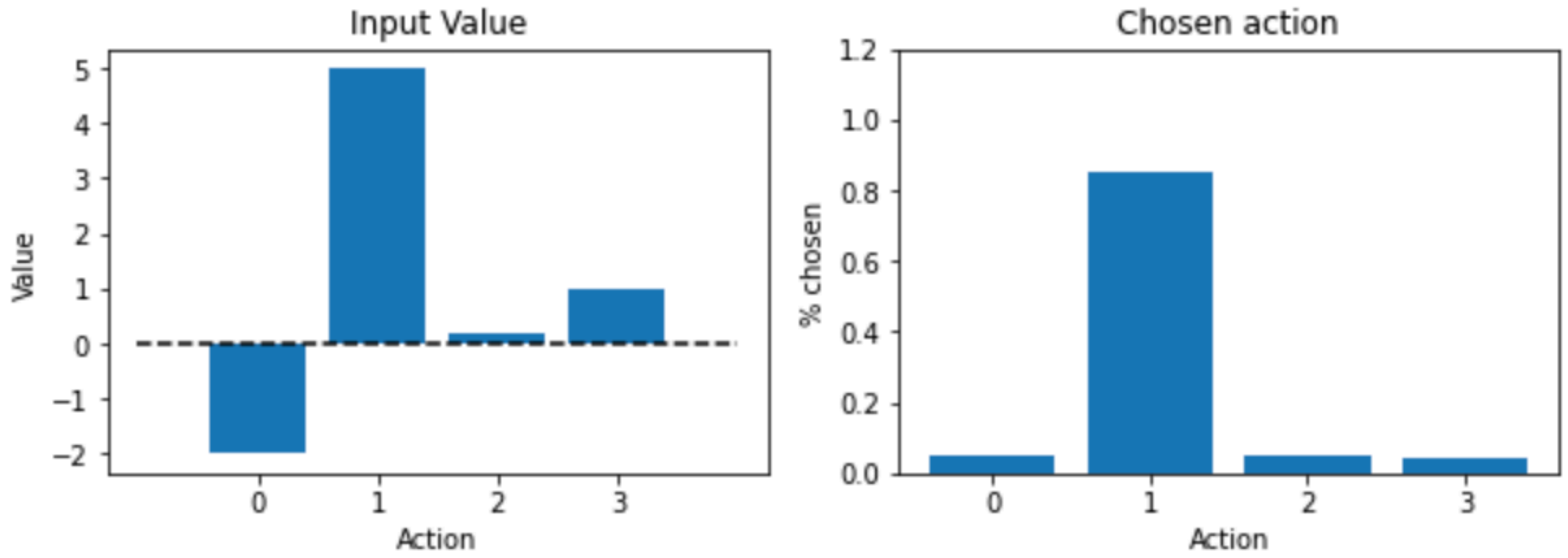
Epsilon-greedy action selection:

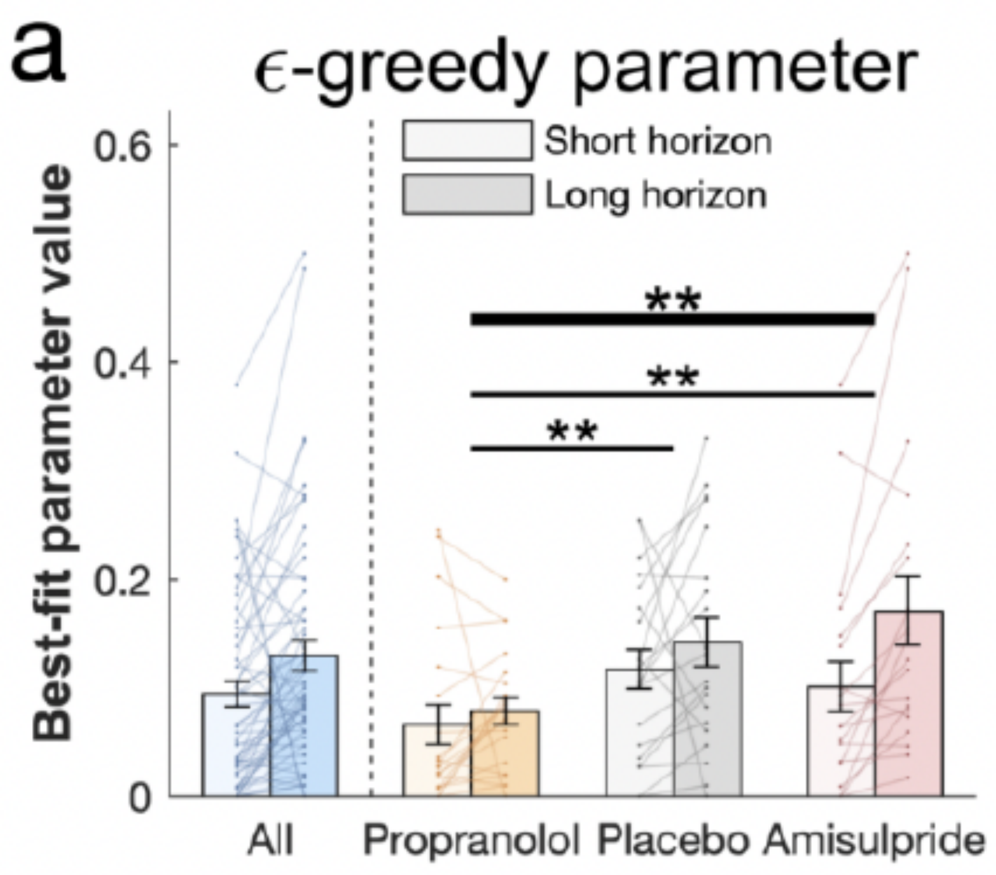
$$P(a_t = a) = \begin{cases} 1 - \epsilon & \text{if } a_t = \text{argmax}_a \, V_t(a) \\ \epsilon / N & \text{otherwise} \end{cases}$$



Action is governed by a **policy**:

$$\pi(a, s) = P(a_t = a \,|\, s_t = s)$$



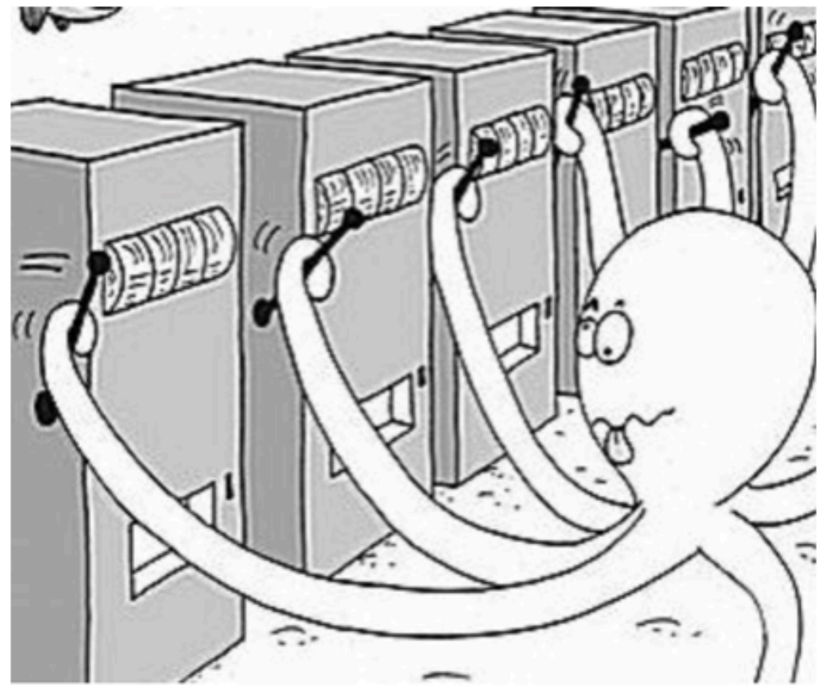


Dubois, … & Hauser, eLife, 2021

How to find good actions?

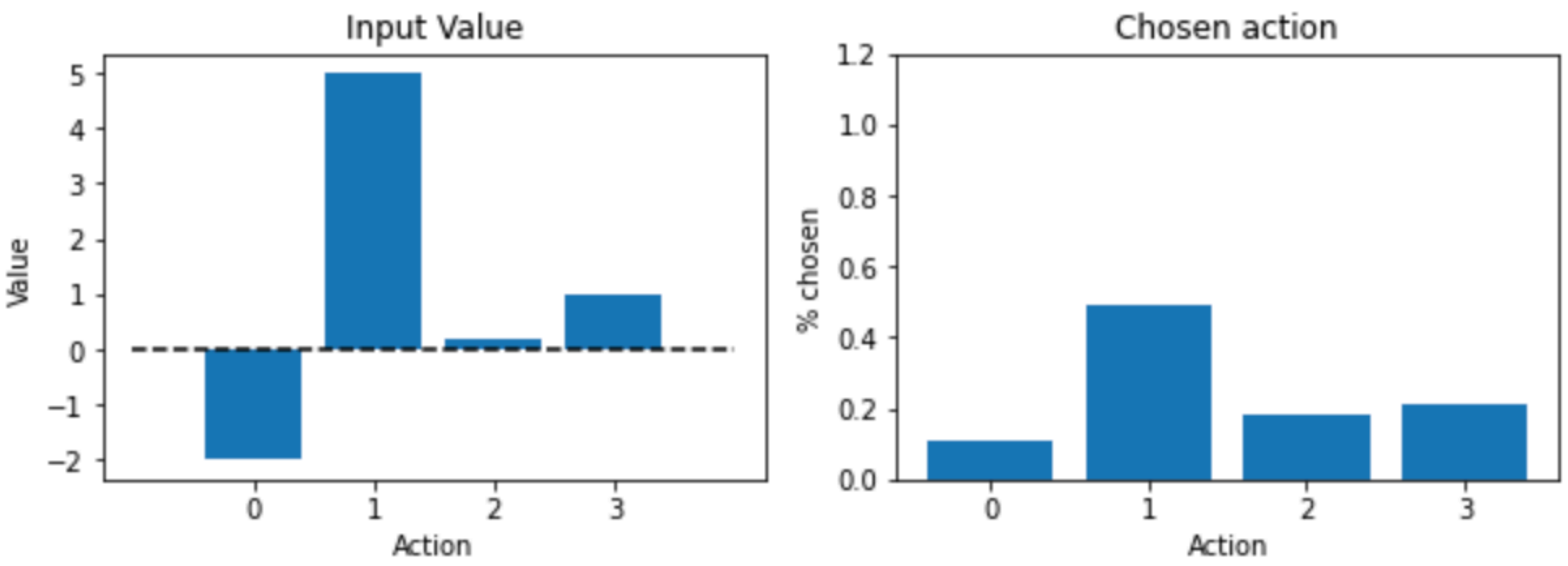
Softmax action selection:

$$P(a_t = a) = \frac{e^{V_t(a)\cdot\beta}}{\sum_{i=1}^{N} e^{V_t(a_i)\cdot\beta}}$$



Action is governed by a **policy**:

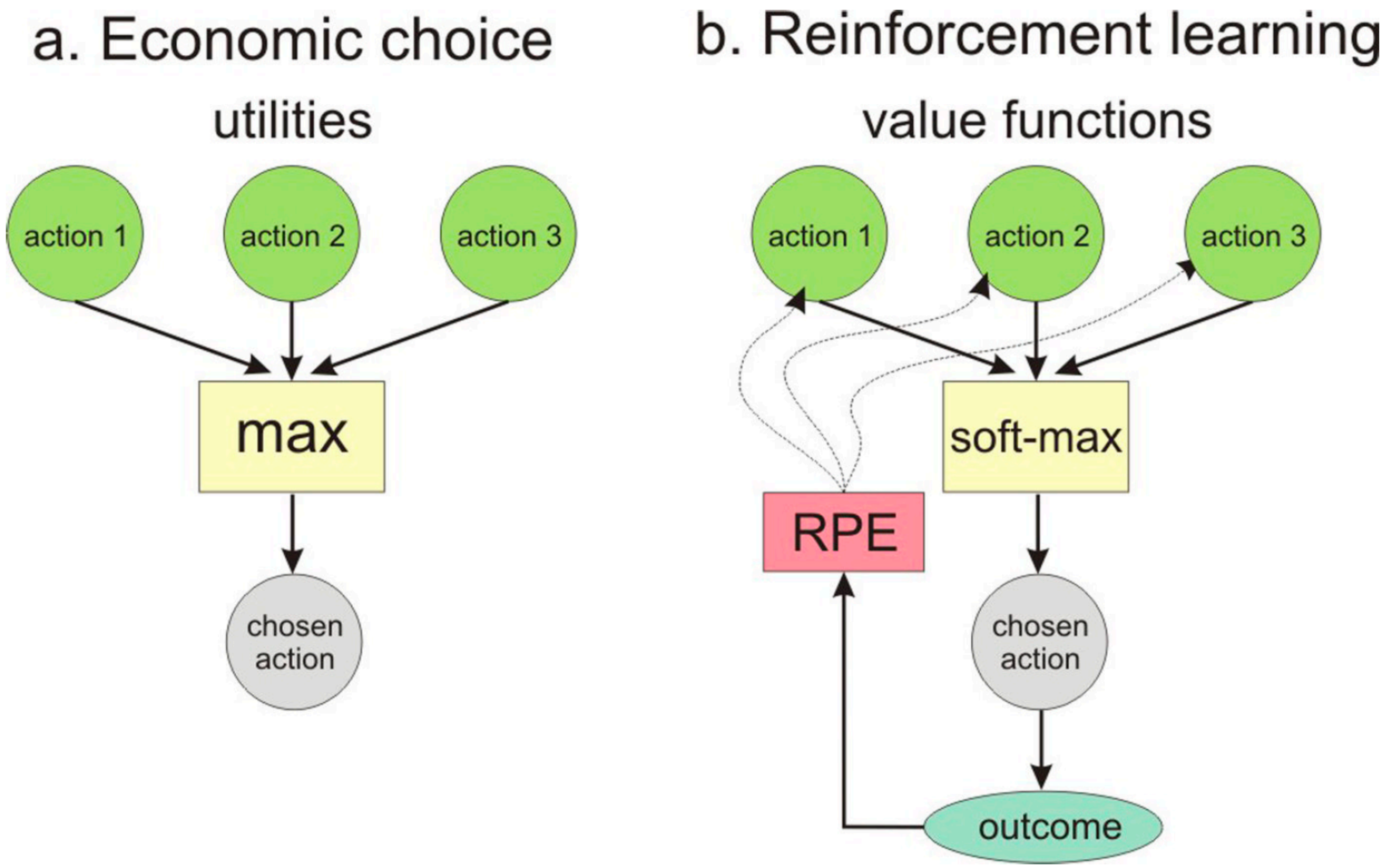
$$\pi(a, s) = P(a_t = a \mid s_t = s)$$



Strongly related to function of neuromodulators (dopamine, norepinephrine)..

How to find good actions?

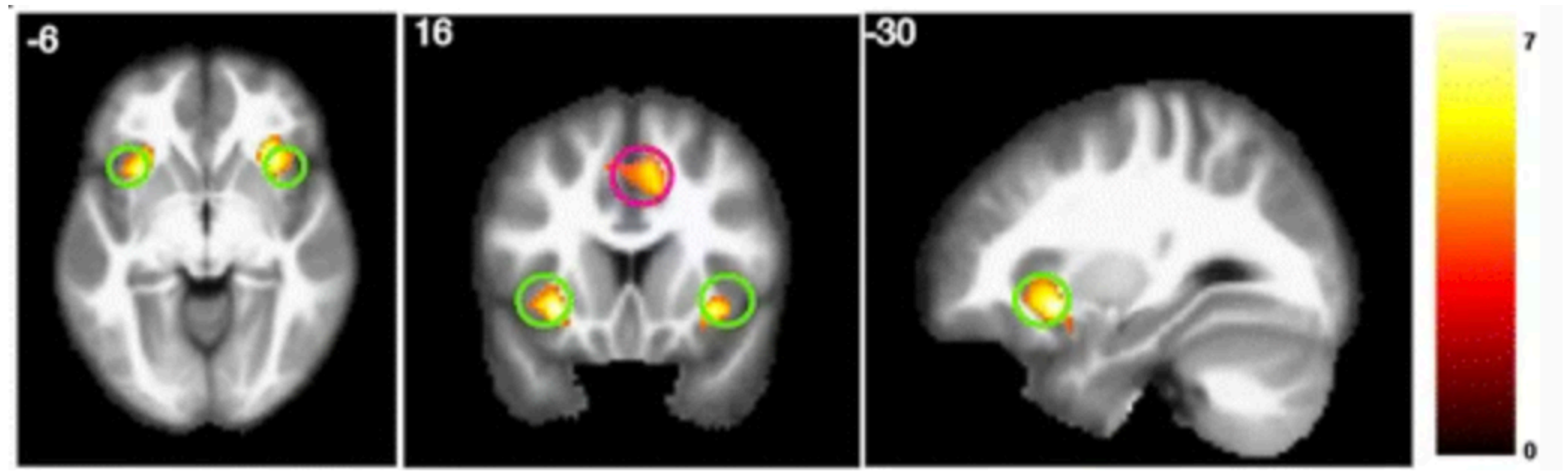
Difference economic choice vs. reinforcement learning (Lee et al., Annu Rev Neurosci 2012):



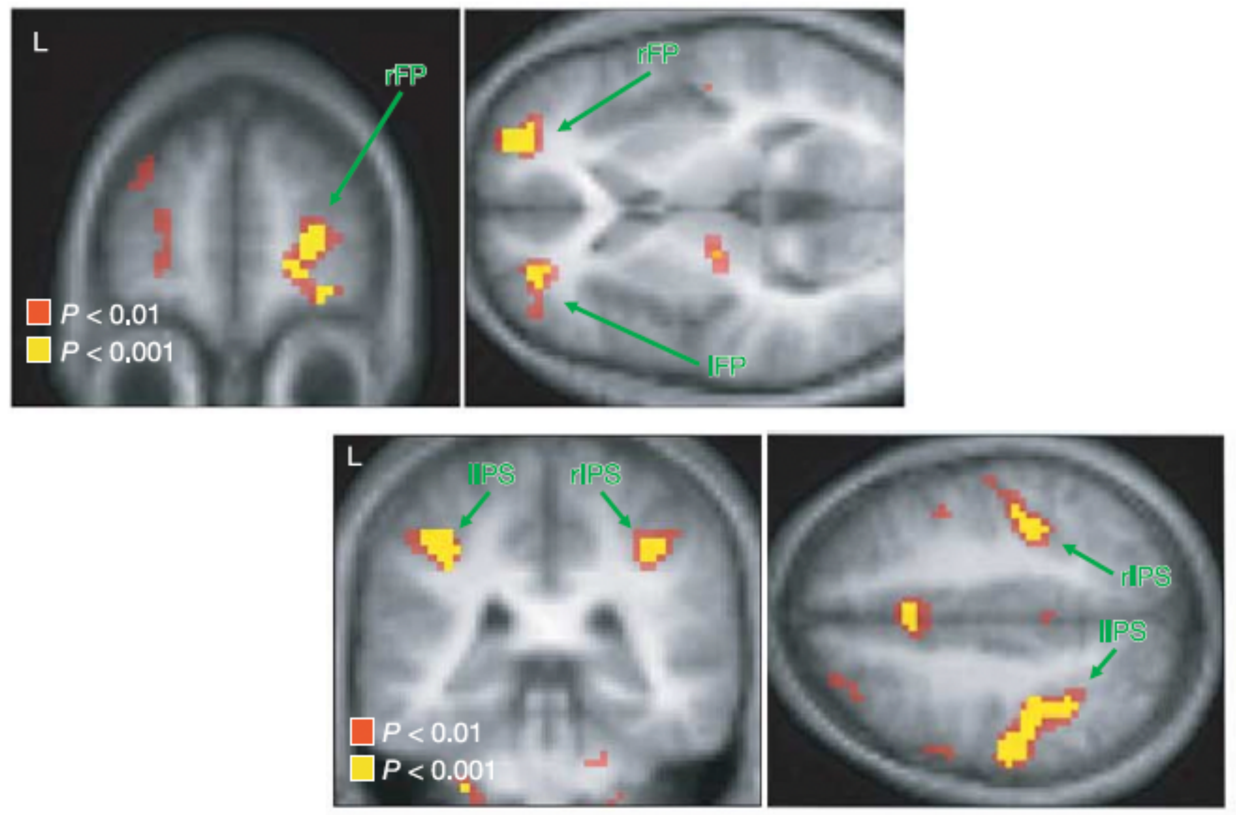
Lee et al., Annu Rev Neurosci. 2012

How to find good actions?

Is there a neural basis for making exploratory decisions?



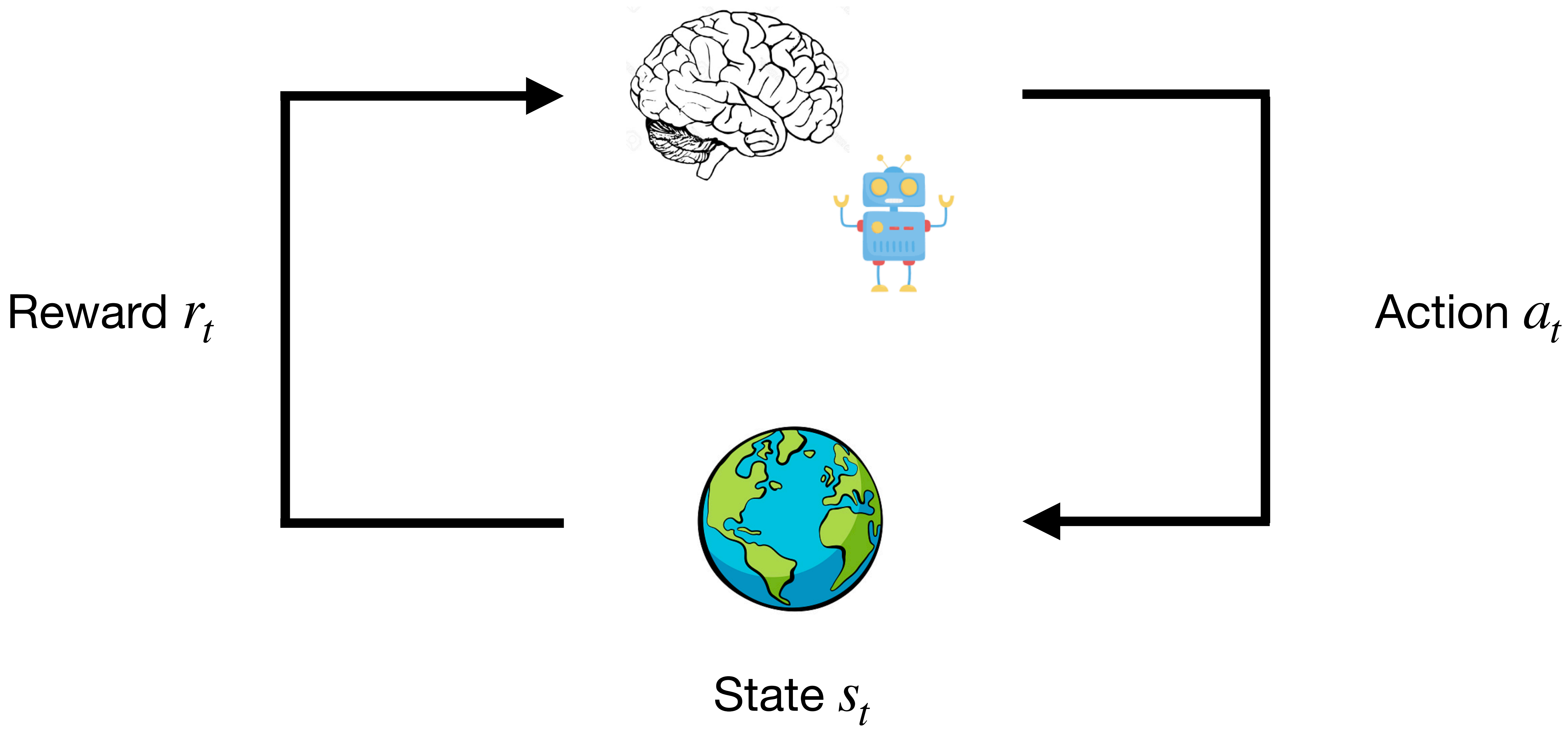
Blanchard & Gershman, Cognitive, Affective & Behavioural Neuroscience, 2018



Daw, … & Dolan, Nature, 2006

3. Model-free vs. model-based RL (in the brain)

Basic setup: how to agents learn to act?



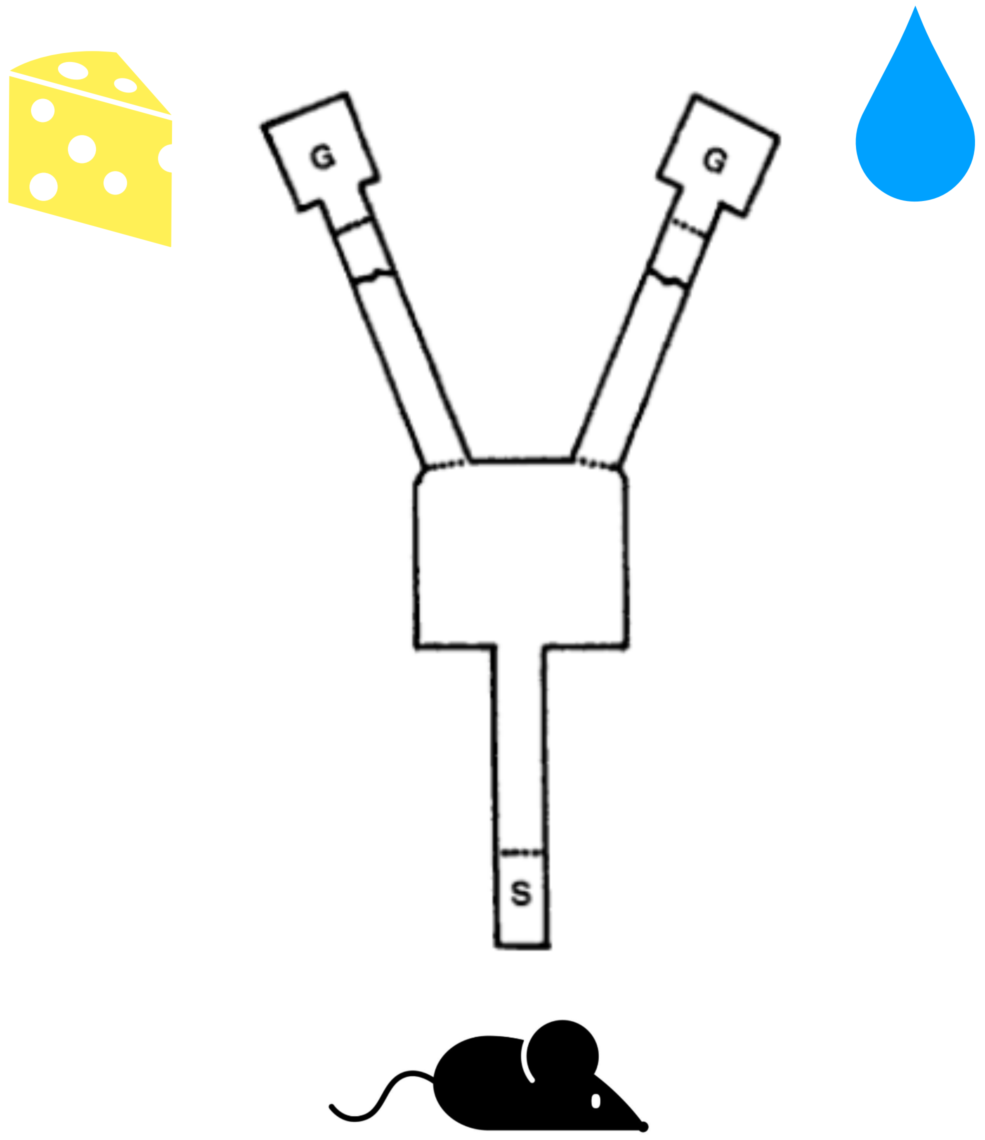
Agents can learn a **model of the environment** to make smarter decisions, e.g.:

$$P(s_{t+1} = s, r_{t+1} = r \,|\, s_t = s, a_t = a)$$

Model-based RL: devaluation

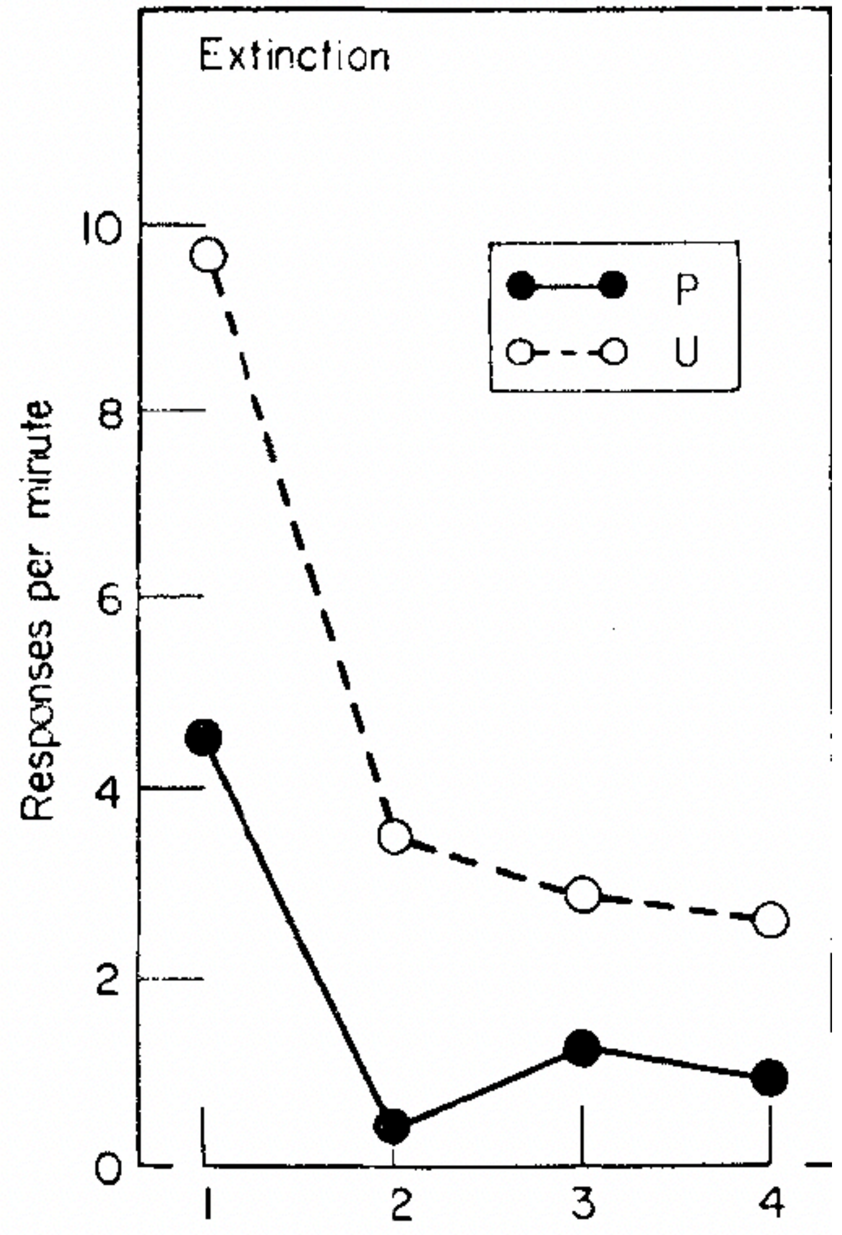
Outcome devaluation (*revaluation*): gold-standard test for forward model predicting outcomes of actions

Animal is trained to perform two different actions, with a different reward:



One reward is then devalued, for example by satiation.

Impact of this devaluation is tested in 'extinction', without providing outcomes.



Adams & Dickinson, Quarterly Journal of Experimental Psychology, 1981
Colwill & Rescorla, Journal of Experimental Psychology, 1985
Akam, Costa, & Dayan, PLOS CB 2015

MDPs basis for model-based RL

$$P(s', r | s, a) = P(s_{t+1} = s', r_{t+1} = r | s_t = s, a_t = a)$$

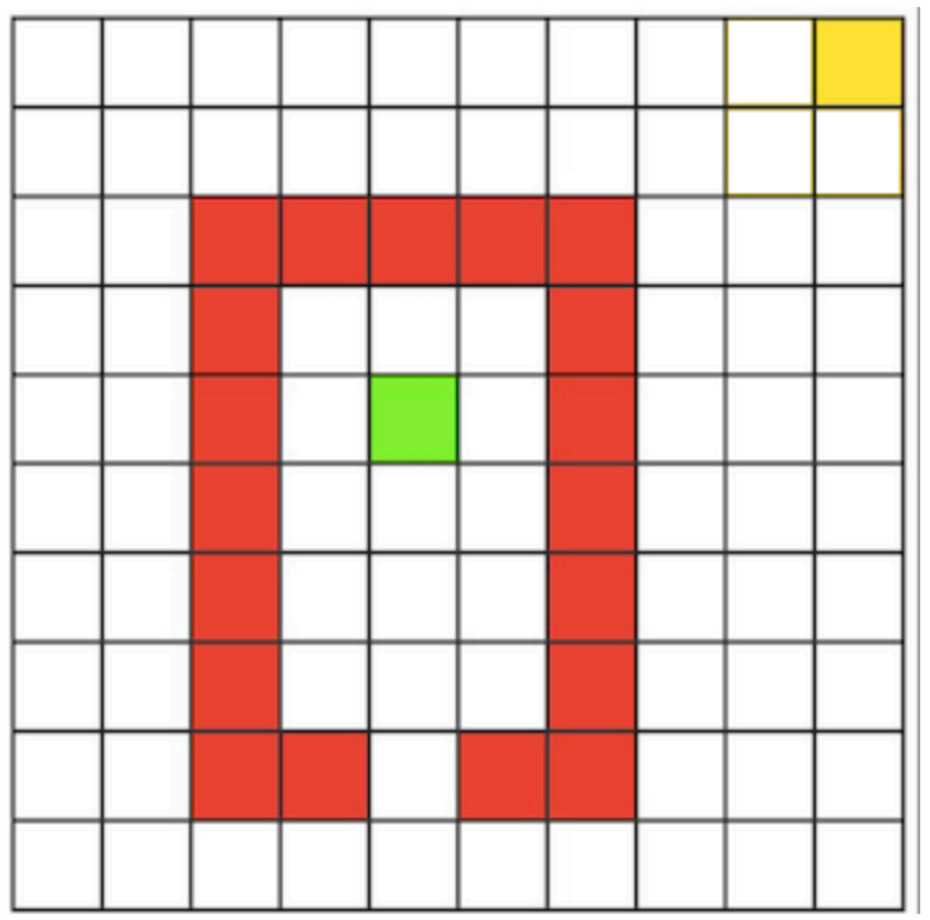
How can we make use of such models of the world?

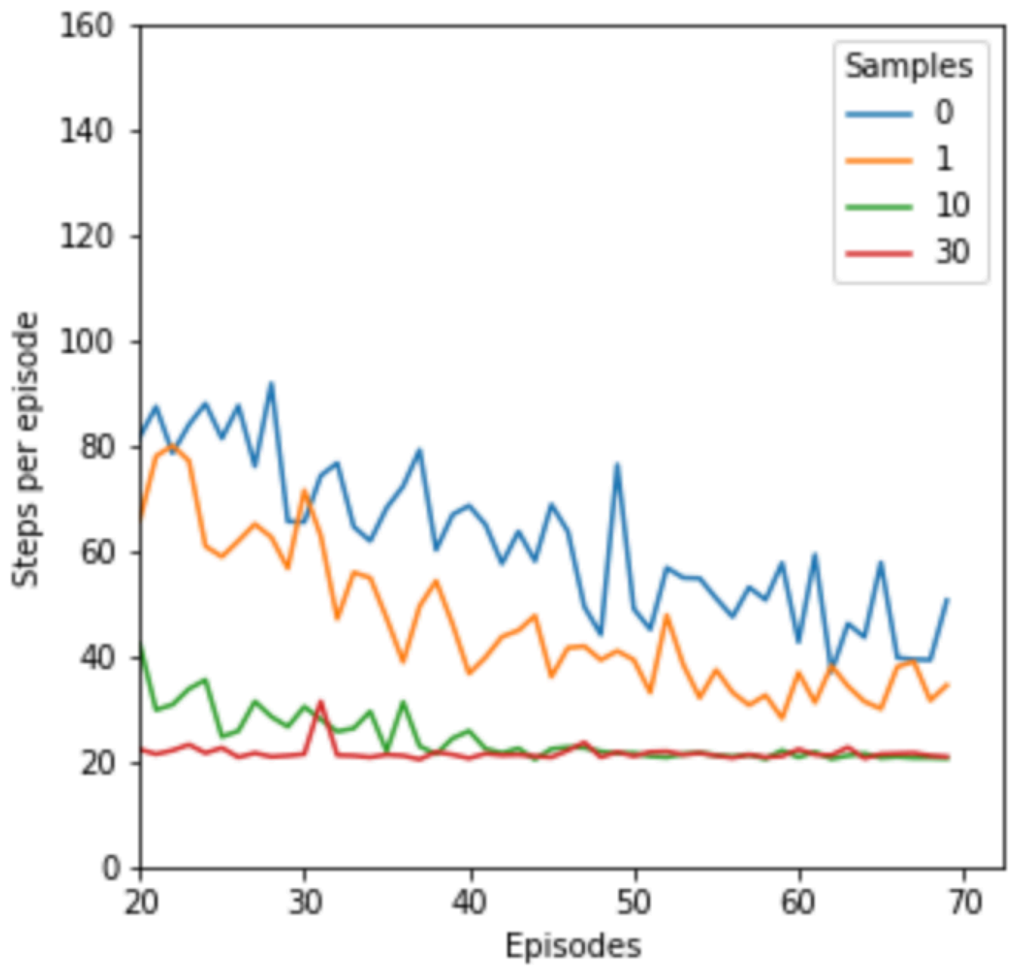
Learning

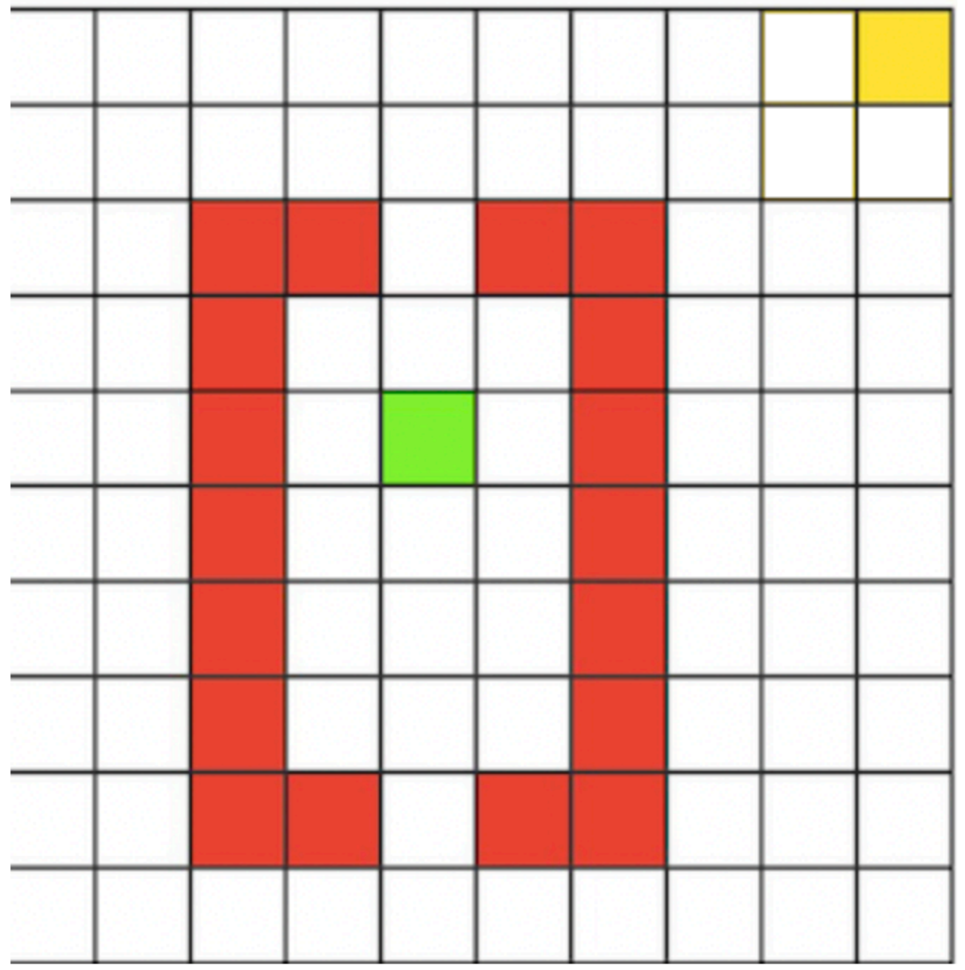
- Key idea: store experiences in world model $P(s', r | s, a)$
- Sample from this model to generate extra learning data
- This is called **DYNA-Q…**

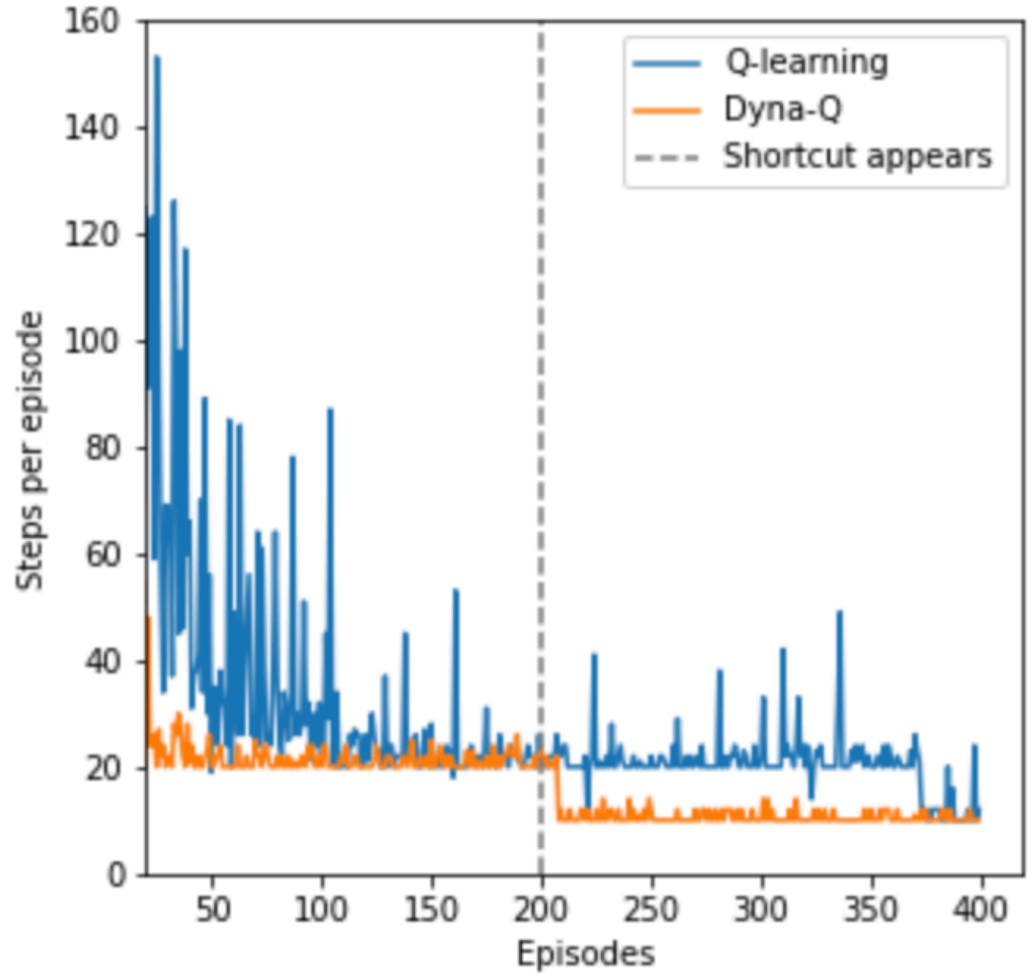
DYNA-Q

Sample from world model $P(s', r \mid s, a)$ to generate extra learning data









And during breaks ('offline rest'), they can sample from this experience and learn from these samples:

$S \leftarrow$ previously observed state

$A \leftarrow$ action previously taken in S

$R, S' \leftarrow Model(S, A)$

$Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma \max_a Q(S', A) - Q(S, A)\right]$

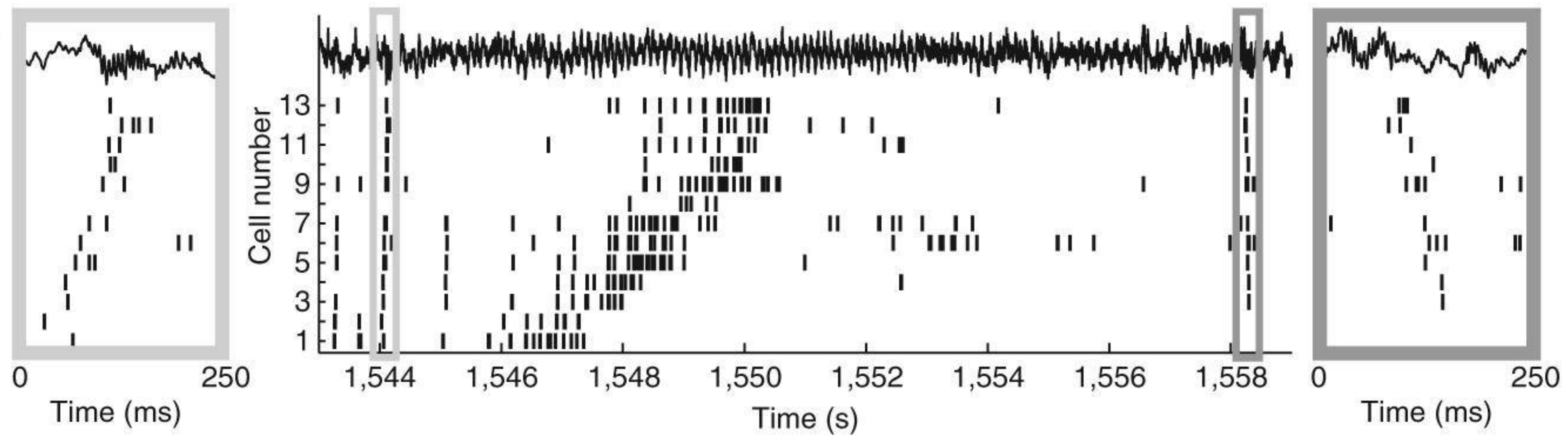
Link to code here

DYNA-Q - Replay as a candidate neural mechanism

DYNA-Q looks a lot like replay.

Replay as a computational mechanism in PFC and hippocampal formation

- i.e. fast reactivation of external states



Diba & Buzsaki (2007) Nature Neuroscience

Implicated in

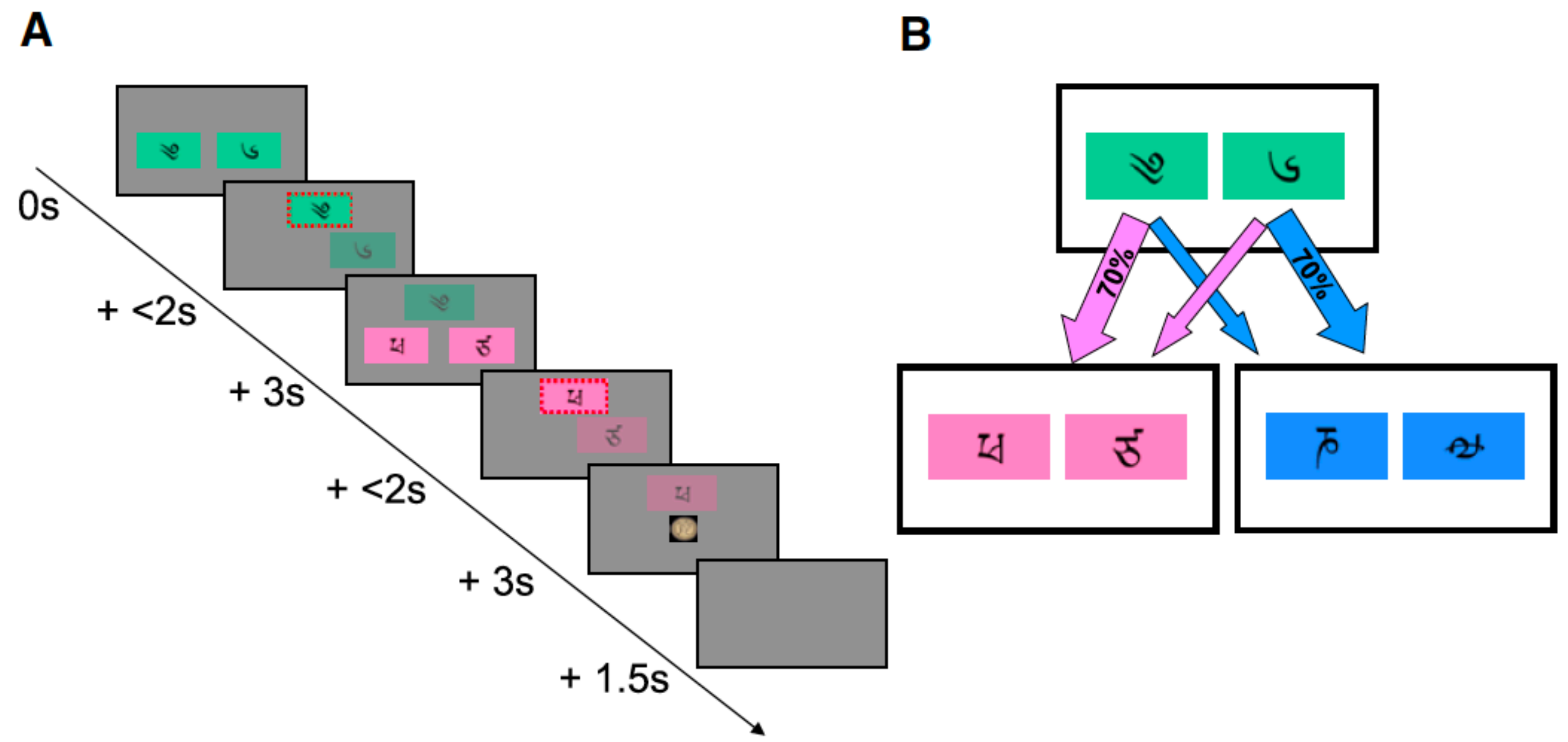
- Learning from the *past* (credit assignment, Ambrose et al. (2016) Neuron)
- Planning *future* trajectories (Pfeiffer & Foster (2013) Nature)

MDPs basis for model-based RL

$$P(s', r \mid s, a) = P(s_{t+1} = s', r_{t+1} = r \mid s_t = s, a_t = a)$$

How can we make use of such models of the world?

Planning and **action selection**

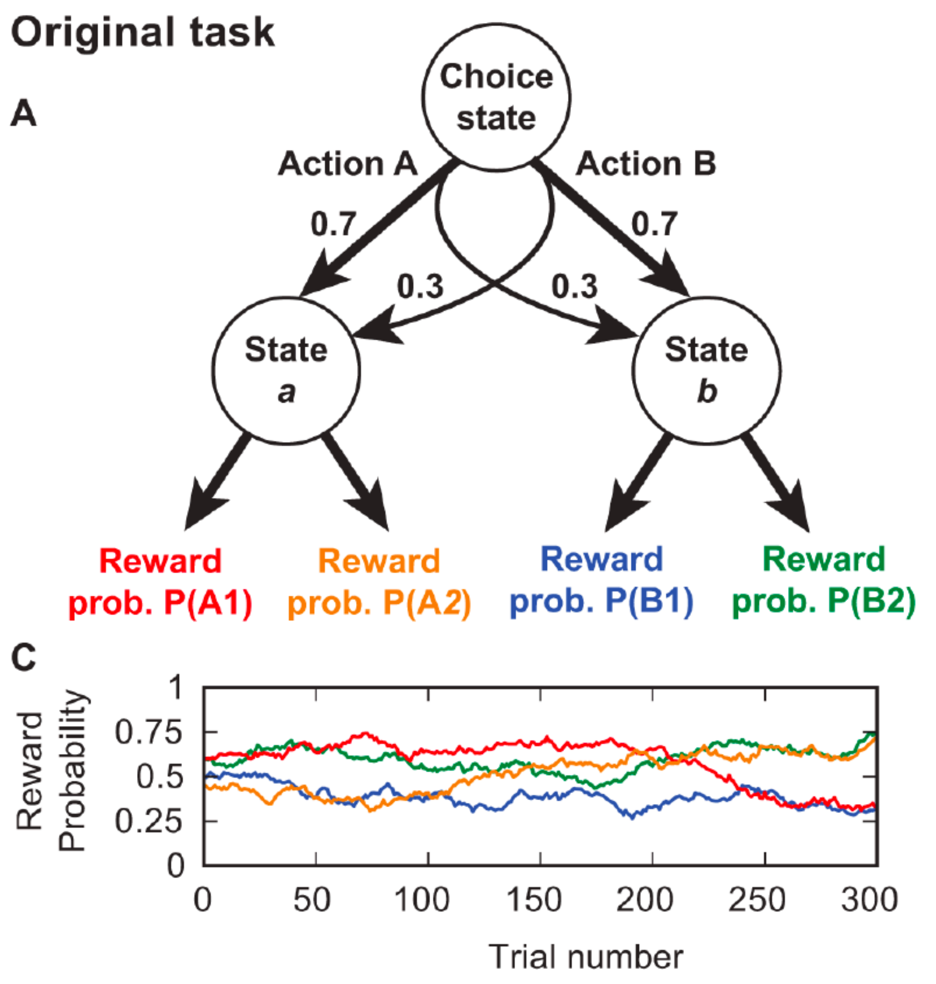


Daw, …, Dolan, Neuron, 2011

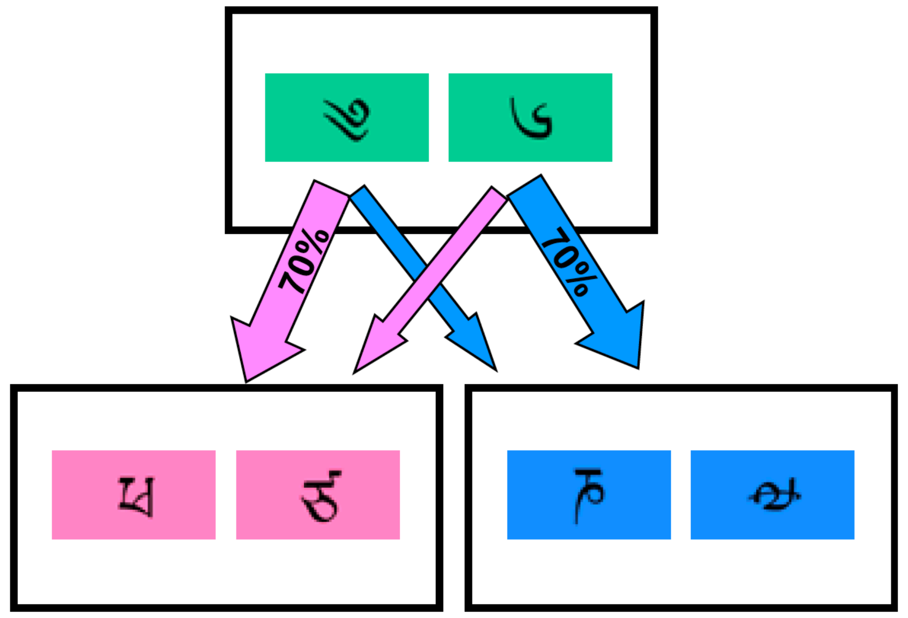
'Two-step task'

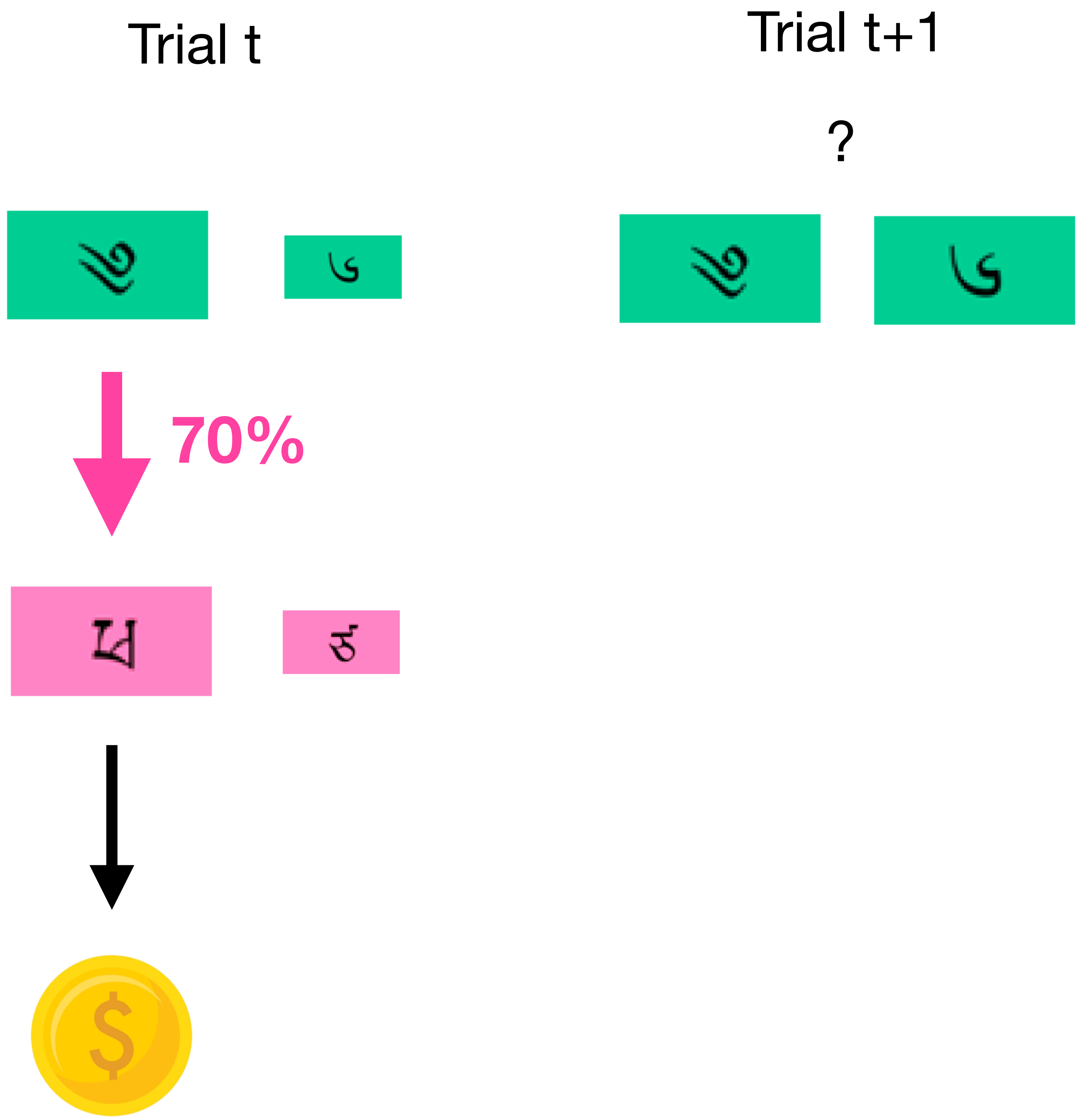
Key manipulation: **common** and **rare** transitions

Two-step task: one of the most iconic RL tasks



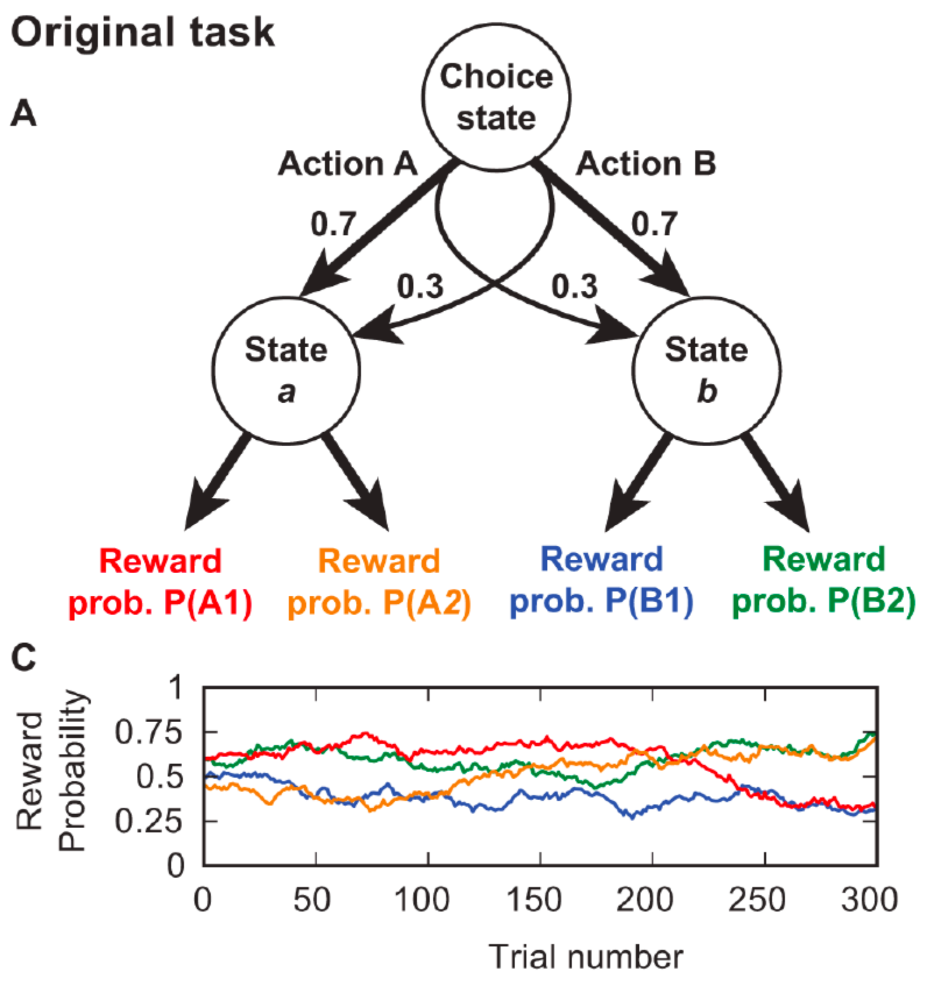
Akam, Costa, Dayan,
PLOS Computational Biology, 2015



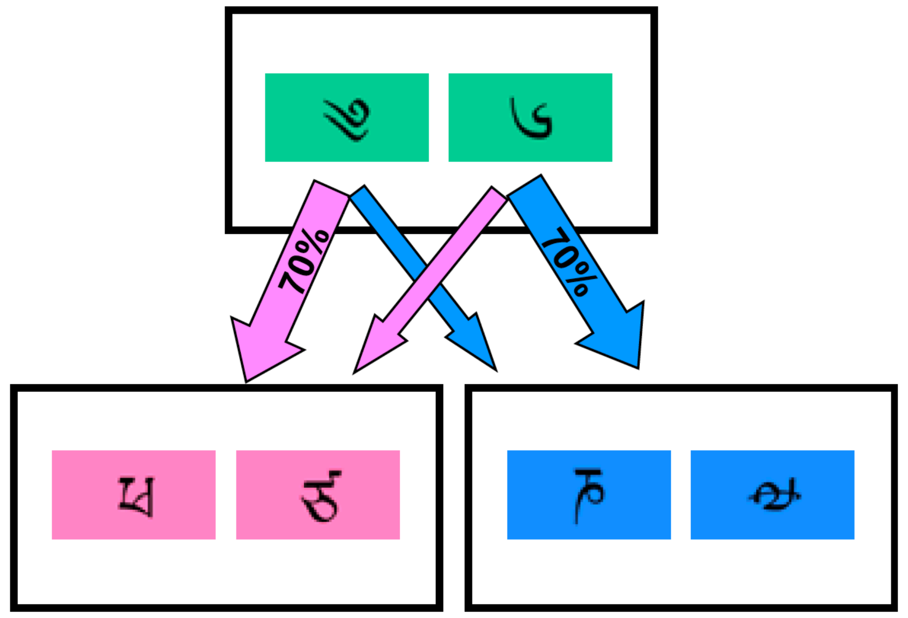


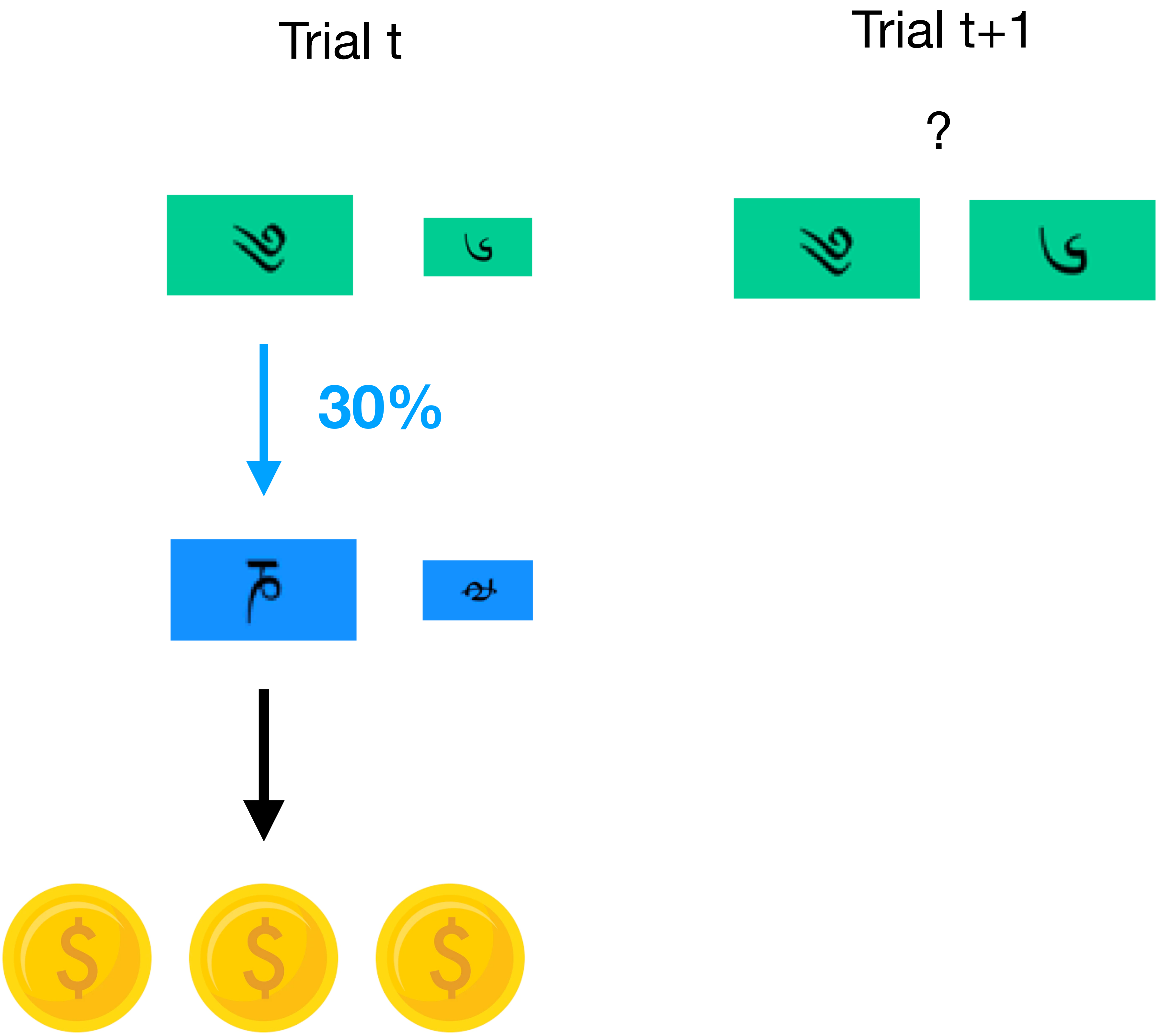
Which green option should the agent choose again at trial t+1?

Two-step task: one of the most iconic RL tasks



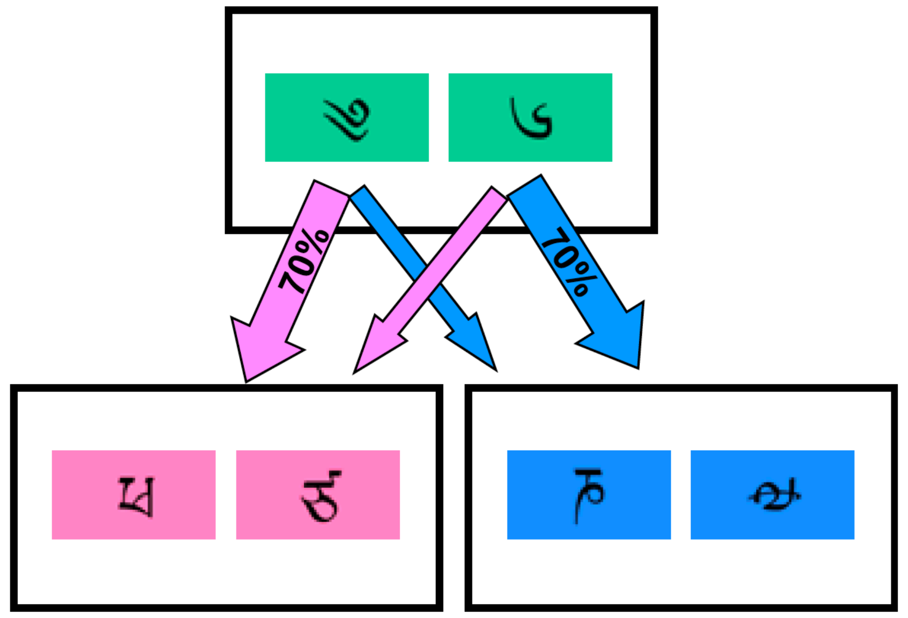
Akam, Costa, Dayan,
PLOS Computational Biology, 2015





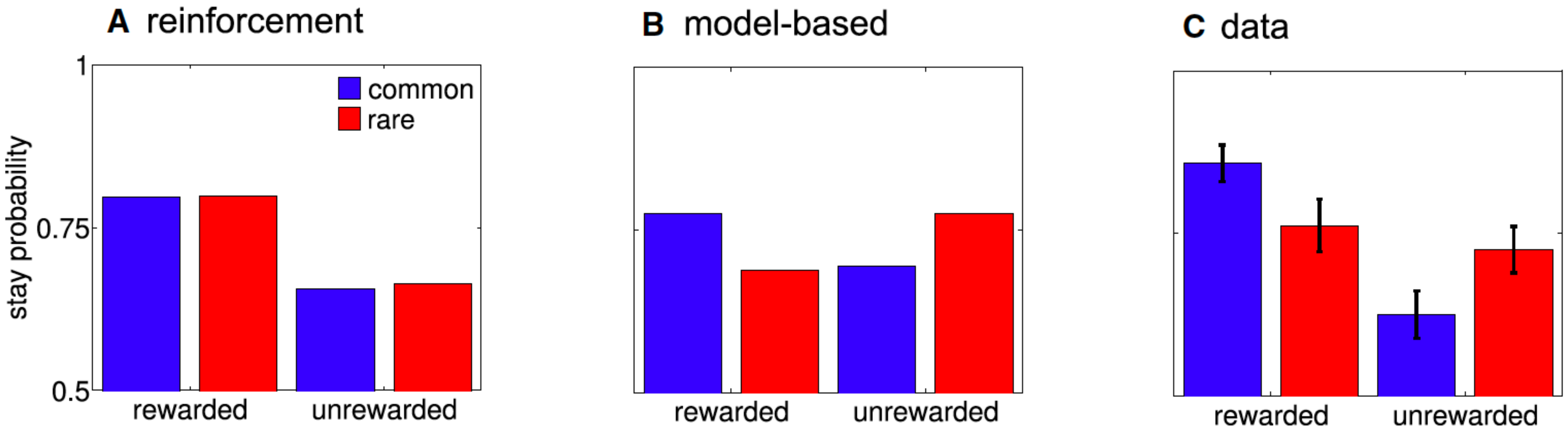
Which green option should the agent choose again at trial t+1?

Two-step task: one of the most iconic RL tasks



Trial t+1

?



Model-free RL agent: repeat what is rewarding

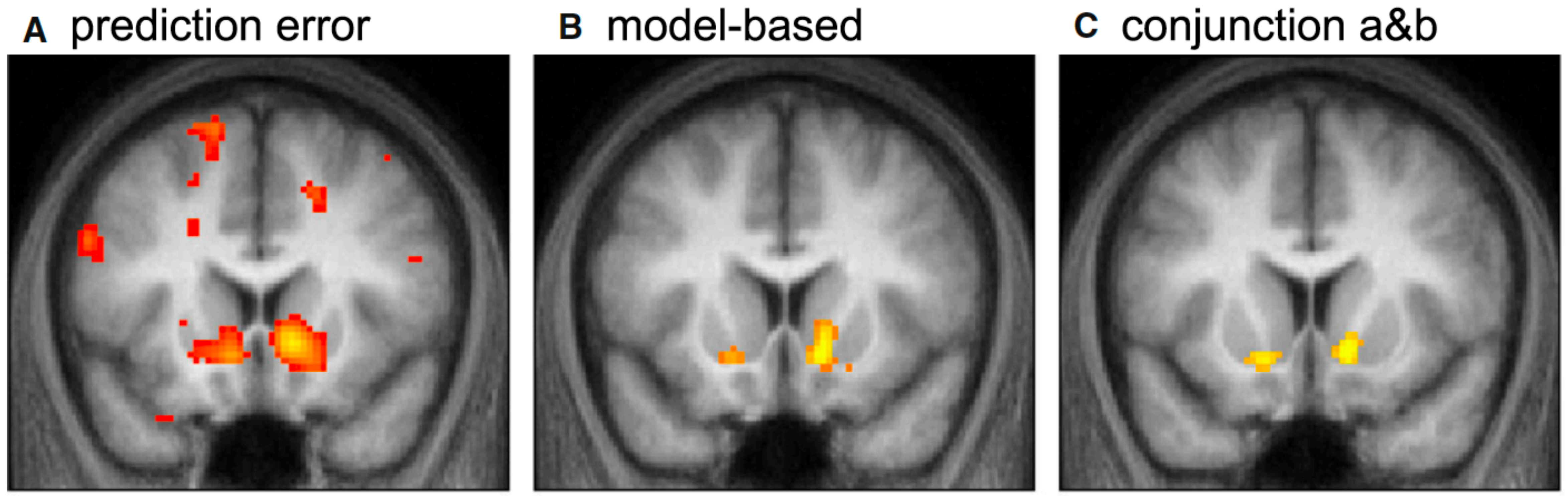
Model-based RL agent: repeat what is rewarding, but be clever

Really data: a mix of both

Link to code <u>here</u>

Two-step task: one of the most iconic RL tasks

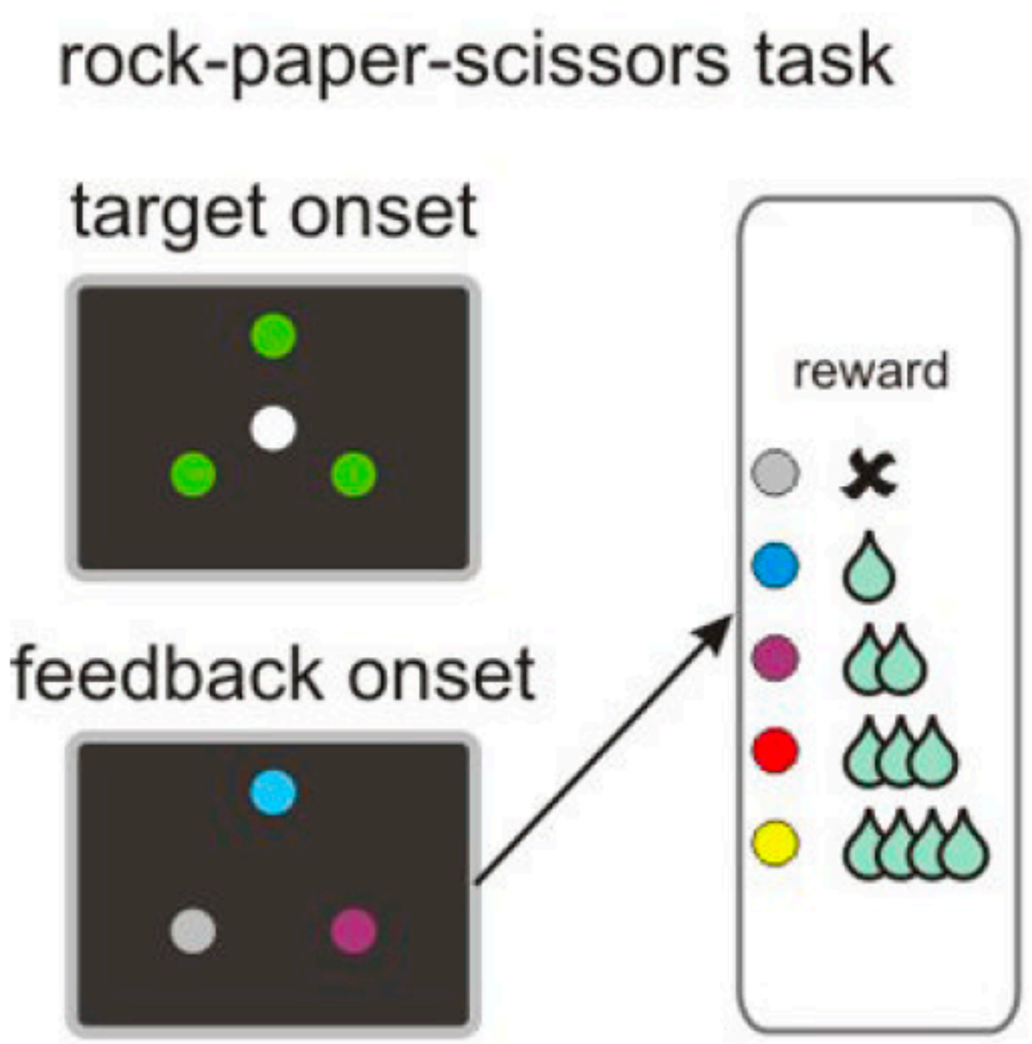
Model-free and model-based prediction errors in ventral striatum

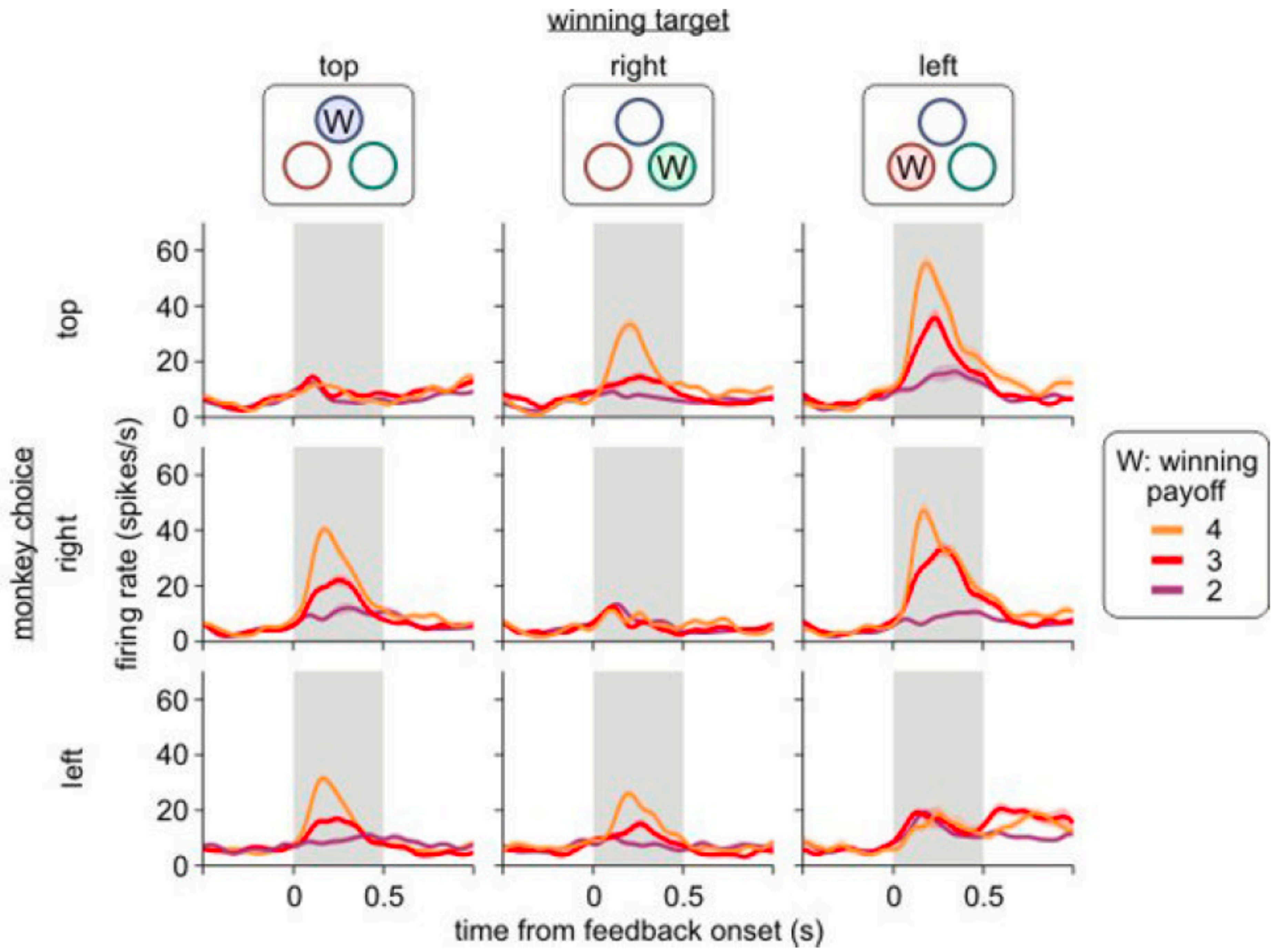


Model-based reasoning: counterfactuals

Some neurons in orbitofrontal cortex encode hypothetical outcomes:

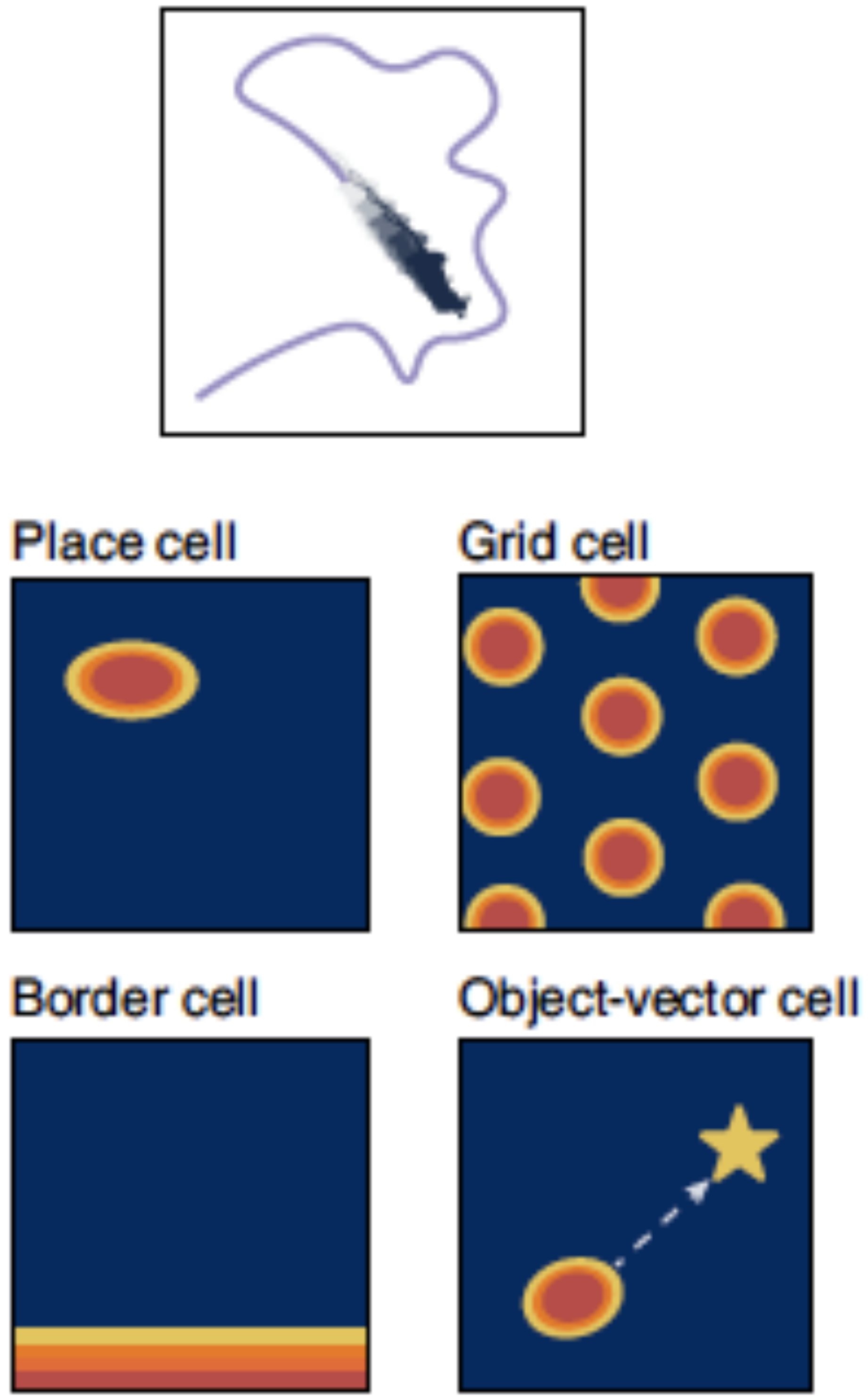
- Fire only if an *unchosen* option was rewarded

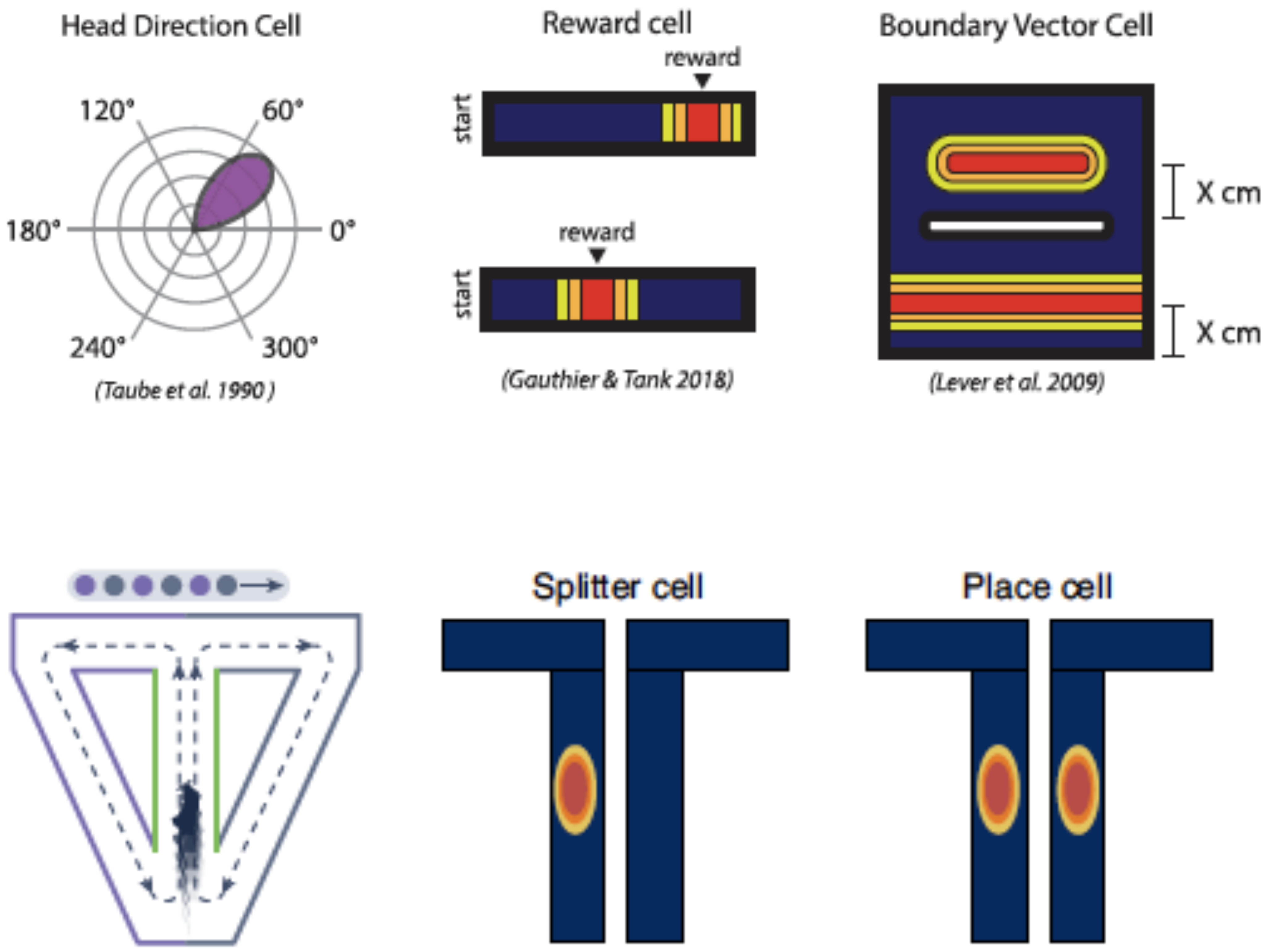




Lee et al., Annu Rev Neurosci. 2012

What is the model in model-based RL?





Is this a **basis set** over world structures?

Whittington et al. (2022). How to build a cognitive map. Nature Neuroscience

Behrens et al. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. Neuron

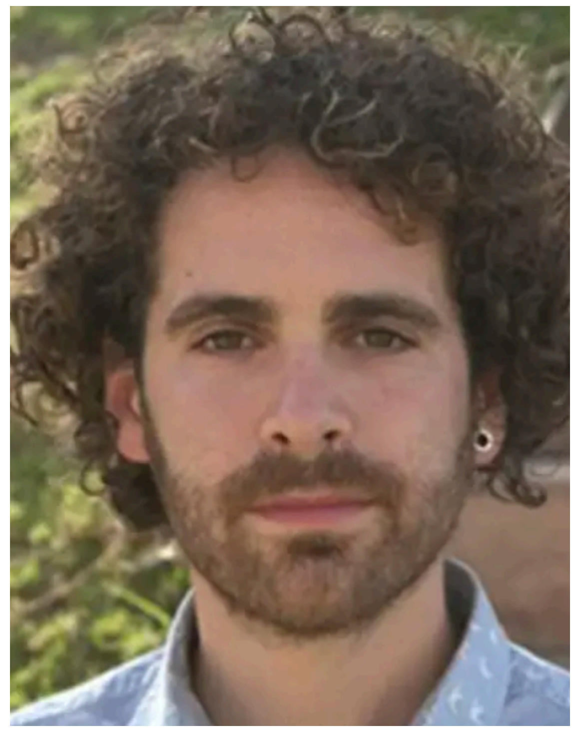
Discussion questions

- Is reward enough? Can you think of limits of RL?
- How are cognitive maps useful in RL?
- Can you think of situations where cognitive maps are useful that are not in a RL context?
- If you were a scientist, what experiment on RL (and perhaps cognitive maps) would you want to run?

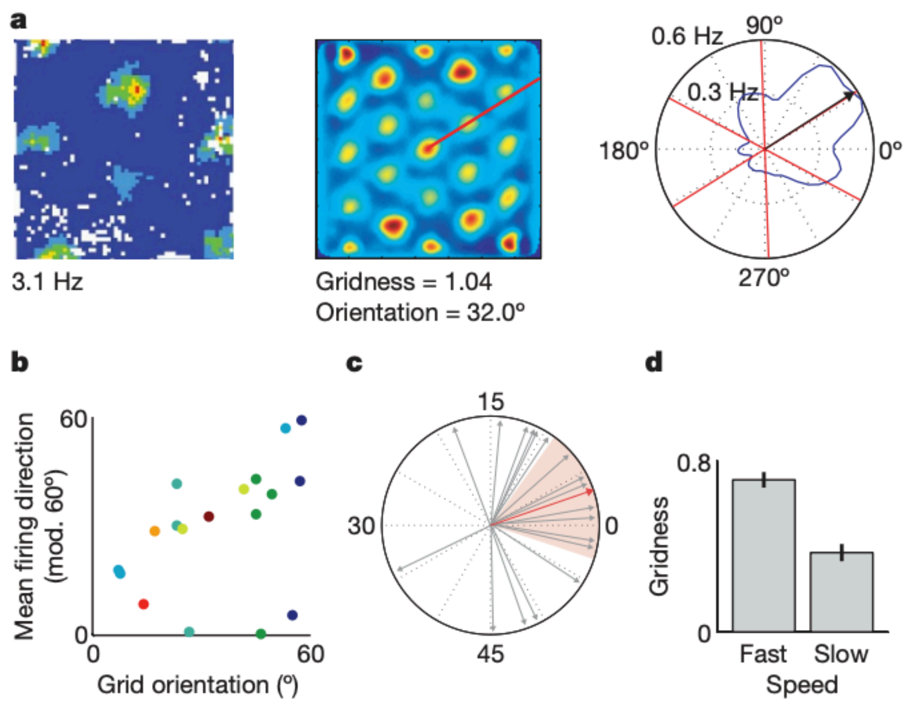
Next week

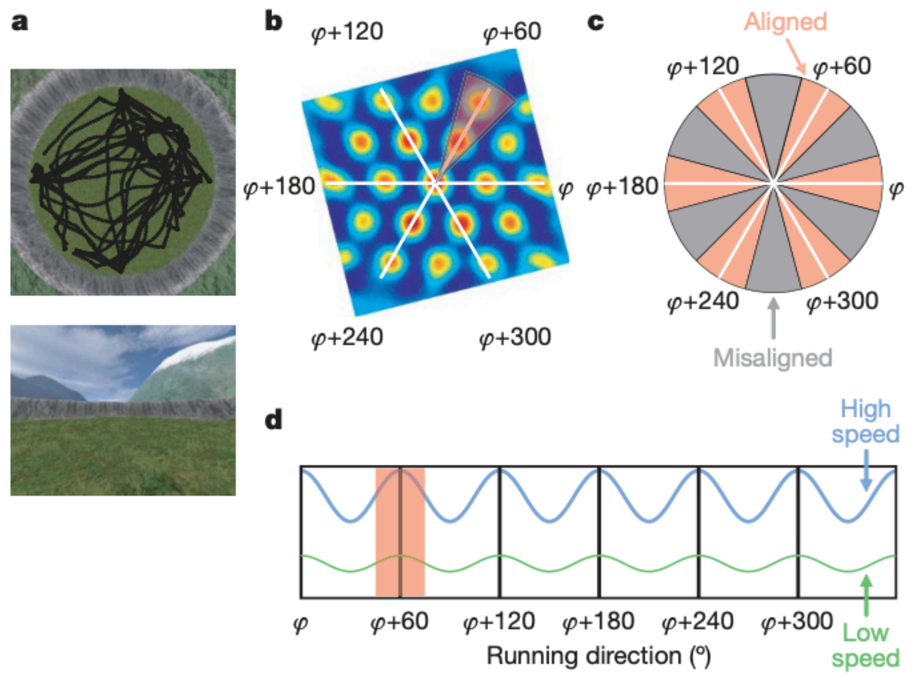
Evidence for grid cells in a human memory network

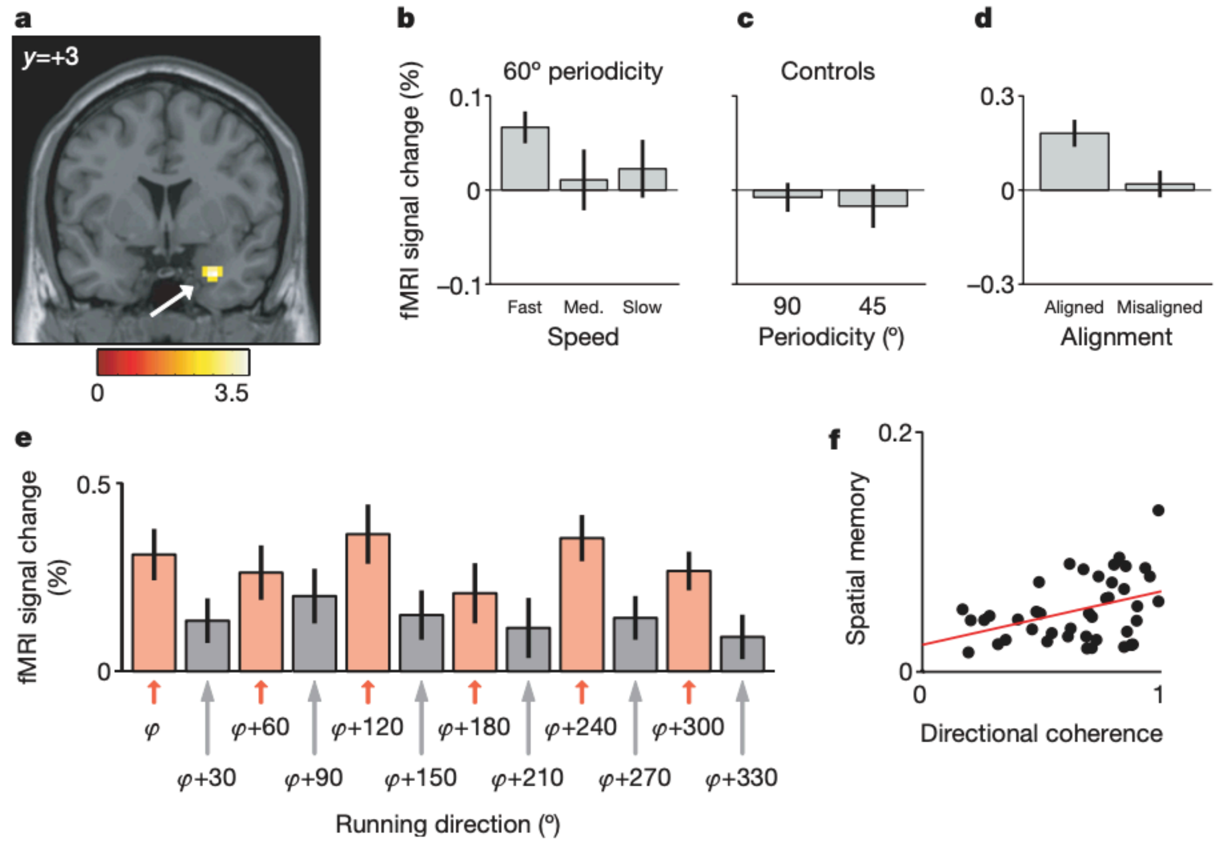
Christian F. Doeller[1,2], Caswell Barry[1,3,4] & Neil Burgess[1,2]



Nir Moneta
PhD student, Max Planck UCL Centre for Computational Psychiatry and Ageing Research







YOUR task:

- Read the paper
- Submit a question AND YOUR NAME <u>here</u>