

DeepMind

REINFORCEMENT LEARNING

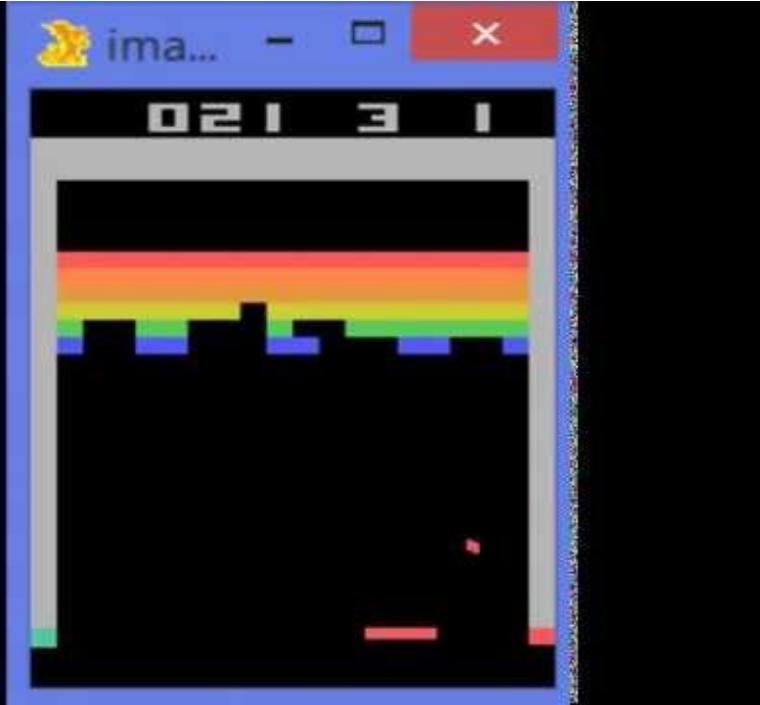
Computational Modeling for Learning and Decision Making

Maria K. Eckstein
mariaeckstein@deepmind.com

BAMB summer school, 21/07/2023



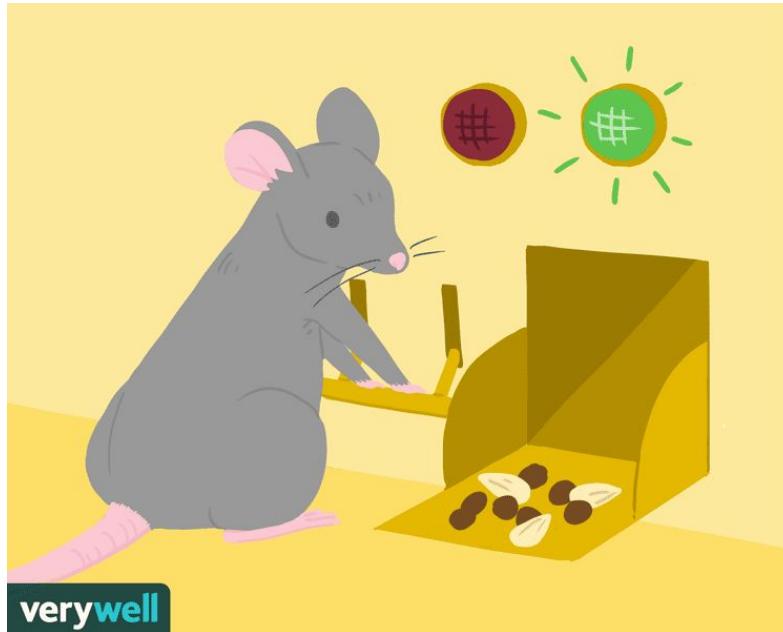
Reinforcement Learning (RL)



-> What do both videos have in common?

What is RL?

Learning from rewards;



and punishment.



How to Use RL (as a Cognitive Model)?

Goal



Reward

+1

Ingredients

action = [→, ←]

state = []

reward = [0, +1]

Algorithm

$$Q(s,a) \leftarrow Q(s,a) + \alpha \text{RPE}$$

$$\text{RPE} = r + \gamma Q(s',a') - Q(s,a)$$



action = [jump, stand]

state = []

reward = [0,]

???

Questions?

Confidential - DeepMind



DeepMind

Lecture Roadmap



Reinforcement Learning (RL)

Confidential - DeepMind

1. **Introduction**
2. RL from a psychology perspective
3. RL from an AI perspective
4. RL from a neuroscience perspective
5. Bringing it all together: RL as a cognitive model
6. Conclusion



Reinforcement Learning (RL)

Confidential - DeepMind

1. Introduction
2. **RL from a psychology perspective**
3. RL from an AI perspective
4. RL from a neuroscience perspective
5. Bringing it all together: RL as a cognitive model
6. Conclusion

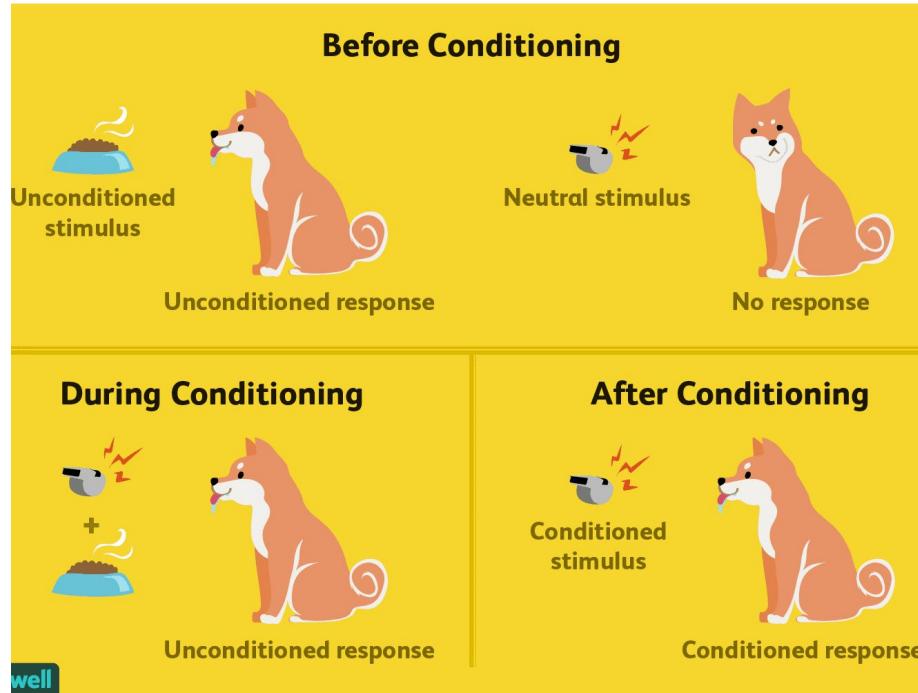
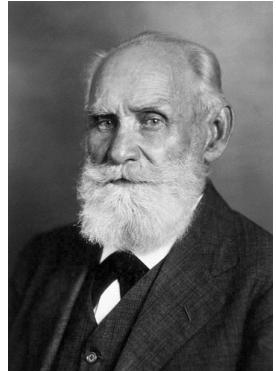


DeepMind

RL from a psychology perspective



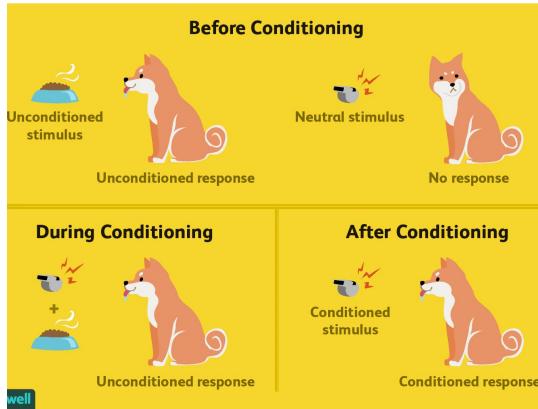
Classical Conditioning



Animals learn associations between US (e.g., food) and neutral CS (e.g., bell) when they reliably co-occur.



The Rescorla-Wagner Model (1972)



$$\boxed{\text{RPE} = \text{reward} - \sum [\text{value}(CS)]}$$

$$\text{value}(CS) \leftarrow \underbrace{\text{value}(CS)}_{\text{Old value (before learning)}} + \underbrace{\alpha_{CS} * \beta_{US} * \text{RPE}}_{\text{New value (after learning)}}$$

Combined predictive value of all stimuli

- Stimuli (CS) have “associative strength” (**value**)
 - Does the stimulus predict a US (**reward**)?
- When reward arrives, there might a “reward prediction error” (RPE)
 - Was the reward predicted by the present stimuli?
- RPEs trigger learning: update values to predict reward better
 - Learning speed depends on salience (α_{CS}) and “association value” (β_{US})



Rescorla-Wagner Example



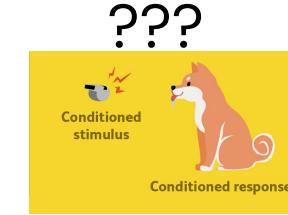
value(bell): 0
reward: 1
RPE: 1
New value(bell): 0.5

value(bell): 0.5
reward: 1
RPE: 0.5
New value(bell): 0.75

value(bell): 0.75
reward: 1
RPE: 0.25
New value(bell): 0.865

$$\text{RPE} = \text{reward} - \sum[\text{value}(\text{CS})]$$
$$\text{value}(\text{CS}) \leftarrow \text{value}(\text{CS}) + \alpha_{\text{CS}} * \beta_{\text{US}} * \text{RPE}$$

[[Assume $\alpha_{\text{CS}} * \beta_{\text{US}} = 0.5$]]



value(bell): 1

"Conditioned response"



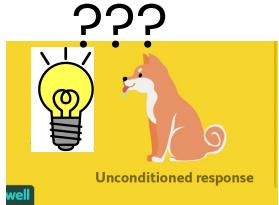
Blocking Example



```
value(bell):      1  
reward:          1  
RPE:             0  
New value(bell): 1 (no change)
```



```
value(bell):      1  
value(light):     0  
 $\Sigma[\text{value(CS)}]$ : 1  
reward:          1  
RPE:             0  
New value(bell): 1 (no change)  
New value(light): 0 (no change)
```



```
value(light):      0
```

No "Conditioned response"

$$\text{RPE} = \text{reward} - \sum[\text{value(CS)}]$$

$$\text{value(CS)} \leftarrow \text{value(CS)} + \alpha_{\text{CS}} * \beta_{\text{US}} * \text{RPE}$$

[[Assume $\alpha_{\text{CS}} * \beta_{\text{US}} = 0.5$]]



Operant conditioning



```
value(press|lev) : 0  
reward: 1  
RPE: 1  
New value(press|lev) : 0.5
```



```
value(press|lev) : 0.5  
reward: 1  
RPE: 0.5  
New value(press|lev) : 0.75
```

...

```
value(press|lev) : 1
```

"Goal-directed response" or "Habit"?

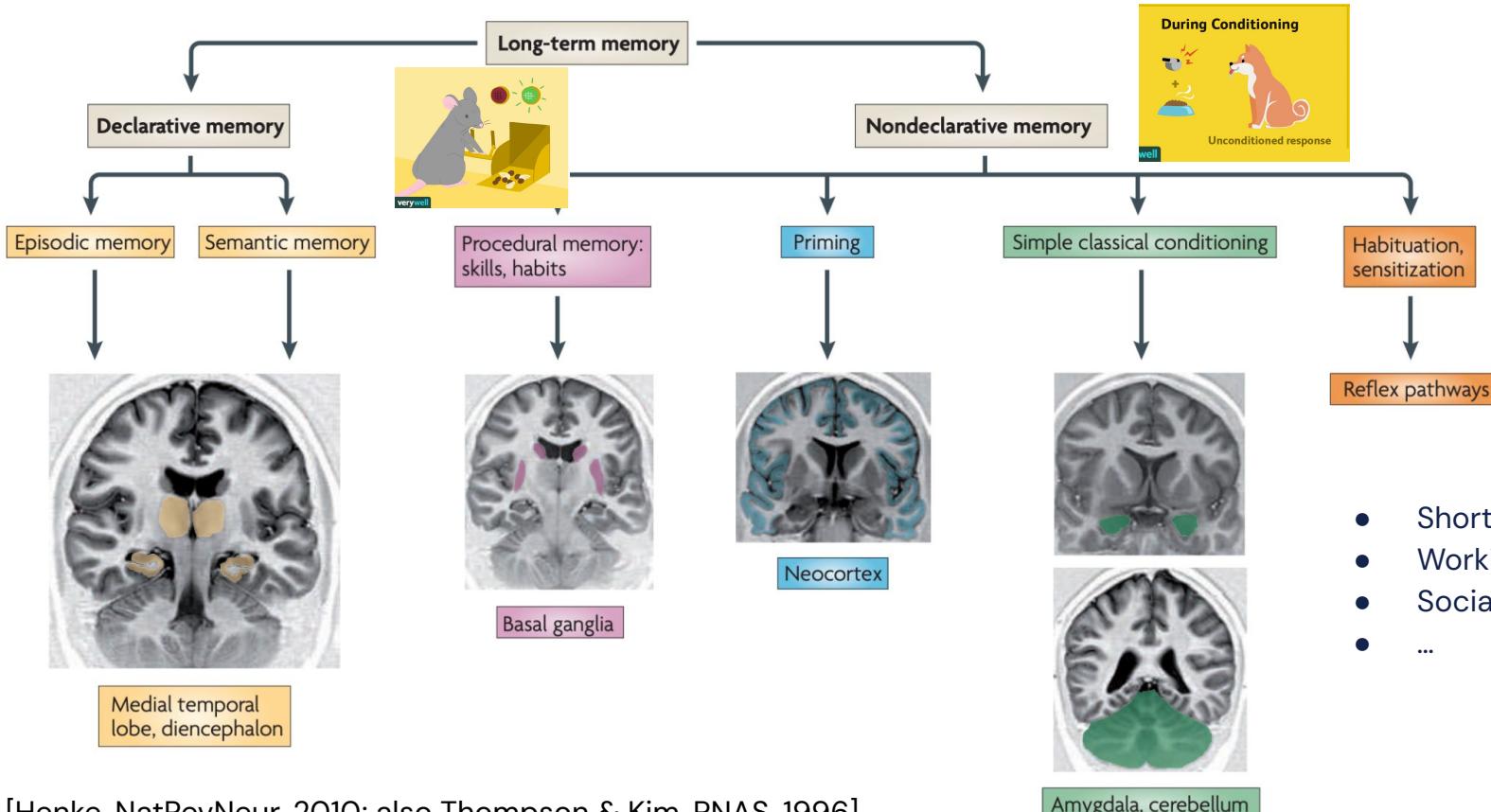


$$\text{RPE} = \text{reward} - \text{value(action|state)}$$

$$\begin{aligned}\text{value(action|state)} &\leftarrow \\ \text{value(action|state)} &+ \alpha * \text{RPE}\end{aligned}$$



Multiple memory systems



Questions?

Confidential - DeepMind



Reinforcement Learning (RL)

Confidential - DeepMind

1. Introduction
2. RL from a psychology perspective
- 3. RL from an AI perspective**
4. RL from a neuroscience perspective
5. Bringing it all together: RL as a cognitive model
6. Conclusion



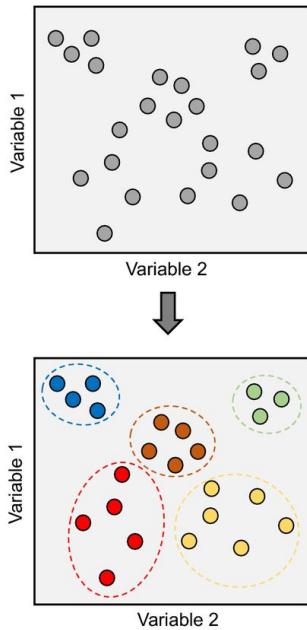
DeepMind

RL from an AI perspective

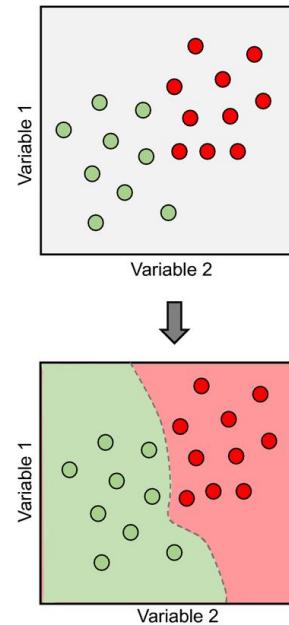


RL in the context of machine learning (ML)

Confidential - DeepMind



Unsupervised learning: Learn patterns or structure in data
(e.g., dimensionality reduction, clustering, ...)



Supervised learning: Learn to predict target(s)
(e.g., regression, classification, ...)

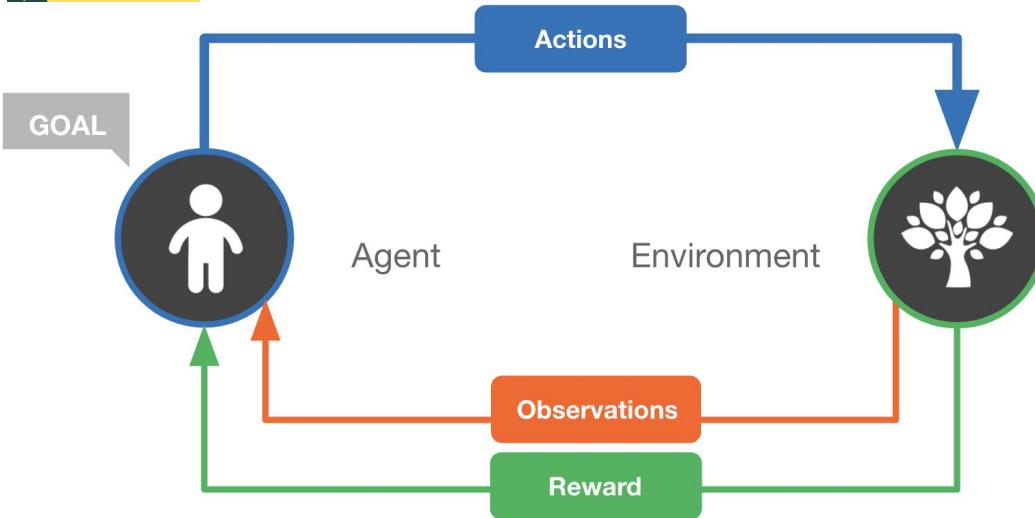


Reinforcement Learning: Learn from interactions in the world, through a scalar reward signal



RL Ingredients

Confidential - DeepMind



Agent: Learns a policy π that maps observations to actions, in order to maximize rewards.

Environment: E.g., experimental task; game (chess, Starcraft); factory (robotics); fusion reactor; ...

Reward:

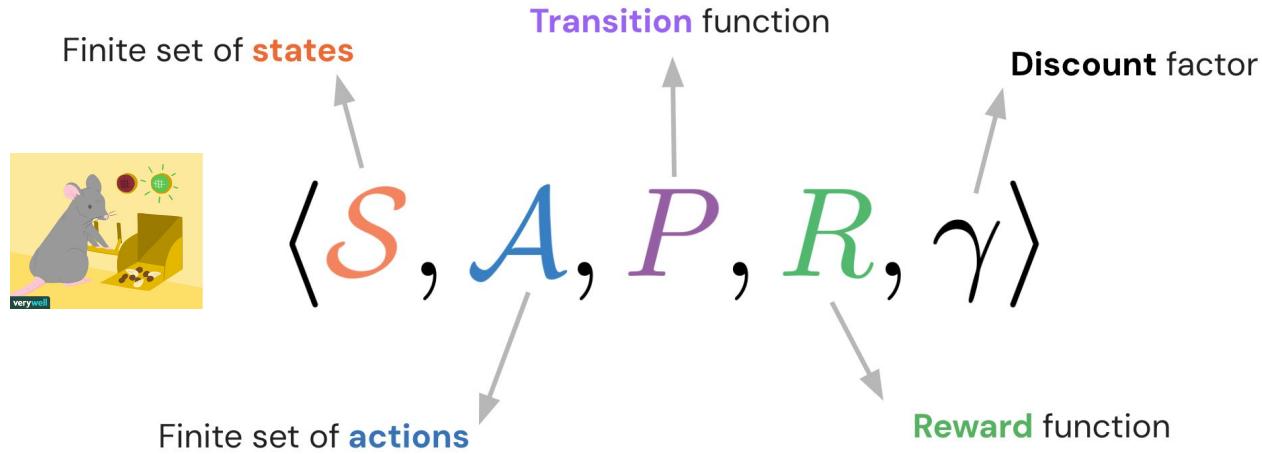
- *Extrinsic* (food, water, hard-coded)
- *Intrinsic* (curiosity, novelty, empowerment, learning progress, compression, explanation, ...)



The Markov Decision Process (MDP)

Confidential - DeepMind

Markov Decision Processes allow us to *formalize* and *solve* the RL problem.

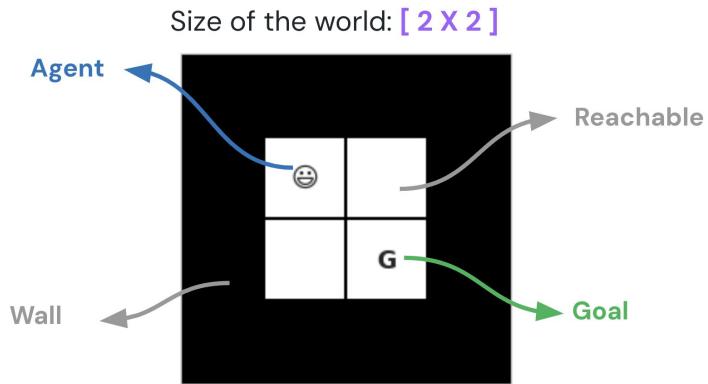


Markov Property: The next state depends only on the current state and action, not on the entire history (e.g., chess).

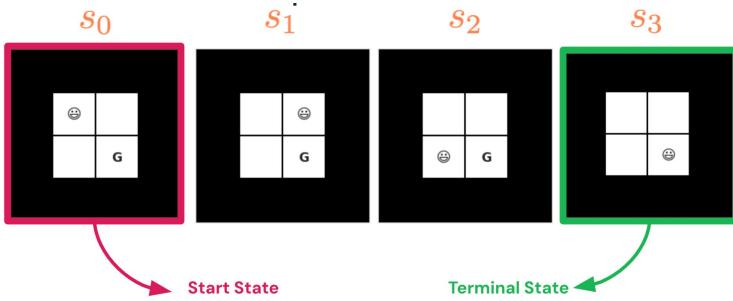
$$P(s_{t+1} | s_t, a_t, s_{t-1}, a_{t-1}, \dots, s_0) = P(s_{t+1} | s_t, a_t)$$



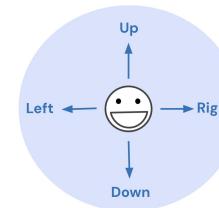
Grid Worlds



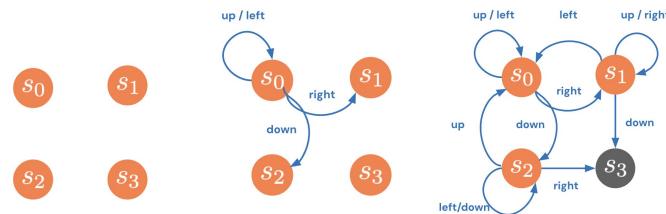
State space \mathcal{S}



Action Space \mathcal{A}



Transition model P

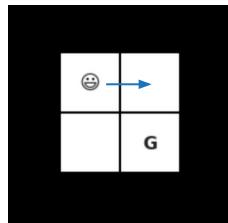


Rewards R

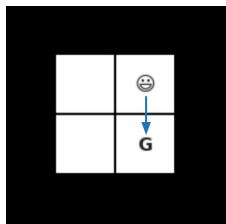
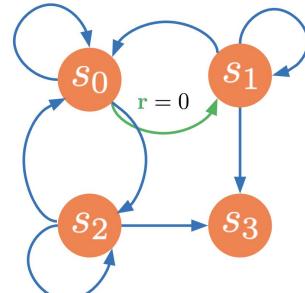
Empty cell: 0
Wall: -5
Goal: +10



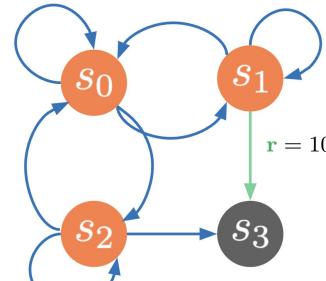
Policy and Values



[assume $\gamma = 0.9$]
 $Q^{\pi^*}(s_0, a_{\text{right}}) = 0 + 0.9 * 10 = 9$



$Q^{\pi^*}(s_1, a_{\text{down}}) = 10 + 0.9 * 0 = 10$



Agent's goal: Maximize (γ -discounted) sum of future rewards:

$$G_t = \underbrace{r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots}_{\text{Return}}$$

To achieve this, the agent learns an action **policy** π :

$$a_t \sim \pi(a_t | s_t)$$

How do we find this policy?



Using values!

$$Q^\pi(s_t, a_t) = \mathbb{E}_\pi [r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots | s_t, a_t]$$

Once we have values, finding the optimal policy is easy:

$$\pi^*(s) = \max_a Q^*(s, a)$$



Temporal Difference (TD) Learning

$$\begin{aligned}\text{Value}(s) &+= \alpha * \text{RPE} \\ \text{RPE} &= r - \text{Value}(s)\end{aligned}$$

$$V_{t+1}(s_t) \leftarrow V_t(s_t) + \eta \delta_t$$

Learning rate

New estimate of value of current state Old estimate of value of current state Reward prediction error

$$\delta_t = R_t + \gamma V_t(s_{t+1}) - V_t(s_t)$$

Reward prediction error Actual observed value of current state (written the recursive way) Old estimate of value of current state

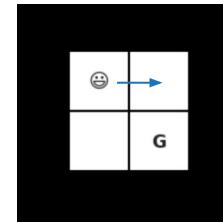
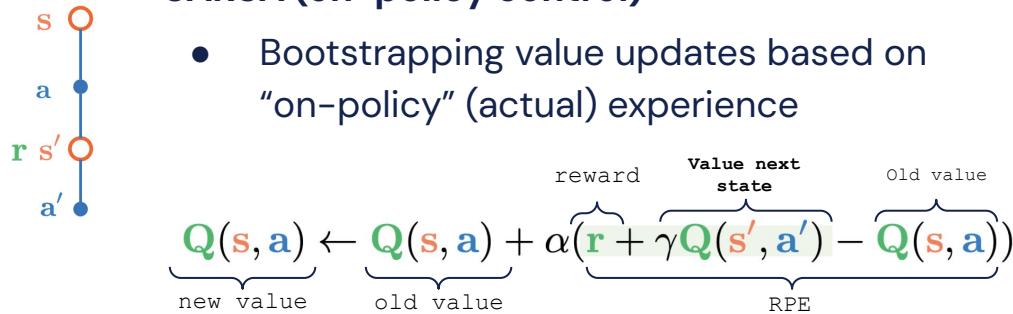


Learning the value function

Problem: We can't predict the future! (And we don't want to...)

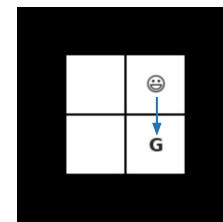
SARSA (on-policy control)

- Bootstrapping value updates based on "on-policy" (actual) experience



All Q's=0

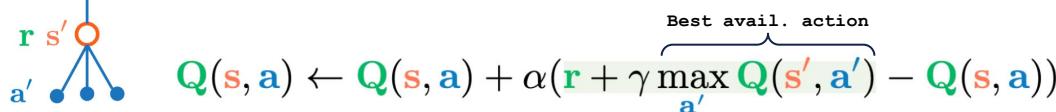
$$Q(s_0, \text{right}) \leftarrow 0 + \alpha (0 + \gamma^* 0 - 0) = 0$$



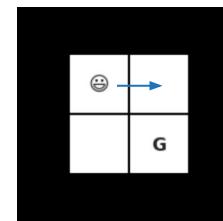
$$Q(s_1, \text{down}) \leftarrow 0 + \alpha (10 + \gamma^* 0 - 0) = 5$$

Q-learning (off-policy control)

- Bootstrapping value updates based on "off-policy" (hypothetical) experience



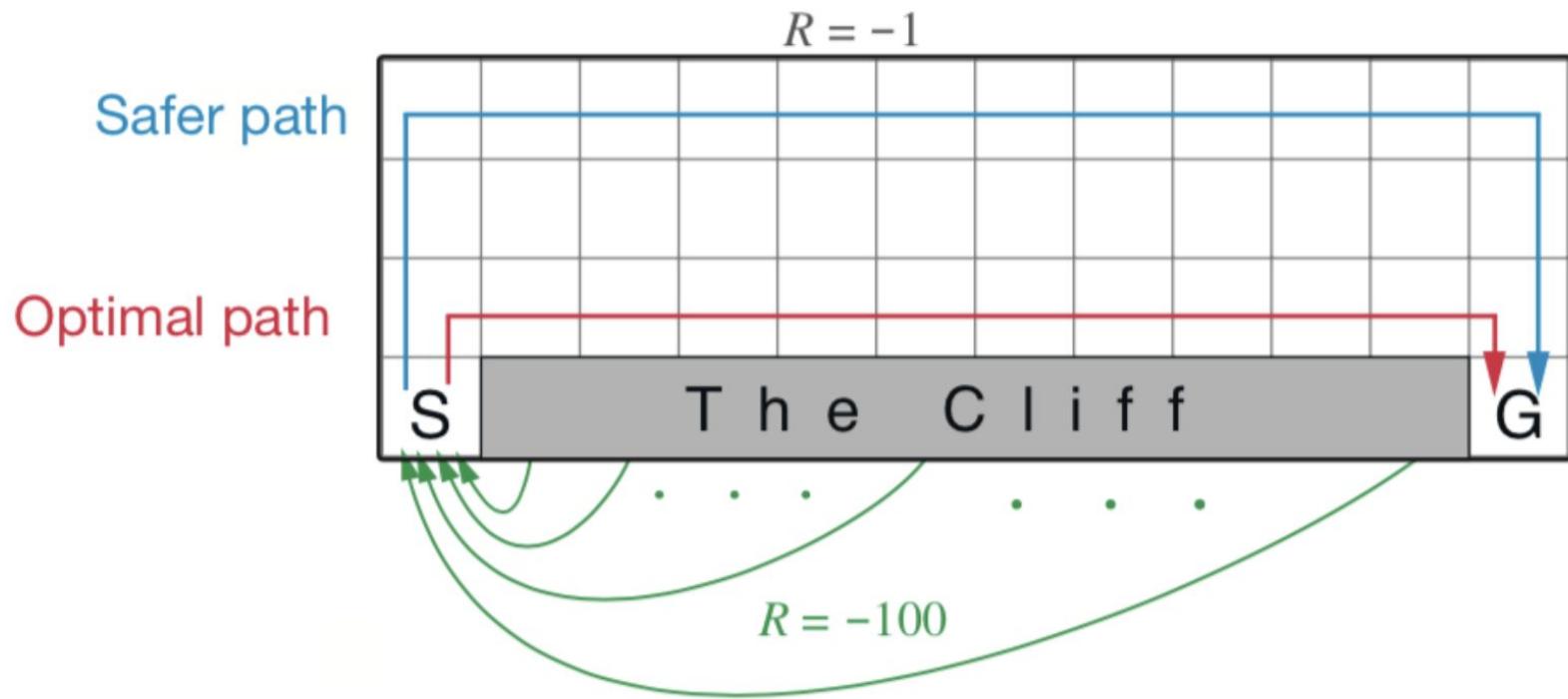
$$\pi^*(s) = \max_a Q^*(s, a)$$



$$Q(s_0, \text{right}) \leftarrow 0 + \alpha (0 + \gamma^* 5 - 0) = 2.25$$



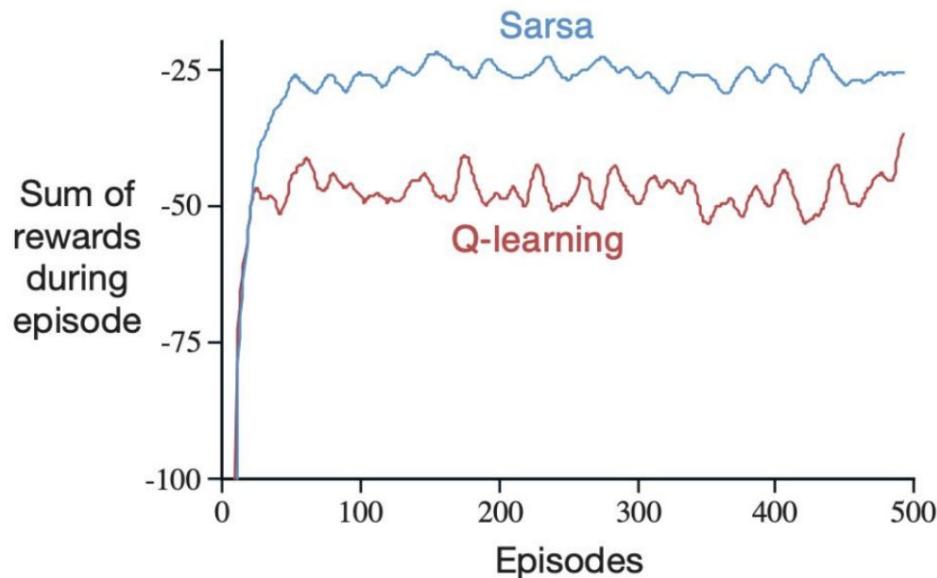
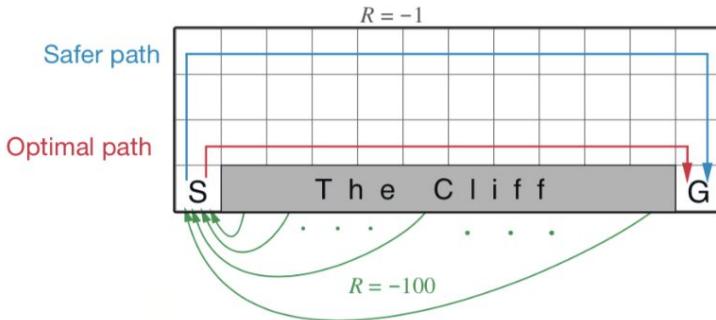
SARSA vs Q-Learning: The Cliff-walking Example



Sutton & Barto. Reinforcement Learning: An Introduction. (Chapter 6)

SARSA vs Q-Learning: The Cliff-walking Example

- **Q-learning** learns the **optimal path** while its online performance is worse than SARSA.
- **SARSA** learns the **safer path**.

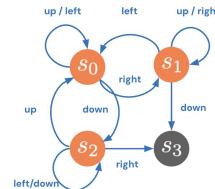


Cheat Sheet

Rescorla Wagner: keep track of reward expectations



TD Learning: +over time



SARSA: +control (on-policy)

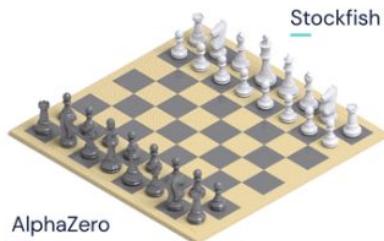


Q-Learning: +control (off-policy)

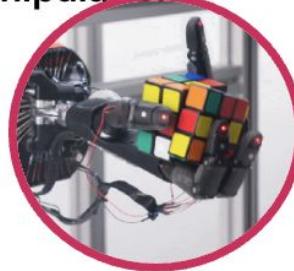


Real-world Reinforcement Learning: Examples

Game playing



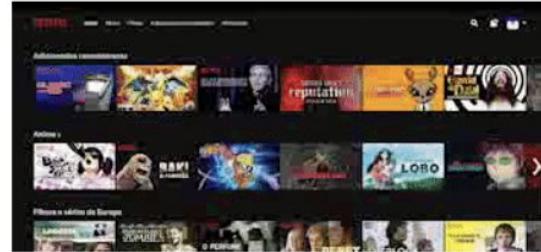
Robotics /
Manipulation



Self-driving cars



User personalization



Managing energy usage



Questions?

Confidential - DeepMind



Reinforcement Learning (RL)

Confidential - DeepMind

1. Introduction
2. RL from a psychology perspective
3. RL from an AI perspective
- 4. RL from a neuroscience perspective**
5. Bringing it all together: RL as a cognitive model
6. Conclusion



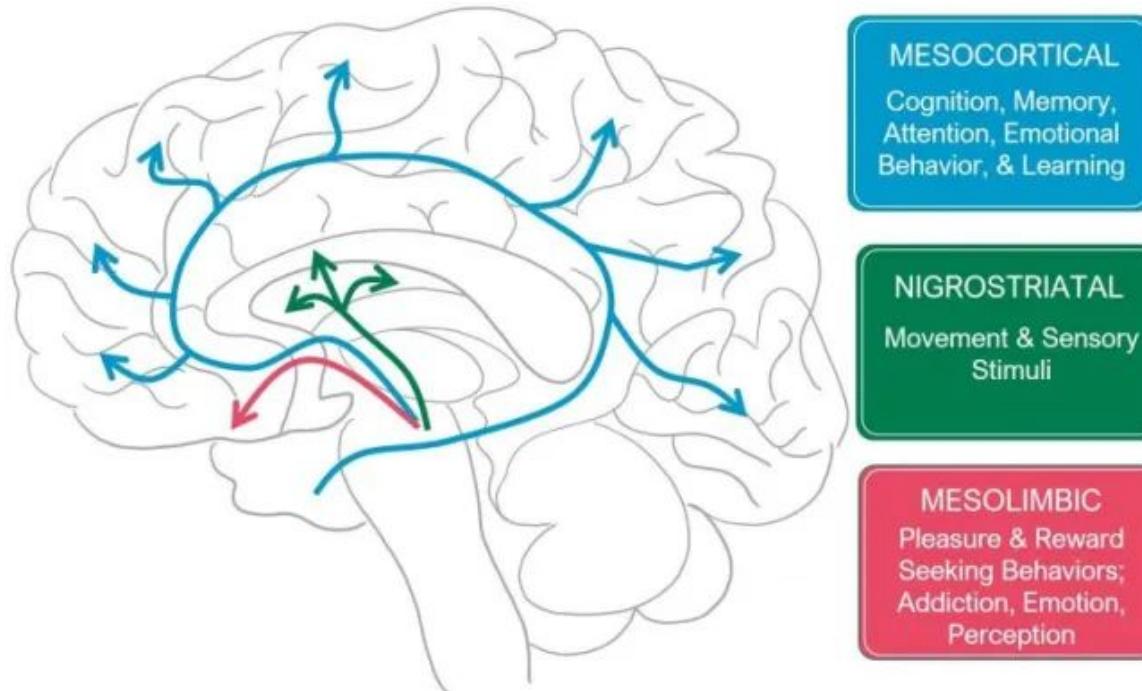
DeepMind

RL in neuroscience



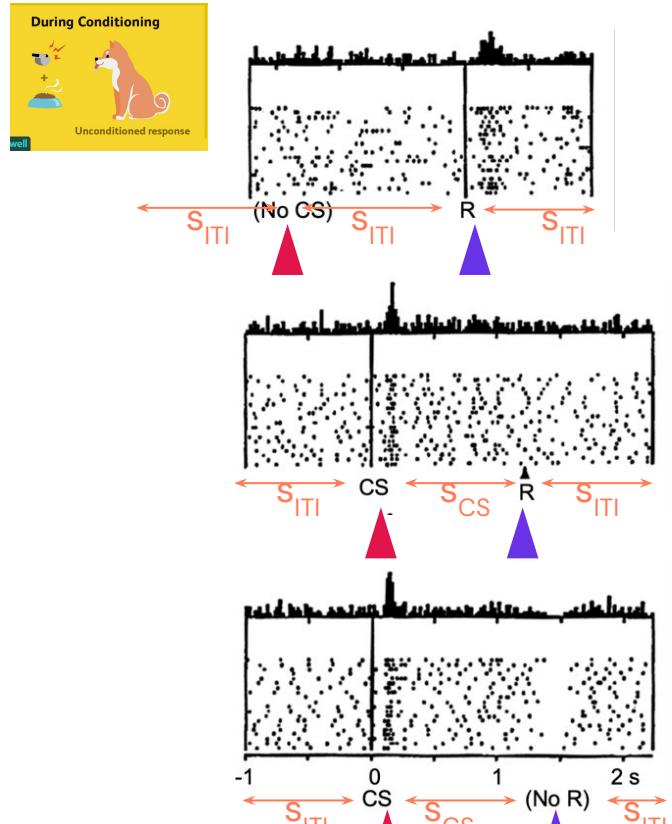
The Neurotransmitter Dopamine

Confidential - DeepMind



Dopamine Reward Prediction Errors

$$RPE = r + \gamma Q(s', a') - Q(s, a)$$



[Montague et al., 1996; Schultz et al., 1997]

$$\begin{aligned} RPE &= r + \gamma Q(s_{ITI}) - Q(s_{ITI}) \\ &= 0 + \gamma 0 - 0 = 0 \end{aligned}$$

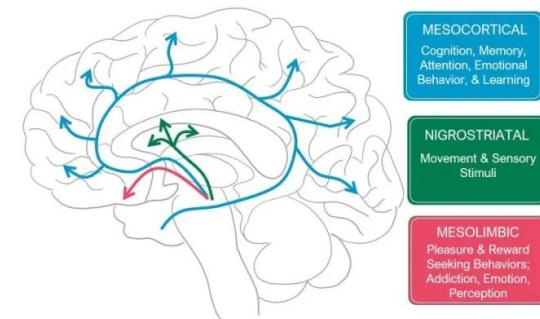
$$\begin{aligned} RPE &= r + \gamma Q(s_{ITI}) - Q(s_{ITI}) \\ &= 1 + \gamma 0 - 0 = 1 \end{aligned}$$

$$\begin{aligned} RPE &= r + \gamma Q(s_{CS}) - Q(s_{ITI}) \\ &= 0 + \gamma 1 - 0 = 0.9 \end{aligned}$$

$$\begin{aligned} RPE &= r + \gamma Q(s_{ITI}) - Q(s_{CS}) \\ &= 1 + \gamma 0 - 0.9 = 0.1 \end{aligned}$$

$$\begin{aligned} RPE &= r + \gamma Q(s_{CS}) - Q(s_{ITI}) \\ &= 0 + \gamma 1 - 0 = 0.9 \end{aligned}$$

$$\begin{aligned} RPE &= r + \gamma Q(s_{ITI}) - Q(s_{CS}) \\ &= 0 + \gamma 0 - 0.9 = -0.9 \end{aligned}$$

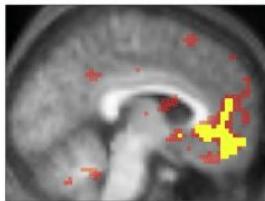


- Converging evidence across studies and species
- Mostly in simple conditioning paradigms

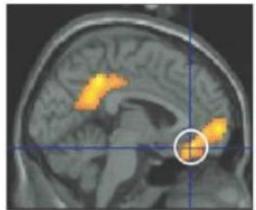
[Niv, 2009]



Human fMRI



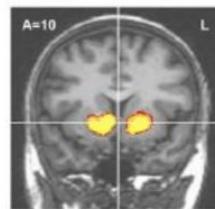
money
value predicted
(Daw et al 2006)



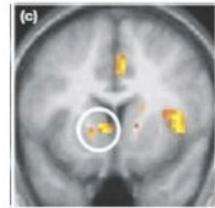
faces
attractiveness
(O'Doherty et al 2003)



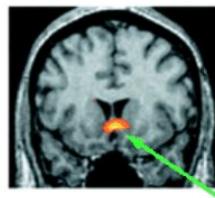
Coke or Pepsi
degree favored
(McClure et al. 2004)



money
gain vs loss
(Kuhnen & Knutson
2005)



food odors
valued vs devalued
(Gottfreid et al 2003)



juice
unpredictable vs
predictable
(Berns et al 2001)

Rewards / reward
anticipation activate:

- Ventromedial prefrontal cortex
- Orbitofrontal cortex
- Striatum

➤ *Generalized
appetitive
function?*



Questions?

Confidential - DeepMind



Reinforcement Learning (RL)

Confidential - DeepMind

1. Introduction
2. RL from a psychology perspective
3. RL from an AI perspective
4. RL from a neuroscience perspective
- 5. Bringing it all together: RL as a cognitive model**
6. Conclusion



DeepMind

RL for Cognitive Modeling



What is Cognitive Modeling?

Goal: Understand behavior, cognitive process



Method:

- Find model (e.g., RL, Regression, DDM, ...)
- “Fit” model (find best parameters)
- Expand model
 - e.g., reward vs punishment [Frank et al., 2004]; WM [Collins & Frank, 2012]; counterfactuals [Boorman et al., 2011]; ...
- Model comparison (AIC, BIC, WAIC, ...)

Result:

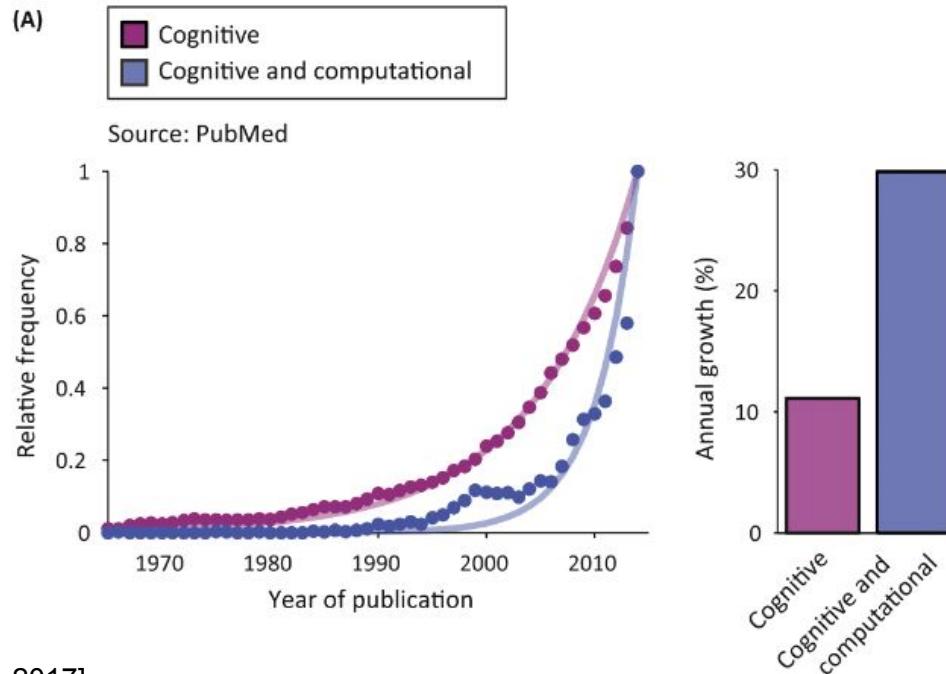
- “Cognitive process”
- Fitted parameters (individual differences)
- Normative understanding (optimality)
- Quantitative methods, statistics
- Complex, multi-step processes
- Precise prediction

RL

$$\begin{aligned} \text{RPE} &= r + \gamma Q(s', a') - Q(s, a) \\ Q(s, a) &\leftarrow Q(s, a) + \alpha * \text{RPE} \end{aligned}$$



Computational modeling is on the rise!



[Palminter et al., 2017]



What is RL Modeling?

Goal

Reward

Ingredients

Algorithm



+1

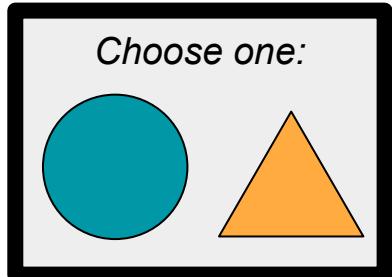
action = [\rightarrow , \leftarrow]

state = []

reward = [0, +1]

$$RPE = r + \gamma Q(s', a') - Q(s, a)$$

$$Q(s, a) \leftarrow Q(s, a) + \alpha * RPE$$



+1

action = [F H]

state = []

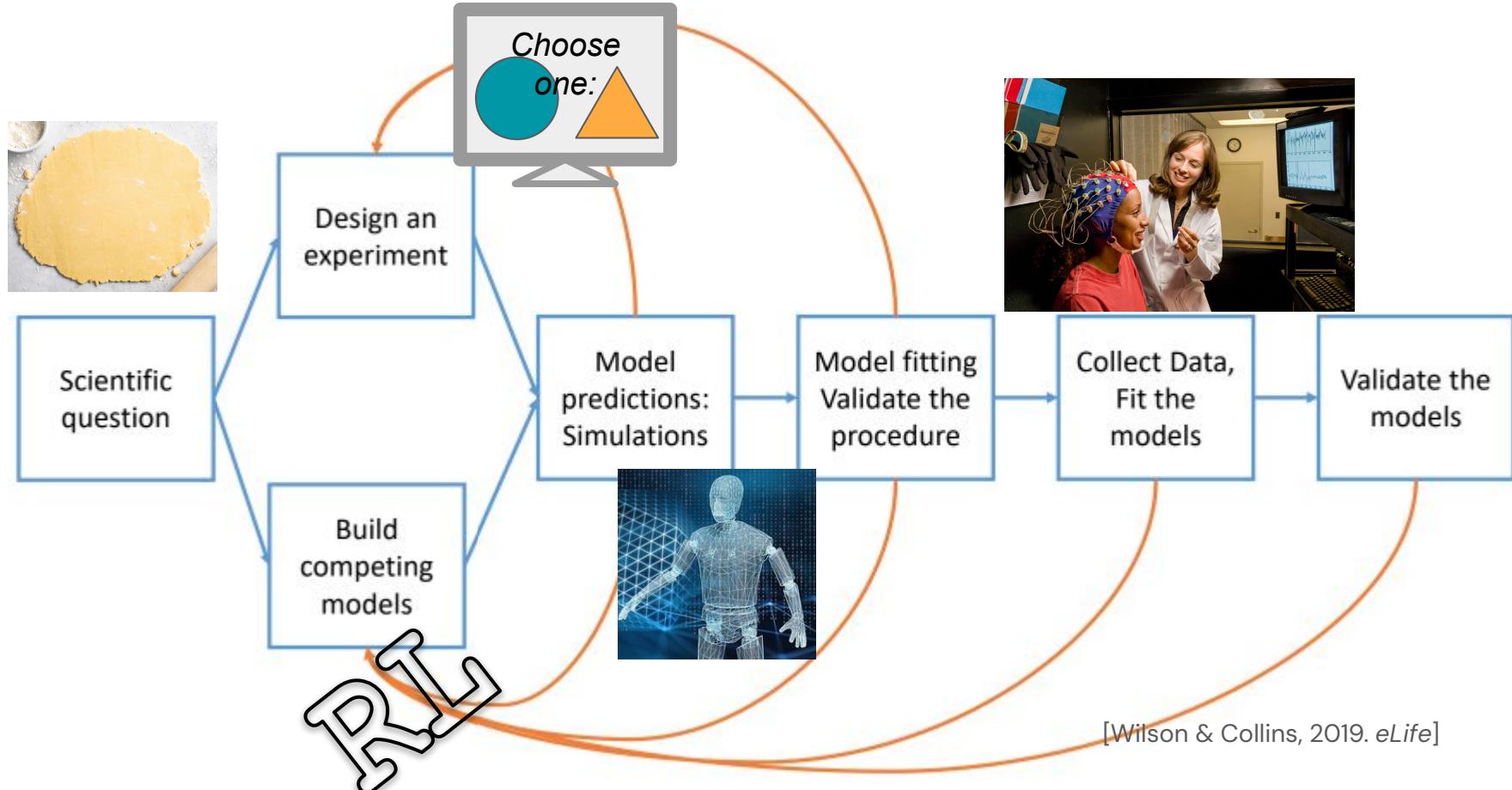
reward = [0, +1]

$$RPE = r + \gamma Q(s', a') - Q(s, a)$$

$$Q(s, a) \leftarrow Q(s, a) + \alpha * RPE$$

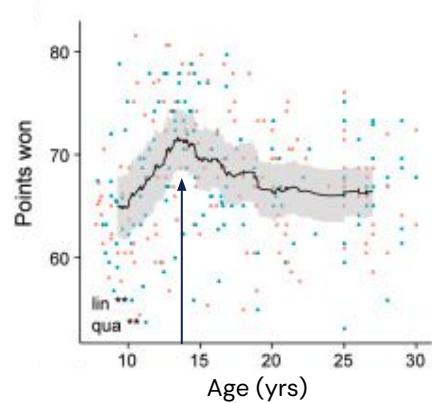
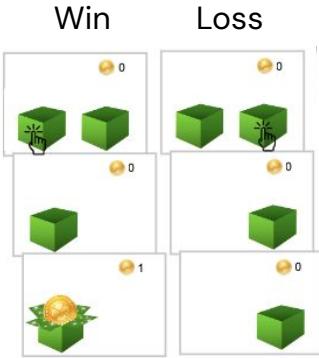
A Recipe for Cognitive Modeling

Confidential - DeepMind



Learning to Reversal Learn

Goal: Understand age trajectory of reversal learning



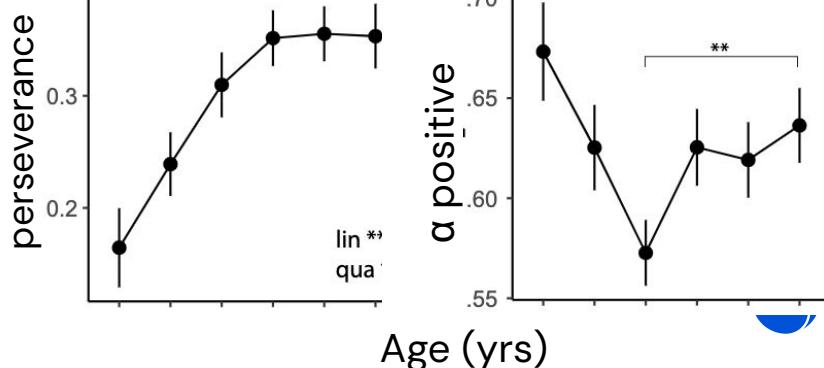
- Best performance at ~13–15

Why? Cognitive mechanism?

RL

$$RPE = r - Q(s,a)$$

$$Q(s,a) \leftarrow Q(s,a) + \alpha * RPE$$



Questions?

Confidential - DeepMind



Reinforcement Learning (RL)

Confidential - DeepMind

1. Introduction
2. RL from a psychology perspective
3. RL from an AI perspective
4. RL from a neuroscience perspective
5. Bringing it all together: RL as a cognitive model
6. **Conclusion**



DeepMind

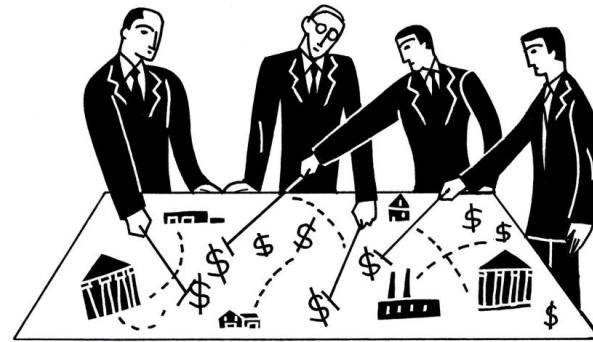
Conclusion



Where do rewards come from?



Evolution?

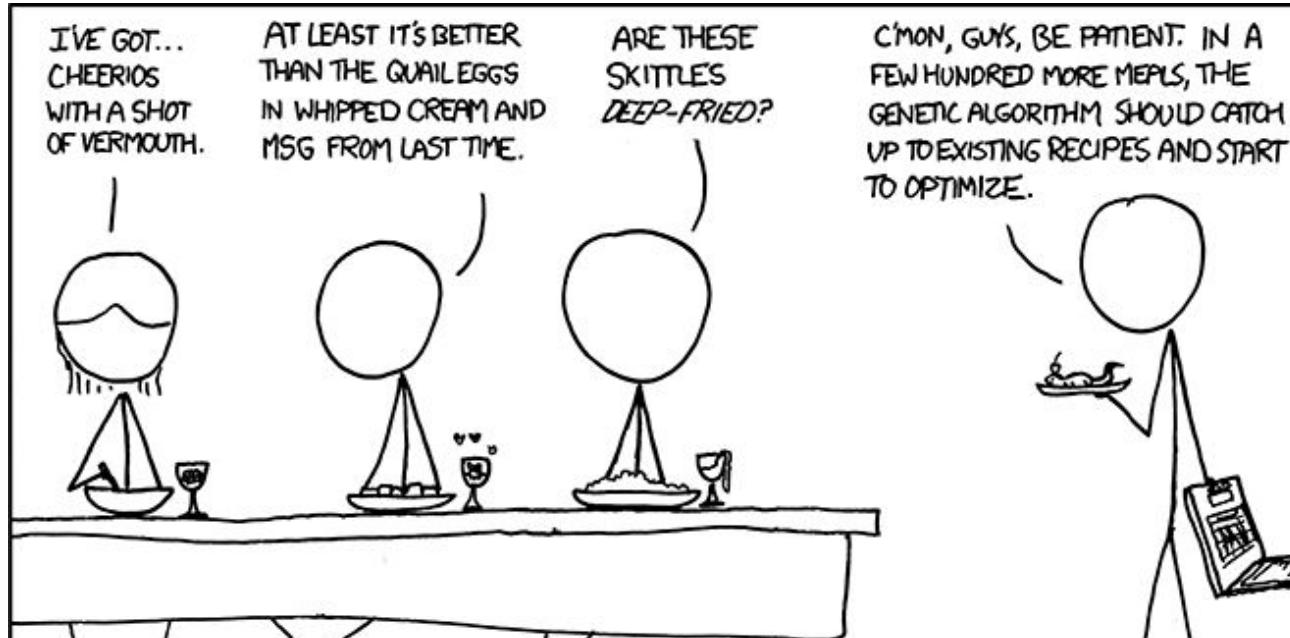


Economists?

- Intrinsic / extrinsic?
- Innate / learned?
- Context-dependent?
- Individual differences?



Exploration



- Epsilon-greedy / softmax?
- Structured exploration?
- Intrinsic goals?
- Sparse rewards



Credit Assignment

Beginning



End



How to link distal outcomes to earlier causes despite many intervening events?

How to generalize over similar + different instances?

How to use knowledge of structure inform credit assignment?



Models as Maps

Original



Model



- Cognitive model = map
 - Smaller, more abstract
 - Loose information
- Different maps
 - Depending on the purpose
 - No one “true” map

Questions?

Confidential - DeepMind



Want to Learn More?

Confidential - DeepMind

Books

- [Reinforcement Learning: an Introduction by Sutton & Barto](#)
- [Algorithms for Reinforcement Learning by Csaba Szepesvari](#)

Lectures and course

- [Neuromatch Lecture on RL by Jane Wang and Feryal Behbahani](#)
- [RL Course by David Silver](#)
- [Reinforcement Learning Course | UCL & DeepMind](#)
- [Emma Brunskill Stanford RL Course](#)
- [RL Course on Coursera by Martha White & Adam White](#)

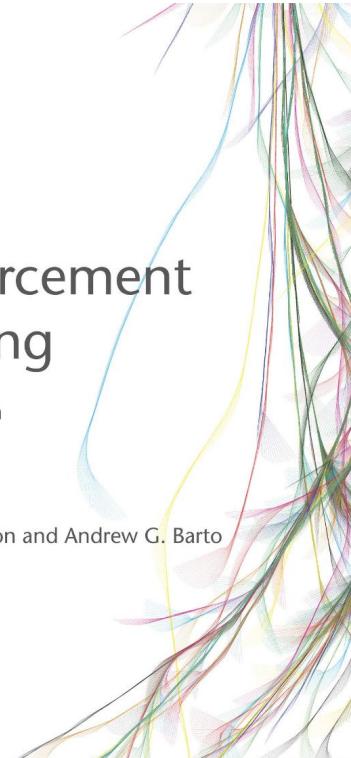
More practical

- [Spinning Up in Deep RL by Josh Achiam](#)
- [Acme white paper & Colab tutorial](#)
- [OpenAI Gym](#)

Reinforcement Learning

An Introduction
second edition

Richard S. Sutton and Andrew G. Barto



Acknowledgements

Kim Stachenfeld, Anne Collins, Jane Wang, Feryal Behbahani, Nathaniel Daw, Chris Knutsen, Kevin Miller, Zeb Kurth-Nelson, Matt Botvinick



Acknowledgements

Slides:



Anne Collins



Kim Stachenfeld

Collaborators at GDM:



Zeb
Kurth-Nelson



Kevin Miller



Nathaniel Daw



Chris
Summerfield





*Dog
tricks
by
Justy*

