DeepMind

REINFORCEMENT LEARNING

Computational Modeling for Learning and Decision Making

Maria K. Eckstein
Google DeepMind
(mariaeckstein@deepmind.com)

Reinforcement Learning (RL)

-> What do both videos have in common?

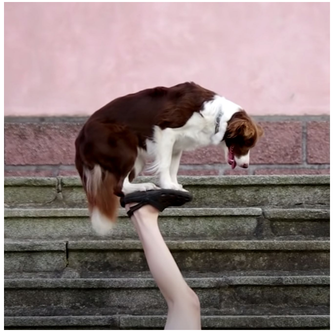
Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

What is RL?

Learning from rewards; and punishment.

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

How to Use RL (as a Cognitive Model)?

Goal	Reward	Ingredients	Algorithm
	+1	**a**ction = [→, ←] **s**tate = [] **r**eward = [0, +1]	$Q(s,a) \leftarrow Q(s,a) + \alpha\, RPE$ $RPE = r + \gamma\, Q(s',a') - Q(s,a)$
		**a**ction = [jump, stand] **s**tate = [] **r**eward = [0, ]	???

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Questions?

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

DeepMind

Lecture Roadmap

Reinforcement Learning (RL)

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

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

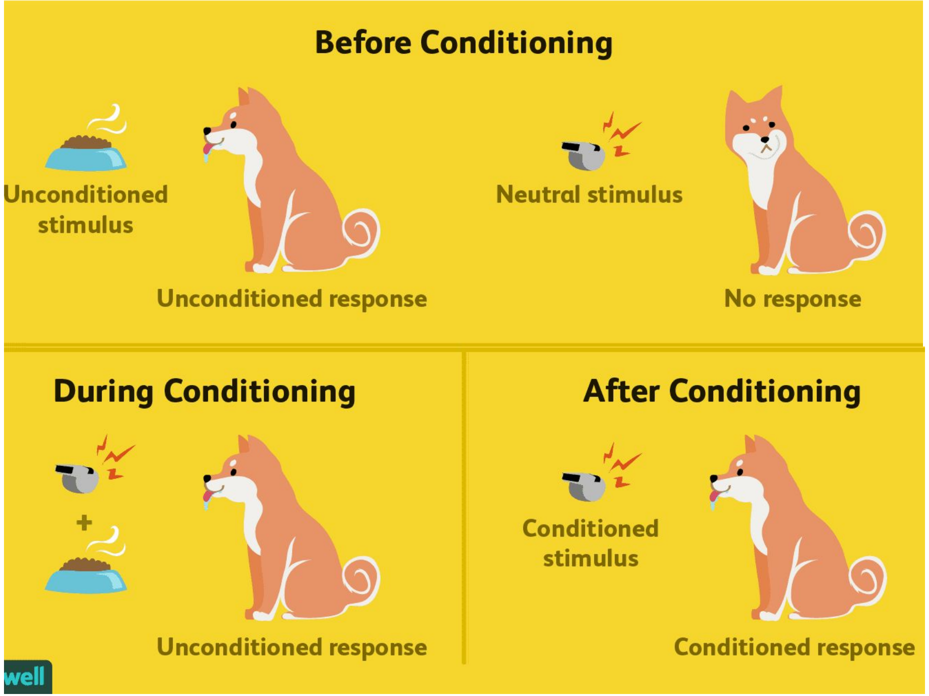
Reinforcement Learning (RL)

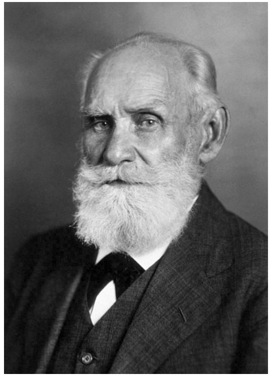
Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

DeepMind

RL from a psychology perspective

Classical Conditioning



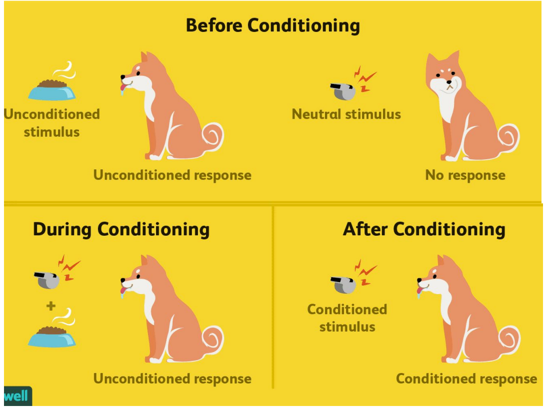


Ivan Pavlov
(1849–1936)

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

redit:
Maria Eckstein
(mariaeckstein@deepmind.com)

The Rescorla-Wagner Model (1972)



$$\text{RPE} = \lambda - \overbrace{\Sigma[\text{value}(\text{CS})]}^{\text{Combined predictive value of all stimuli}}$$

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

- 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
    - λ is the maximum conditioning possible for the US
    - Learning speed depends on "salience" (α_{CS}) and "association value" (β_{US})

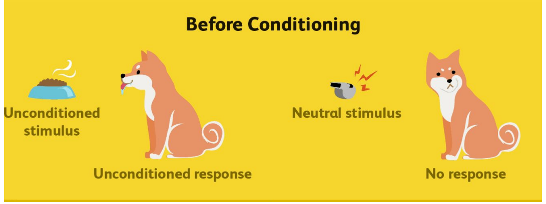
credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Rescorla-Wagner Example

$$\text{RPE} = \lambda - \Sigma[\text{value}(\text{CS})]$$

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

[[Assume $\alpha_{CS} * \beta_{US} = 0.5$ and $\lambda = 1$]]

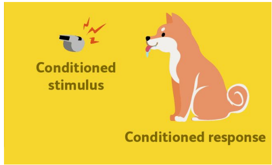


value(bell):	0
λ:	1
RPE:	1
New value(bell):	0.5

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

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

???



value(bell): 1

“Conditioned response”

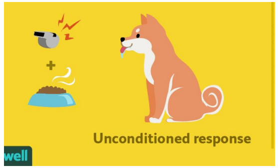
Credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Blocking Example

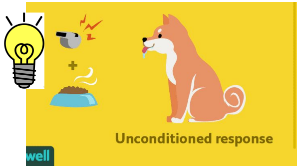
$$RPE = \lambda - \Sigma[value(CS)]$$

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

[[Assume $\alpha_{CS} * \beta_{US} = 0.5$ and $\lambda = 1$]]

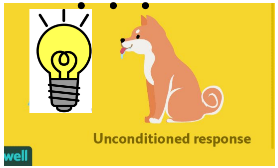



```
value(bell):        1
λ:                  1
RPE:                0
New value(bell):    1 (no change)
```




```
value(bell):        1
value(light):       0
Σ[value(CS)]:       1
λ:                  1
RPE:                0
New value(bell):    1 (no change)
New value(light):   0 (no change)
```

???




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

No "Conditioned response"

redit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Operant conditioning

RPE = reward - value(action|state)

value(action|state) <-

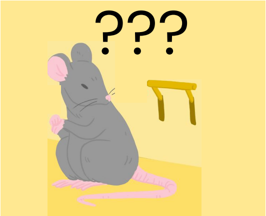
value(action|state) + α * RPE

Mind

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

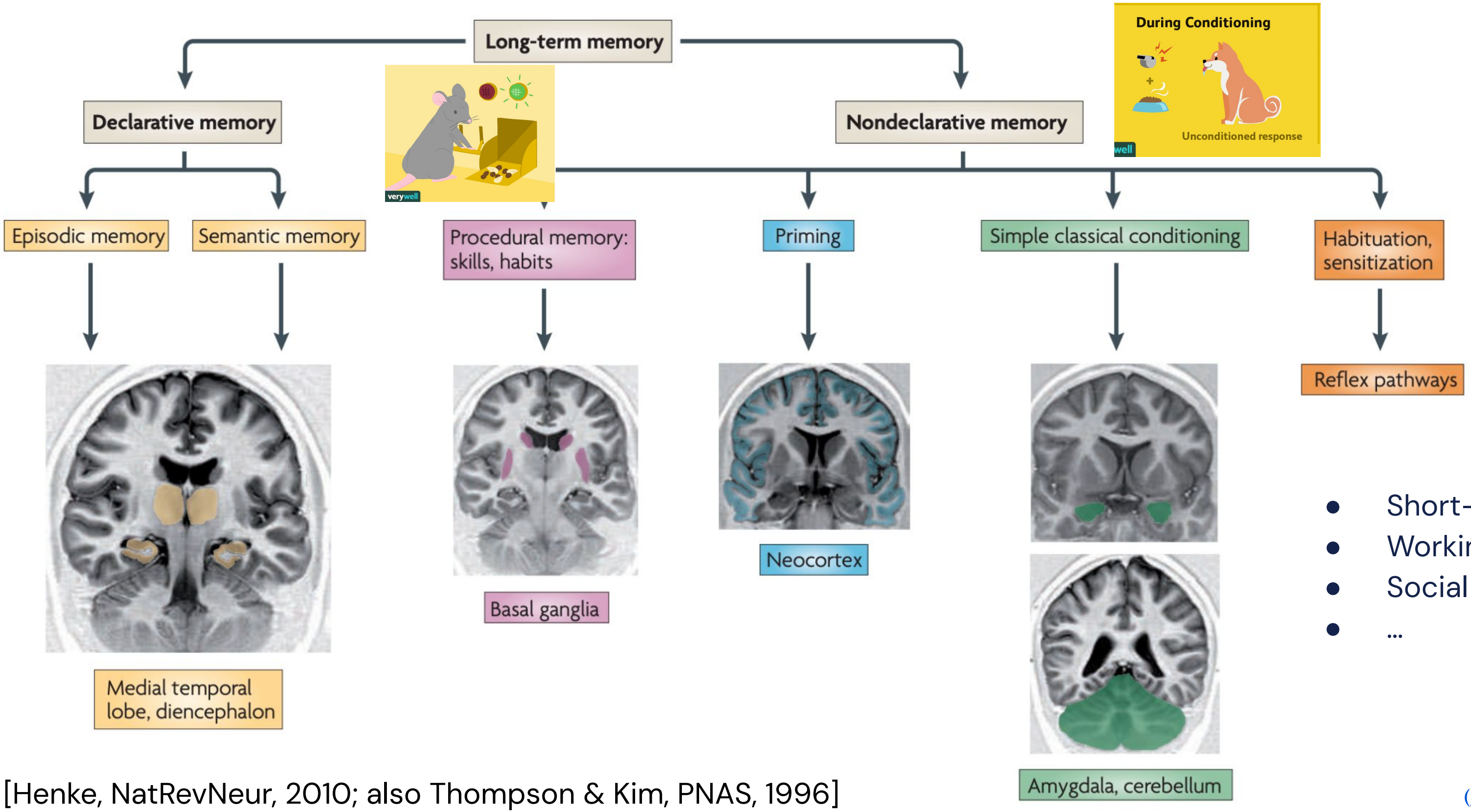
Quizz: According to this theory, what would the trained rat do when it is fully satiated and sees the lever?

A) Press the lever
B) Not press the lever

-> Link to "*habitual*" versus "*goal-directed*" behavior.

...redit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Multiple memory systems



- Short-term memory
- Working memory
- Social memory
- …

[Henke, NatRevNeur, 2010; also Thompson & Kim, PNAS, 1996]

Credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Questions?

Reinforcement Learning (RL)

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

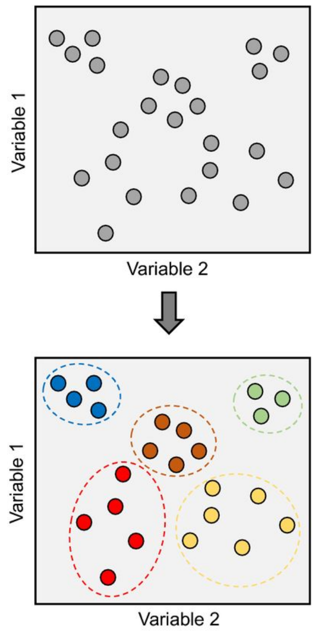
Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

DeepMind

RL from an AI perspective

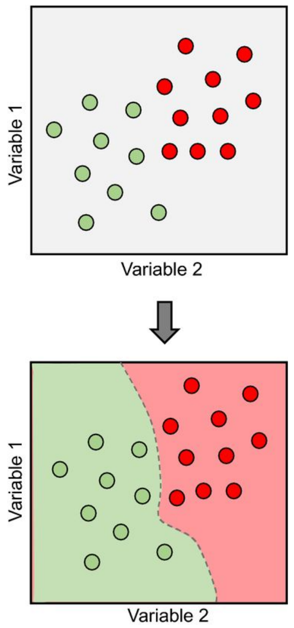
Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

RL in the context of machine learning (ML)



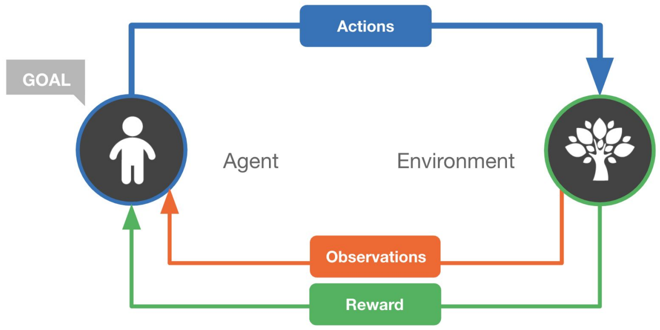
Unsupervised learning: Learn patterns or structure in data

(e.g., dimensionality reduction, clustering, …)



Supervised learning: Learn to predict target(s)

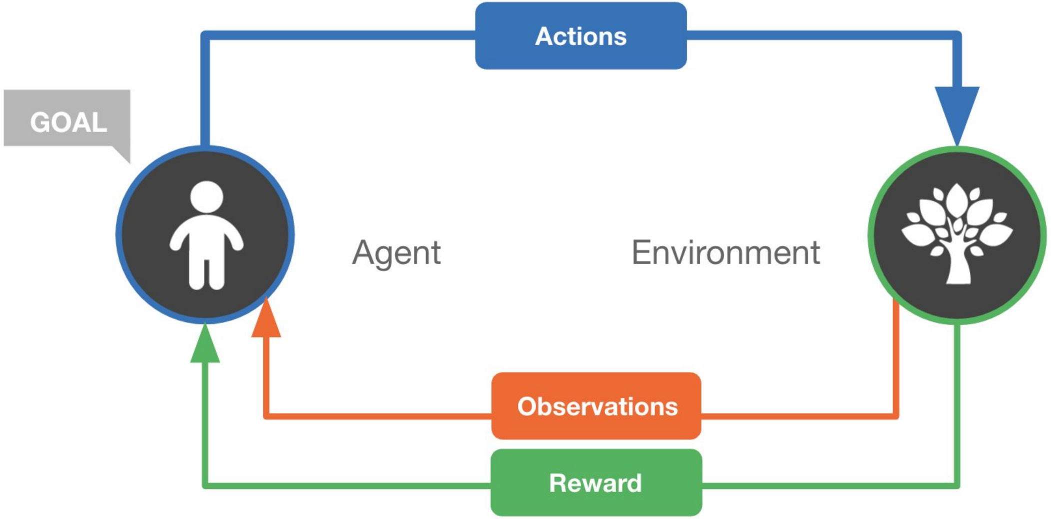
(e.g., regression, classification, …)



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

redit:
Maria Eckstein
(mariaeckstein@deepmind.com)

RL Ingredients



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, …)

redit:
Maria Eckstein, Jane Wang, Feryal Behbahani (mariaeckstein@deepmind.com)

The Markov Decision Process (MDP)

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

Finite set of **states**

Transition function

Discount factor

$$\langle S, \mathcal{A}, P, R, \gamma \rangle$$

Finite set of **actions**

Reward function

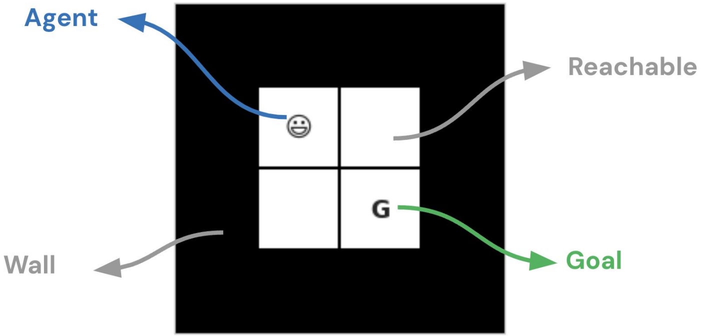
Markov Property: The next state depends only on the current state and action, not on the entire history (e.g., chess).

$$P(\underbrace{s_{t+1}}_{\text{Future}} | \underbrace{s_t, a_t}_{\text{Present}}, \underbrace{s_{t-1}, a_{t-1}, \ldots, s_0}_{\text{Past}}) = P(\underbrace{s_{t+1}}_{\text{Future}} | \underbrace{s_t, a_t}_{\text{Present}})$$

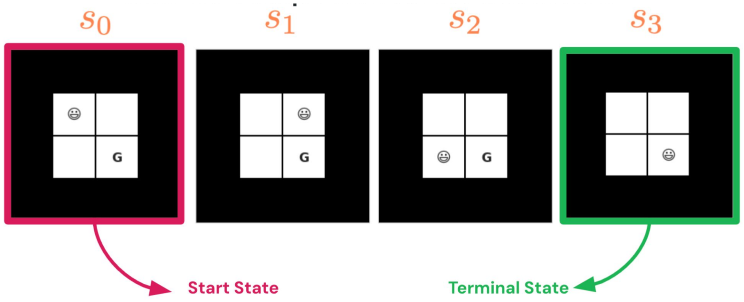
Slide credit:
Maria Eckstein, Jane Wang, Feryal Behbahani
(mariaeckstein@deepmind.com)

Grid Worlds

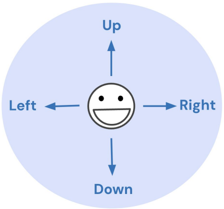
Size of the world: [2 X 2]



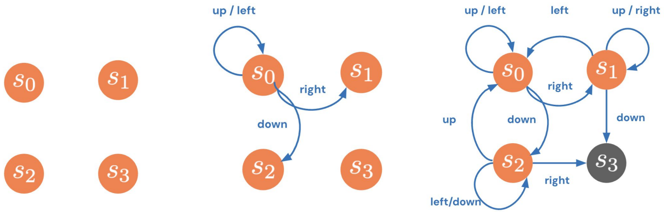
State space S



Action Space $\mathcal{A}$



Transition model P



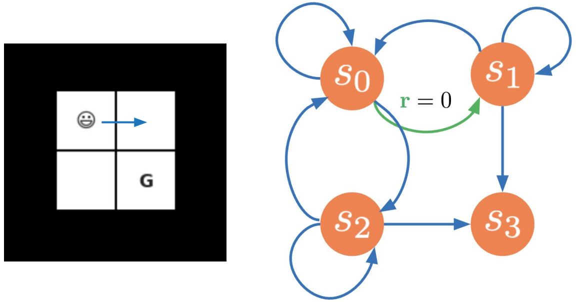
Rewards R

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

redit:
Maria Eckstein, Jane Wang, Feryal Behbahani
(mariaeckstein@deepmind.com)

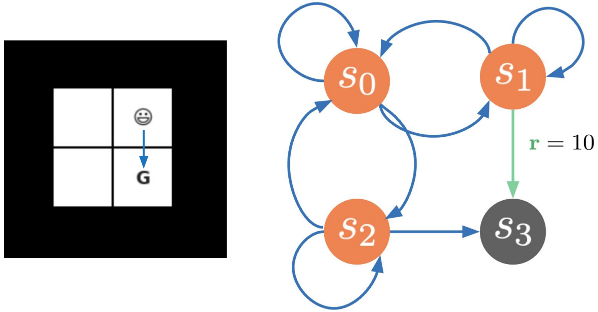
Policy and Values

In MDP terms:



[assume γ = 0.9]

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



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

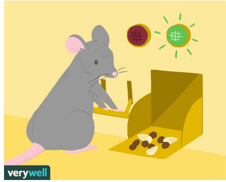
Agent's goal: Maximize (γ-discounted) sum of future rewards:

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

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! Once we have (optimal) values, executing the optimal policy is easy:

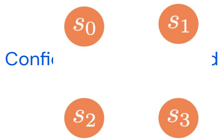
$$\pi^*(s) = \max_a Q^*(s, a)$$

This works because values are defined as:

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

credit: Maria Eckstein, Jane Wang, Feryal Behbahani (mariaeckstein@deepmind.com)

Learning the value function

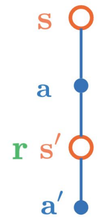


$$Q^{\pi}(s_t, a_t) = \mathop{\mathbb{E}}_{\pi}\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots | s_t, a_t\right]$$

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

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

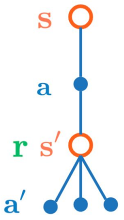
SARSA (on-policy control)



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

$$\underbrace{Q(s,a)}_{\text{new value}} \leftarrow \underbrace{Q(s,a)}_{\text{old value}} + \alpha\underbrace{(\overbrace{r}^{\text{reward}} + \overbrace{\gamma Q(s',a')}^{\text{Value next state}} - \overbrace{Q(s,a)}^{\text{Old value}})}_{\text{RPE}}$$

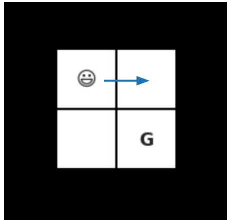
Q-learning (off-policy control)



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

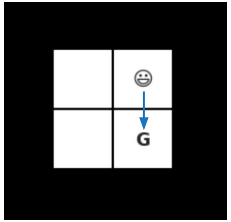
$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \overbrace{\gamma \max_{a'} Q(s',a')}^{\text{Best avail. action}} - Q(s,a))$$

$$\pi^*(s) = \max_{a} Q^*(s,a)$$

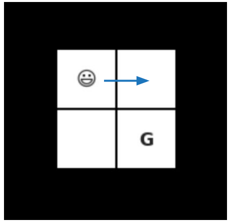


All Q's=0

Q(s0,right) <- 0+α (0+γ*0-0) = 0



Q(s1,down) <- 0+α (10+γ*0-0) = 5



Q(s0,right) <- 0+α (0+γ*5-0) = 2.25

redit:
Maria Eckstein, Jane Wang, Feryal Behbahani
(mariaeckstein@deepmind.com)

Temporal Difference (TD) Learning

```
Value(s) += α * RPE
RPE = r - Value(s)
```

Learning rate

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

Reward prediction error

New estimate of value of current state

Old estimate of value of current state

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

Reward prediction error

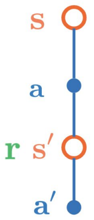
Actual observed value of current state (written the recursive way)

Old estimate of value of current state

Slide credit:
Jane Wang, Feryal Bebahani (@deepmind)

Learning the value function

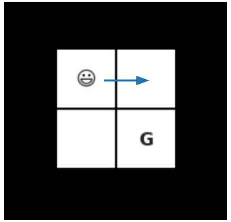
Problem: We can't predict the future! (And we don't want to…)



SARSA (on-policy control)

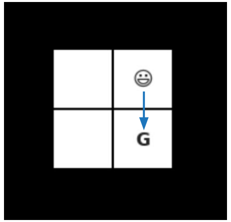
- Bootstrapping value updates based on "on-policy" (actual) experience

$$\underbrace{Q(s,a)}_{\text{new value}} \leftarrow \underbrace{Q(s,a)}_{\text{old value}} + \alpha\underbrace{(\overbrace{r}^{\text{reward}} + \gamma\overbrace{Q(s',a')}^{\text{Value next state}} - \overbrace{Q(s,a)}^{\text{Old value}})}_{\text{RPE}}$$

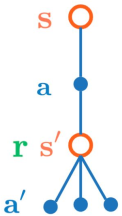


All Q's=0

Q(s0,right) <- 0+α (0+γ*0-0) = 0



Q(s1,down) <- 0+α (10+γ*0-0) = 5

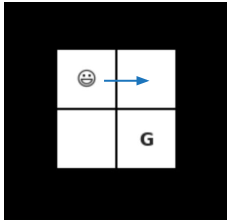


Q-learning (off-policy control)

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

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma\overbrace{\max_{a'} Q(s',a')}^{\text{Best avail. action}} - Q(s,a))$$

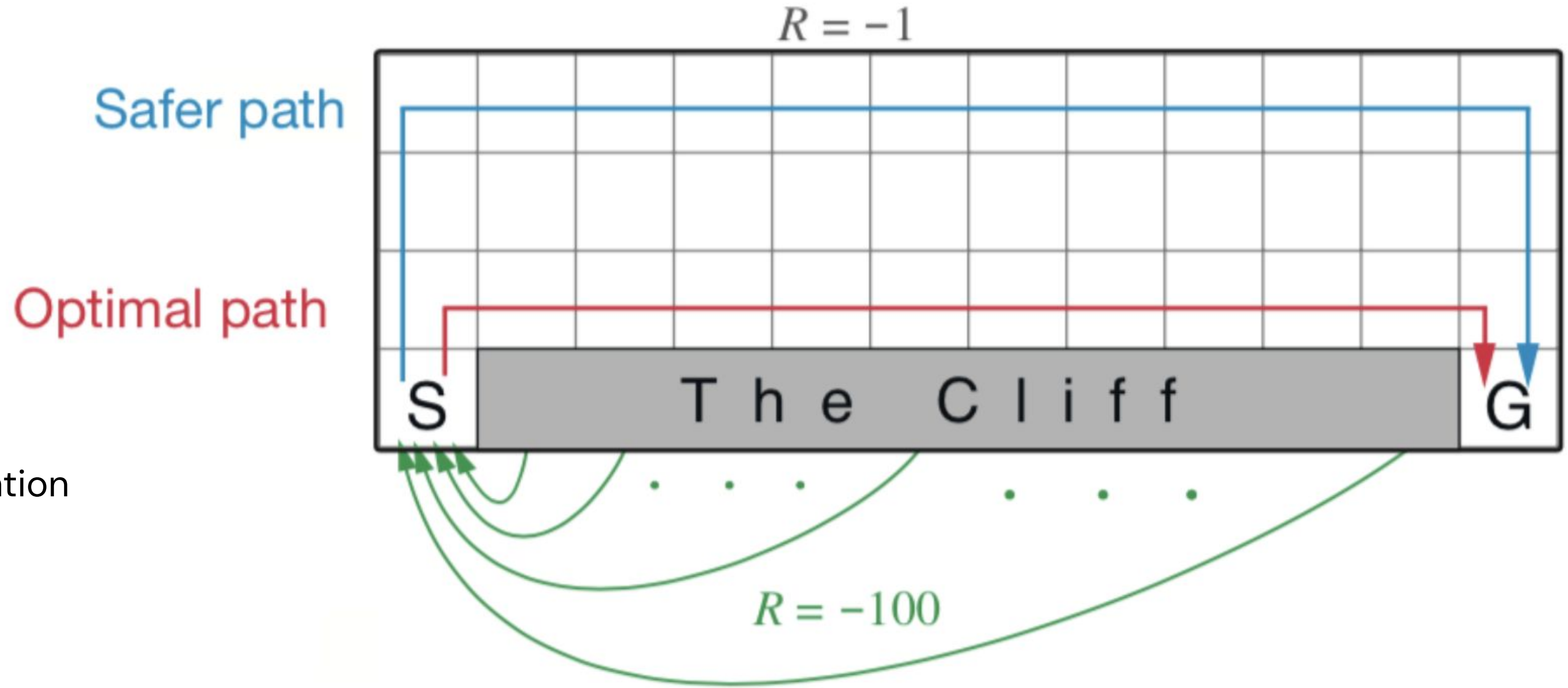
$$\pi^*(s) = \max_{a} Q^*(s,a)$$



Q(s0,right) <- 0+α (0+γ*5-0) = 2.25

Credit:
Maria Eckstein
(mariaeckstein@deepmind.com)
Jane Wang, Feryal Bebahani (@deepmind)

SARSA vs Q-Learning: The Cliff-walking Example



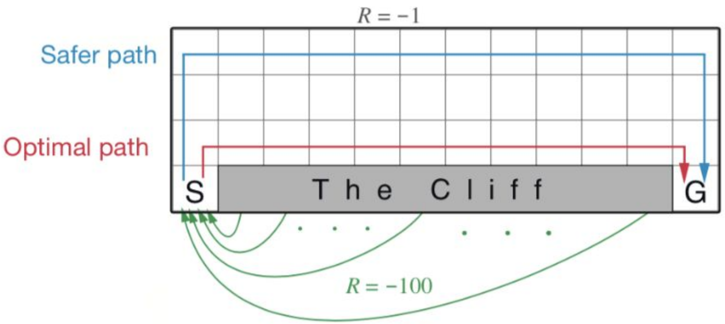
Exploration

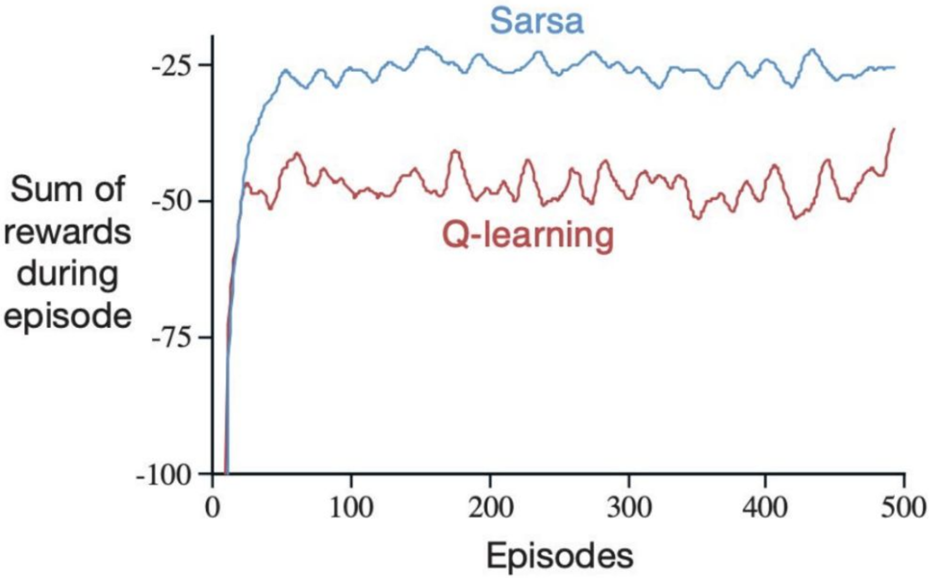
Sutton & Barto. Reinforcement Learning: An Introduction. (Chapter 6)

Slide credit:
Jane Wang, Feryal Bebahani
(@deepmind)

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**.

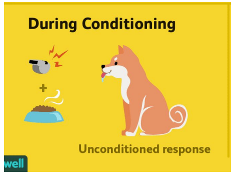




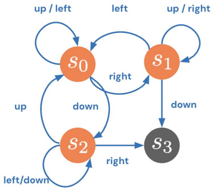
Slide credit:
Jane Wang, Feryal Behbahani
(@deepmind)

Cheat Sheet

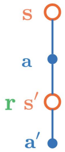
Rescorla Wagner: keep track of reward expectations



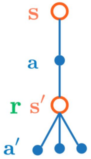
TD Learning: +over time



SARSA: +control (on-policy)



Q-Learning: +control (off-policy)



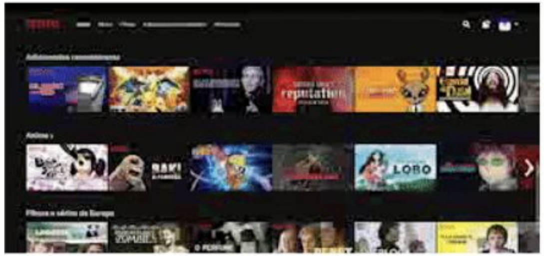
…redit:
Maria Eckstein, Kim Stachenfeld
(mariaeckstein@deepmind.com,
stachenfeld@deepmind.com)

Real-world Reinforcement Learning: **Examples**

Game playing

Robotics / Manipulation

User personalization



Self-driving cars

Managing energy usage

Slide credit:
Jane Wang
(wangjane@deepmind.com)

Questions?

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Reinforcement Learning (RL)

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

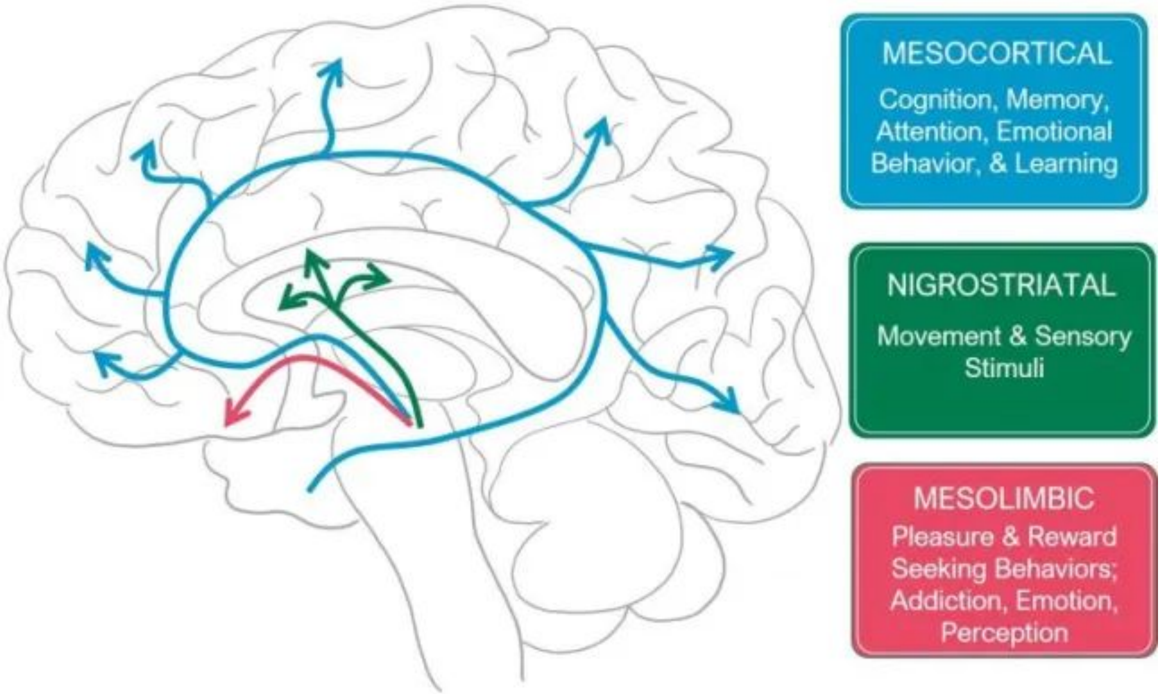
Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

DeepMind

RL in neuroscience

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

The Neurotransmitter Dopamine

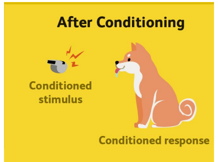


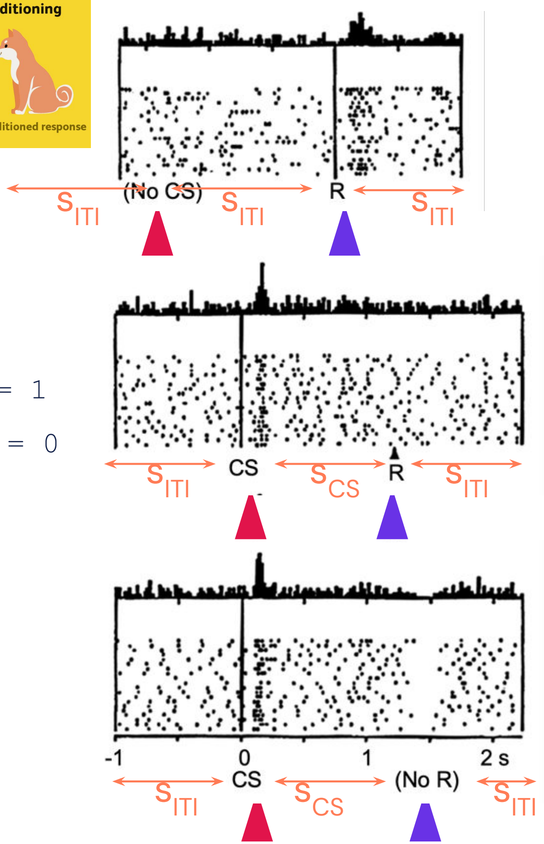
Essential for theory of reinforcement learning!

Dopamine Reward Prediction Errors

Quizz: What does dopamine firing represent?

TD RPE = $r + \gamma V(s') - V(s)$





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

Values:

$V(s_{CS}) = 1$

$V(s_{ITI}) = 0$

$$\text{RPE} = r + \gamma V(s_{ITI}) - V(s_{ITI}) = 0 + \gamma\, 0 - 0 = 0$$

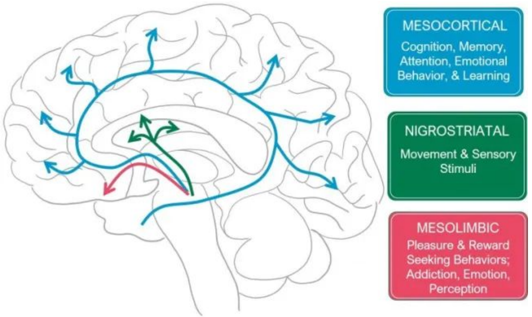
$$\text{RPE} = r + \gamma V(s_{ITI}) - V(s_{ITI}) = 1 + \gamma\, 0 - 0 = 1$$

$$\text{RPE} = r + \gamma V(s_{CS}) - V(s_{ITI}) = 0 + \gamma\, 1 - 0 = 0.9$$

$$\text{RPE} = r + \gamma V(s_{ITI}) - V(s_{CS}) = 1 + \gamma\, 0 - 1 = 0$$

$$\text{RPE} = r + \gamma V(s_{CS}) - V(s_{ITI}) = 0 + \gamma\, 1 - 0 = 0.9$$

$$\text{RPE} = r + \gamma V(s_{ITI}) - V(s_{CS}) = 0 + \gamma\, 0 - 1 = -1$$

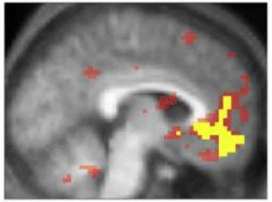


- Converging evidence across studies and species
- Mostly in simple conditioning paradigms

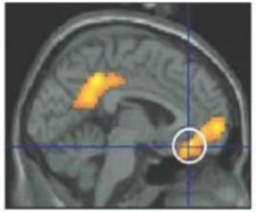
[Niv, 2009]

Credit: Maria Eckstein (mariaeckstein@deepmind.com)

Human fMRI

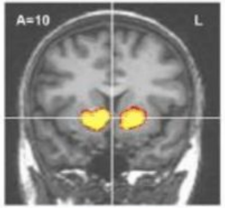


money
value predicted
(Daw et al 2006)

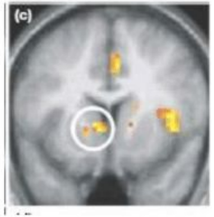


faces
attractiveness
(O' Doherty et al 2003)

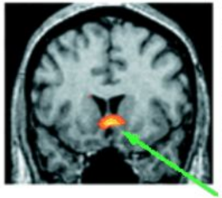
Coke or Pepsi
degree favored
(McClure et al. 2004)



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?*

Slide credit:
Maria Eckstein, Anne Collins
(mariaeckstein@deepmind.com,
annecollins@berkeley.edu)

Questions?

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Reinforcement Learning (RL)

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

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

RL for Cognitive Modeling

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

What is Cognitive Modeling?

RL

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

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

Goal: Understand behavior, cognitive process

Method:

- Find model (e.g., RL, Regression, DDM, …)
- "Fit" model (find best parameters, using cross-entropy loss / negative log likelihood)
- Expand model
  - e.g., forgetting; 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

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

What is RL Modeling?

Goal | Reward | Ingredients | Algorithm

+1

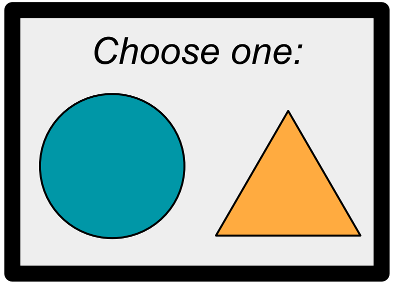
action = [→, ←]

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 = []

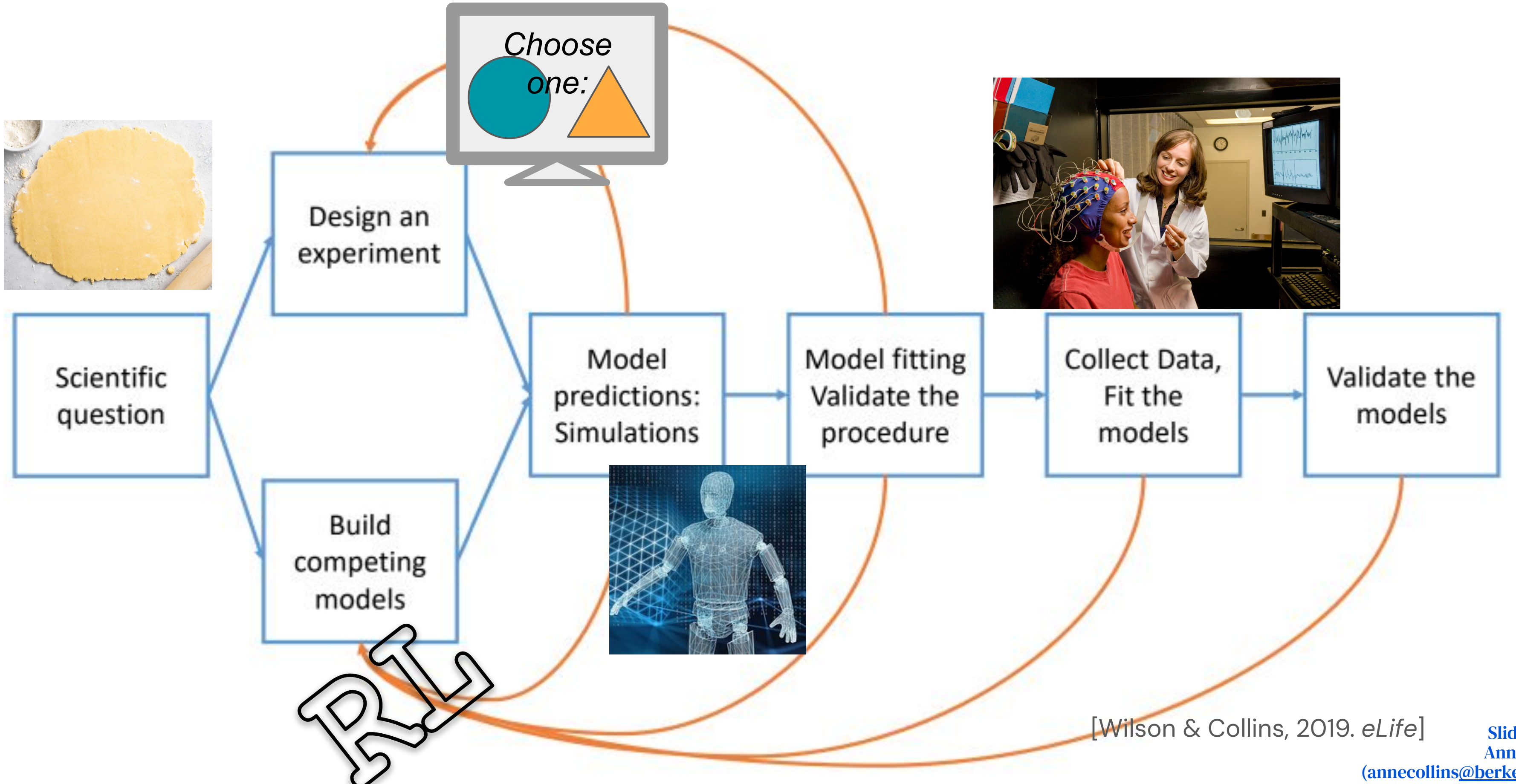
reward = [0, +1]

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

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

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

A Recipe for Cognitive Modeling

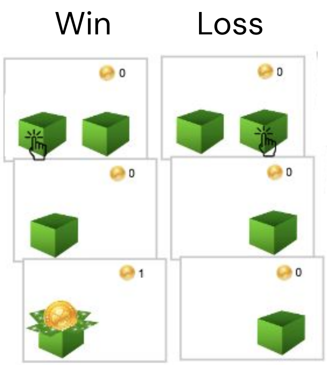


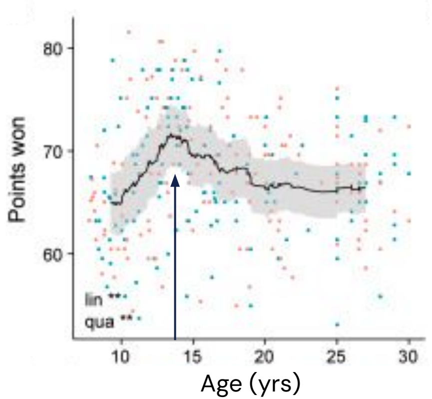
[Wilson & Collins, 2019. *eLife*]

Slide credit: Anne Collins (annecollins@berkeley.edu)

Learning to Reversal Learn

Goal: Understand age trajectory of reversal learning



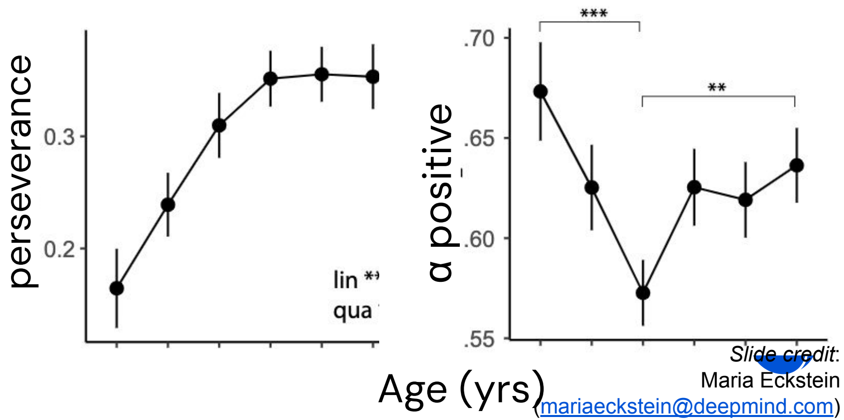


- Best performance at ~13-15

Why? Cognitive mechanism?

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

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



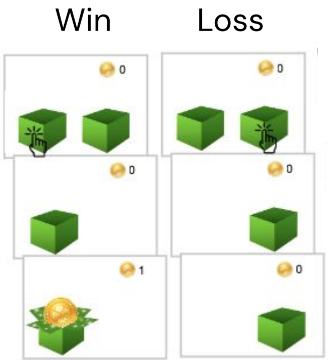
Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

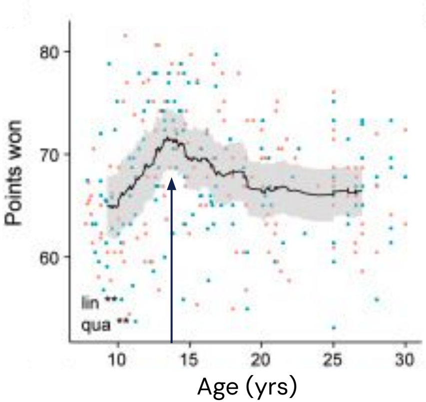
[Eckstein, Master, Dahl, Wilbrecht & Collins, 2022. *DCN*]

Learning to Reversal Learn

$$\mathbf{p(s_t|a_t, r_t)} \propto p(a_t, r_t|s_t) * p(s_t)$$

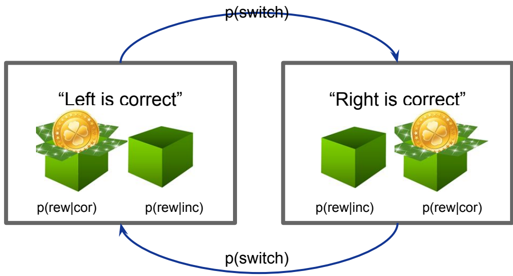
Goal: Understand age trajectory of reversal learning

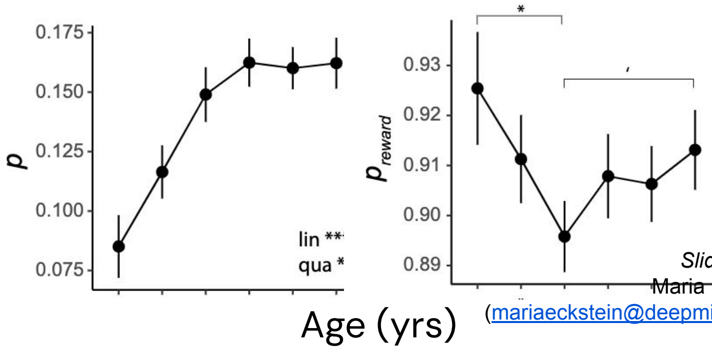




- Best performance at ~13–15

Why? Cognitive mechanism?

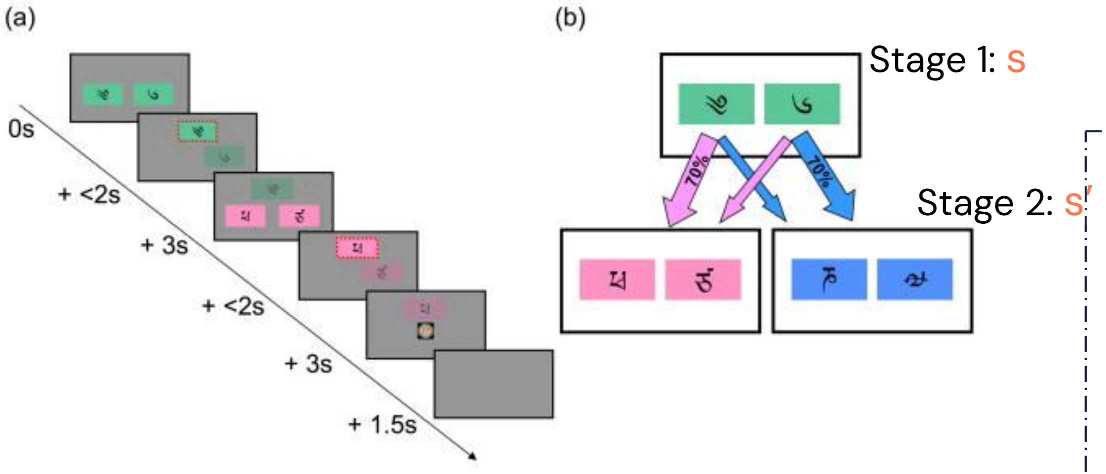


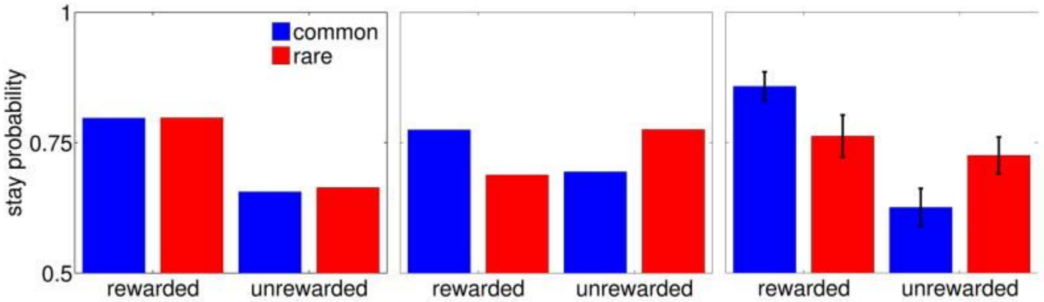


[Eckstein, Master, Dahl, Wilbrecht & Collins, 2022. *DCN*]

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Model-based or model-free RL?





Model-free: SARSA

At both stages:

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

$$Q_{MF}(s,a) \leftarrow Q(s,a) + \alpha * RPE$$

Model-based

Stage 2:

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

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

Stage 1:

$$Q_{MB}(s,a) = p(s_A'|s,a) * \max_a Q(s_A',a') + p(s_B'|s,a) * \max_a Q(s_B',a')$$

Hybrid

$$Q(s,a) = w * Q_{MF}(s,a) + (1 - w) * Q_{MB}(s,a)$$

Slide credit: Maria Eckstein (mariaeckstein@deepmind.com)

Model-based or model-free RL?

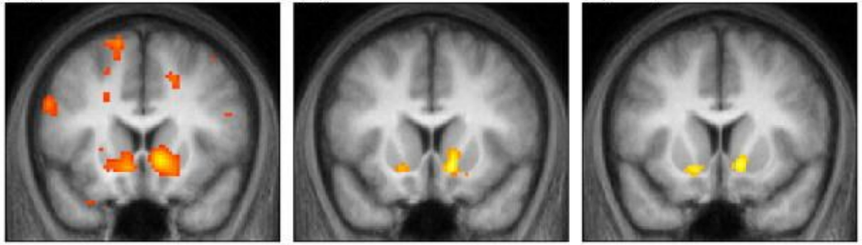
Results:

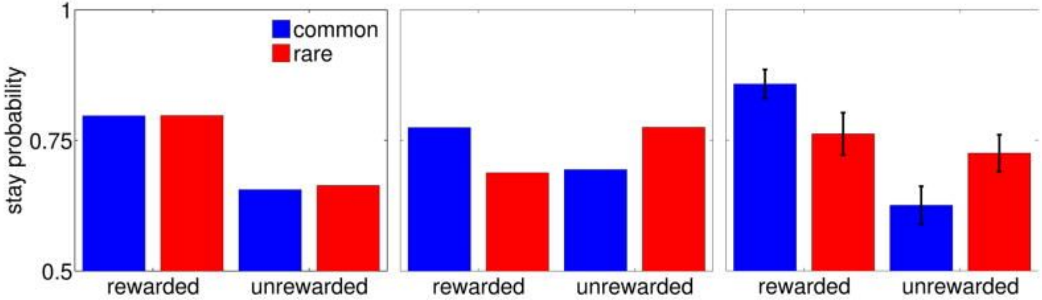
- Hybrid models (LL=3.364) fits data better than MF alone (LL=3.418) or MB alone (LL=3.501)
- Fitted value of w (median across subjects): 0.39

MF RPE

MB difference regressor

Conjunction MF and MB





Model-free: SARSA

At both stages:

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

$$Q_{MF}(s,a) \leftarrow Q(s,a) + \alpha * RPE$$

Model-based

Stage 2:

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

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

Stage 1:

$$Q_{MB}(s,a) = p(s_A'|s,a) * \max_a Q(s_A',a') + p(s_B'|s,a) * \max_a Q(s_B',a')$$

Hybrid

$$Q(s,a) = w * Q_{MF}(s,a) + (1 - w) * Q_{MB}(s,a)$$

Slide credit: Maria Eckstein (mariaeckstein@deepmind.com)

Questions?

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Reinforcement Learning (RL)

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**

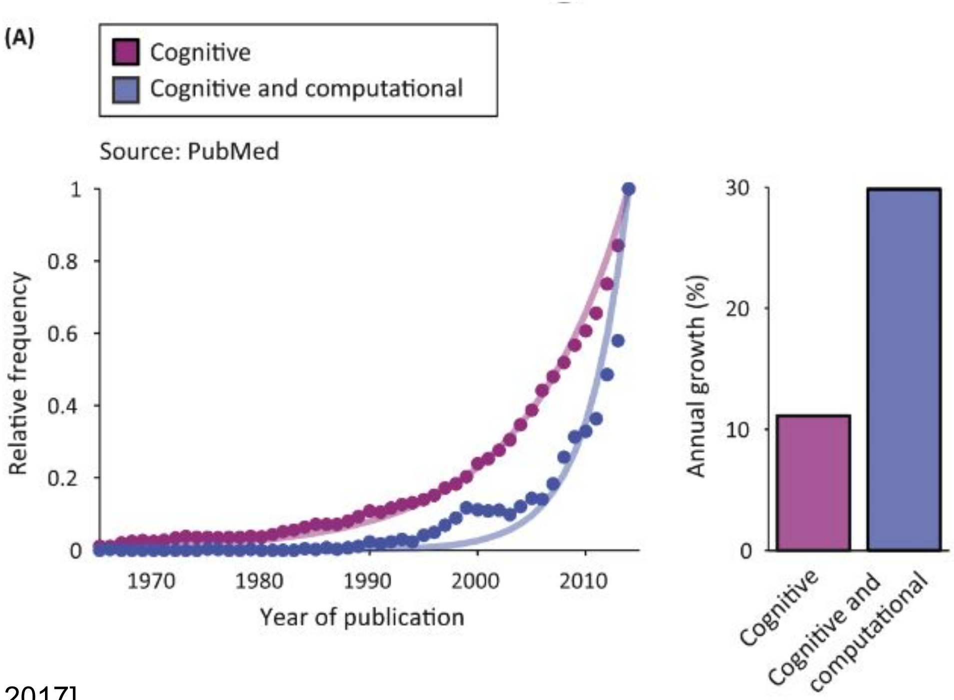
Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

DeepMind

Conclusion

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Computational modeling is on the rise!

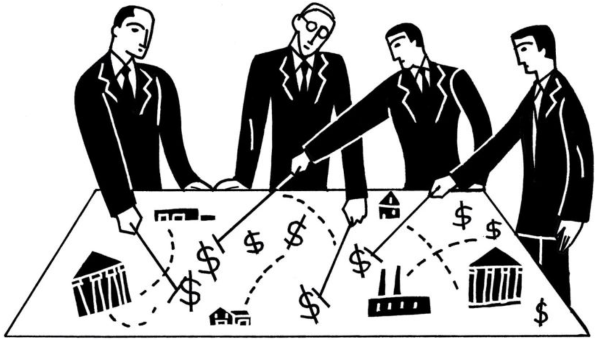


[Palminter et al., 2017]

Slide credit:
Anne Collins
(annecollins@berkeley.edu)

Where do rewards come from?

Evolution?

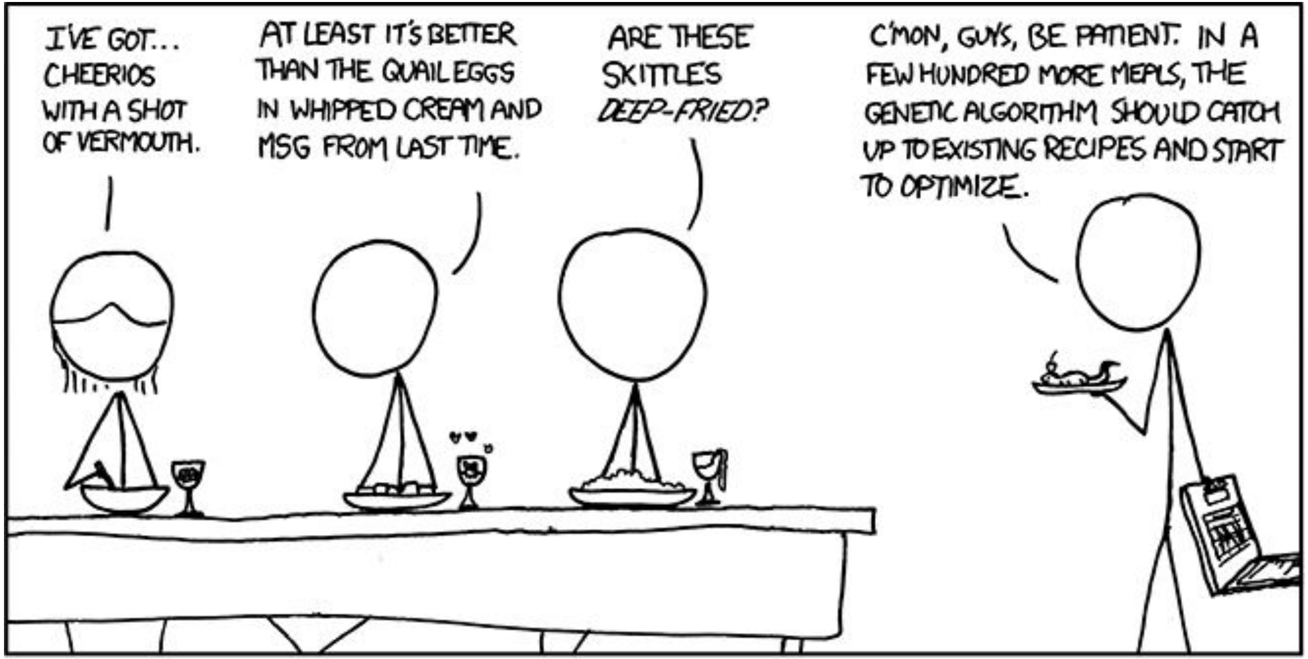


Economists?

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

Slide credit:
Kim Stachenfeld
(stachenfeld@deepmind.com)

Exploration



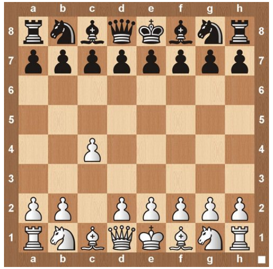
WE'VE DECIDED TO DROP THE CS DEPARTMENT FROM OUR WEEKLY DINNER PARTY HOSTING ROTATION.

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

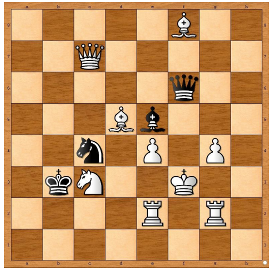
Slide credit:
Kim Stachenfeld
(stachenfeld@deepmind.com)

Credit Assignment

Beginning



End



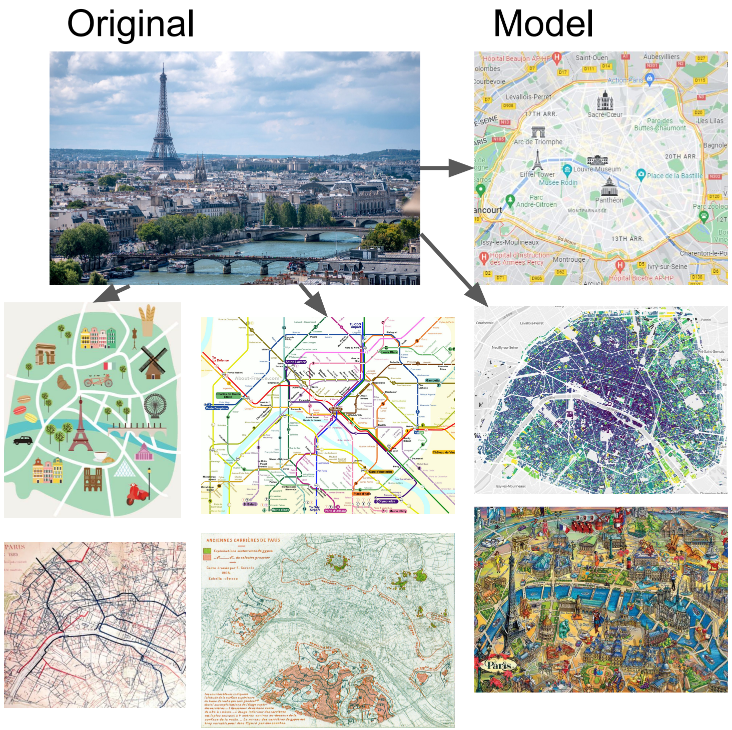
How to link distal outcomes to earlier causes despite many intervening events?

How to generalize over similar + different instances?

How to use knowledge of structure inform credit assignment?

Slide credit:
Kim Stachenfeld
(stachenfeld@deepmind.com)

Models as Maps



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

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Questions?

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

Want to Learn More?

Books

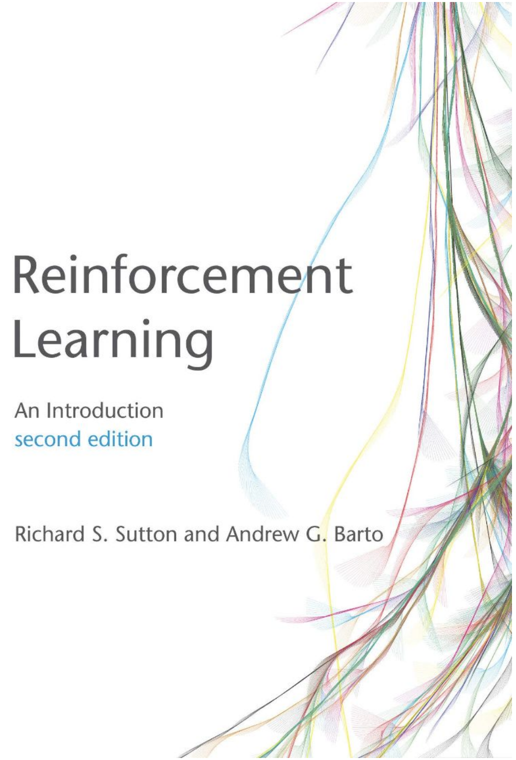
- Reinforcement Learning: an Introduction by Sutton & Barto
- Algorithms for Reinforcement Learning by Csaba Szepesvari

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



Slide credit:
Maria Eckstein, Kim Stachenfeld
(mariaeckstein@deepmind.com,
stachenfeld@deepmind.com)

Acknowledgements

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

credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

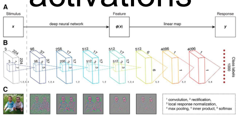

Dog
tricks
by
Justy

DeepMind

Bonus

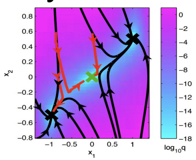
Theory-driven vs Data-driven Models

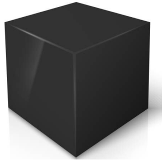
Analyze activations

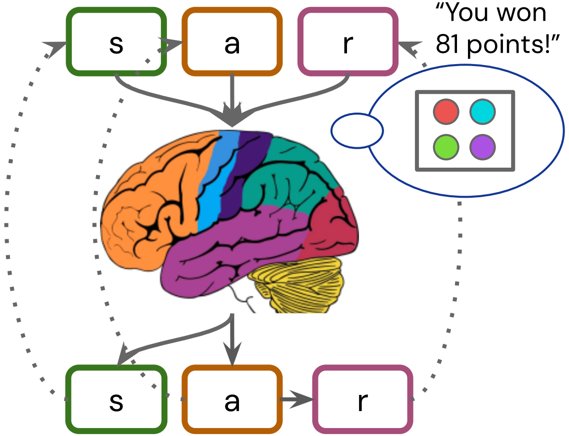


Explainability…

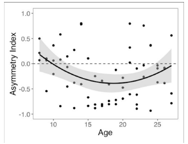
Analyze dynamics







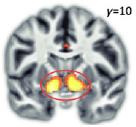
Cognitive development



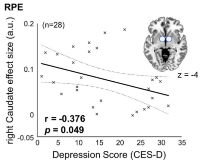
[van den Bos et al., 2012; Lefebvre et al., 2017; Nussenbaum & Hartley, 2019; Master et al., 2020; Eckstein et al., 2022]

Brain function

[Daw et al., 2006; O'Doherty et al., 2007; Dayan & Niv, 2008; Miller et al., 2017; Starkweather et al., 2018]



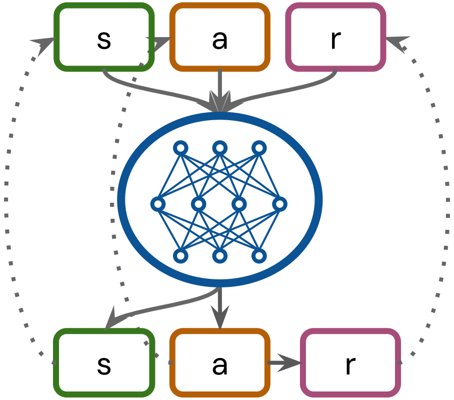
Psychiatry

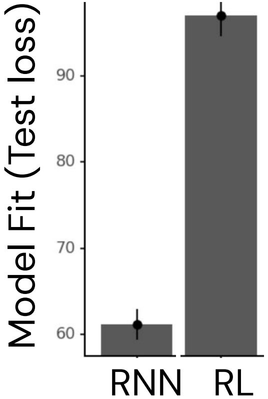


[Maia & Frank, 2011; Montague et al., 2012; Huys et al., 2016; Redish & Gordon, 2016; Hauser et al., 2019]

Abstraction, Model-based, Habits, Exploration, Sequences, …

Vanilla RNN





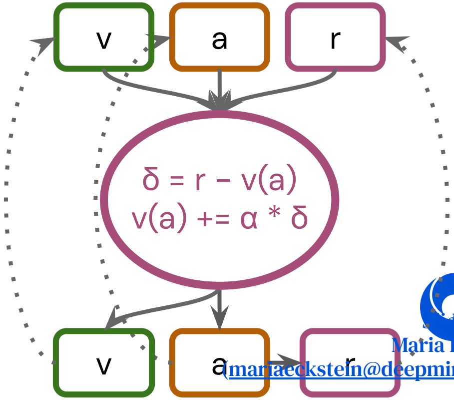
Trade-offs

- Predictive power (RNN) vs Interpretability (RL)
- What makes a good model?
  [Navarro, 2019; Box, 1979; Eckstein et al., 2021]

Uncover the cognitive process

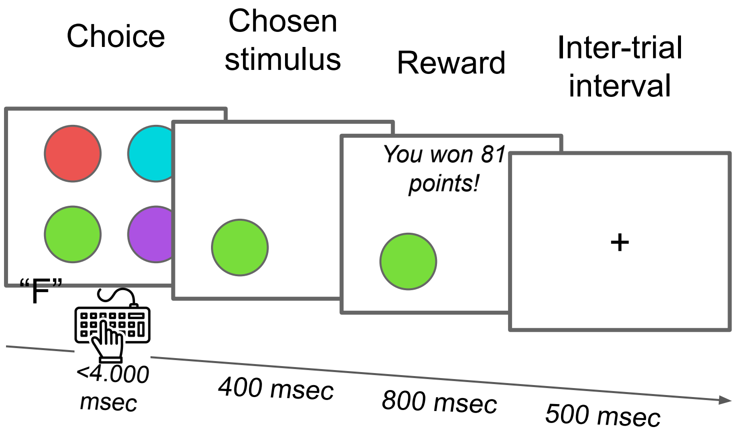
- *Why* is RL underperforming?
- *Which* cognitive processes are missing?
- *Which* assumptions are wrong?

Classic RL



Credit: Maria Eckstein (mariaeckstein@deepmind.com)

Dataset



Key points

- 4-armed bandit
- Arms drift independently

Original task

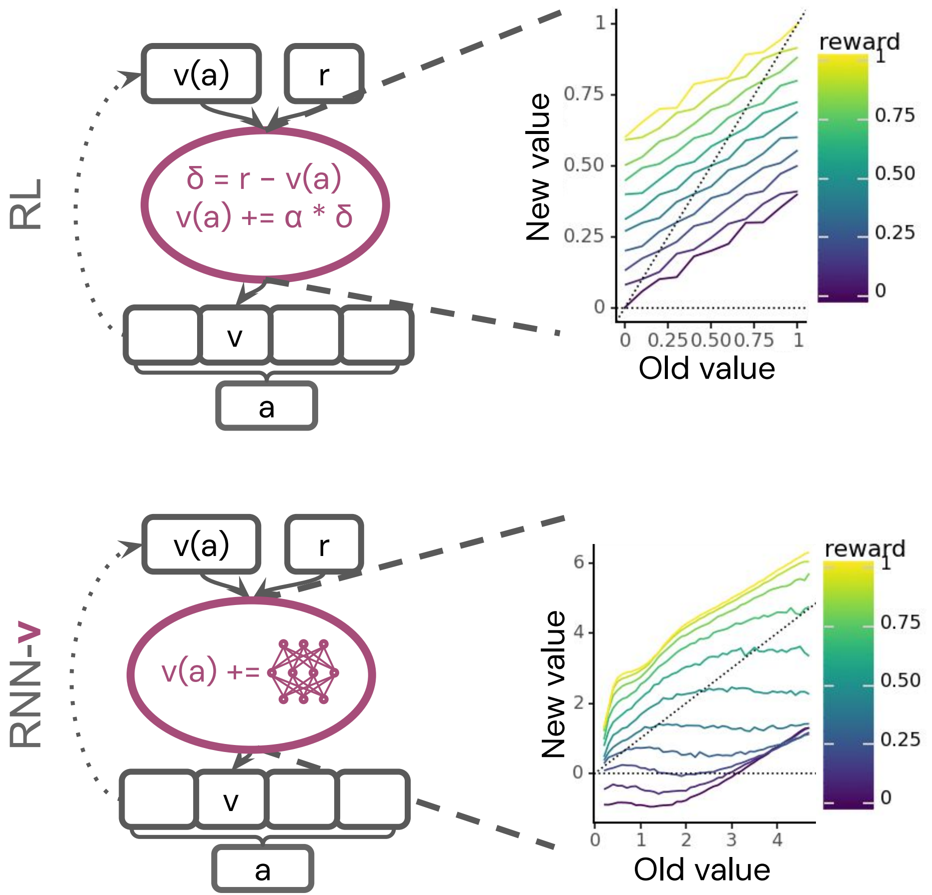
- 14 participants, 150 trials, fMRI [Daw et al., 2006]

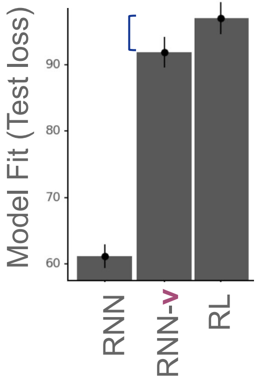
Our version

- 880 participants
- Several blocks (1 training block, several testing blocks à 150 trials)
- Online (prolific)
- Exclusion: 2% of participants, 0.6% of blocks

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

A Different Value Update, Learned from Data





Conclusion

- RNN-**v** fits better than RL
- Human learning is different from pure RL theory
- But still a big gap in model fit
- Test other assumptions of RL

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

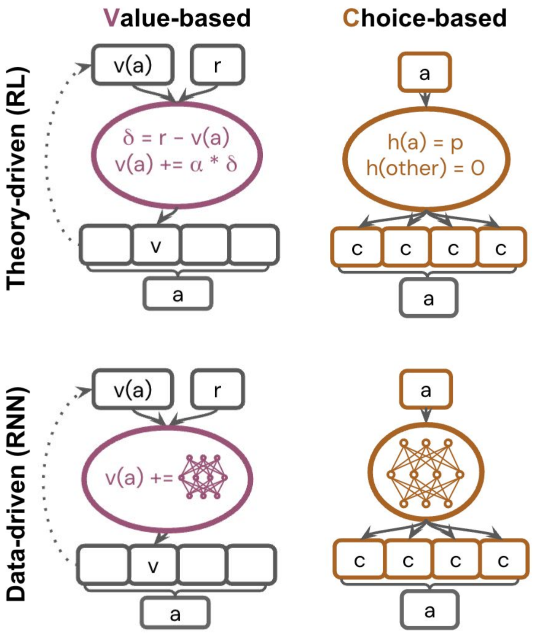
[Eckstein, Daw, Summerfield, & Miller, 2023, *CogSci*]

Testing other assumptions of RL

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)

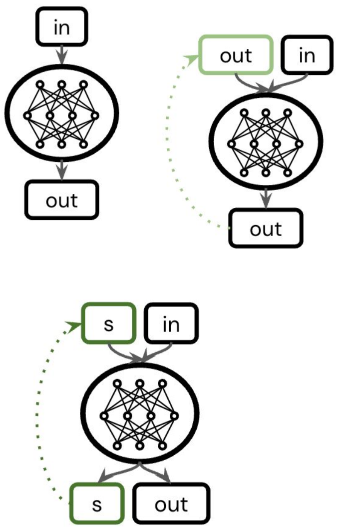
Reward-independent processes

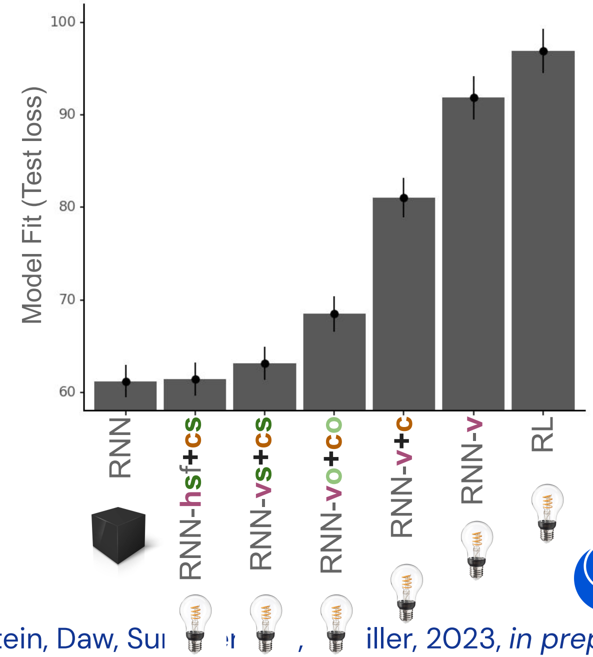
[e.g., Gillan et al., 2015; Miller et al., 2019; Sugawara & Katahira, 2021; ...]



Memory / Context

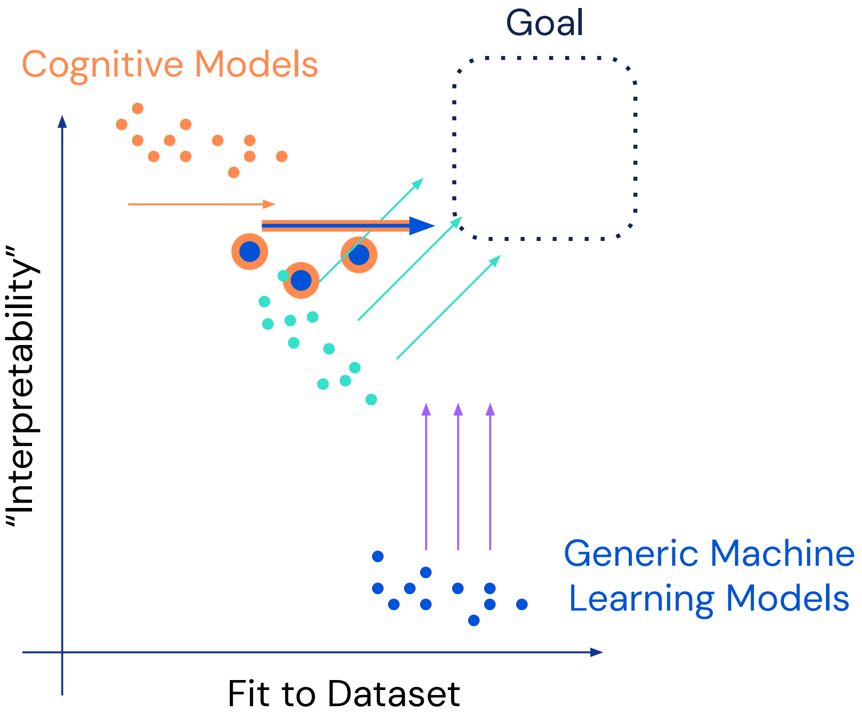
[e.g., Collins & Frank, 2012; Palminteri et al., 2015; Davidow et al., 2016; Gershman & Daw, 2017; Wang et al., 2018; Ramani, 2019; ...]





[Eckstein, Daw, Su ... iller, 2023, *in prep*]

Conclusions: A Landscape of Possibilities



- Quantitative models of behavior: A key tool for Comp. Cog. Neuro.

- Classic Cognitive modeling
- ML models as benchmarks
- ML models for post-hoc interpretability
- Interpretability-encouraging architectures
- Hybrid models

- Combining them!

- Your ideas?

redit:
Maria Eckstein, Kevin Miller
(mariaeckstein@deepmind.com,
kevinjmiller@deepmind.com)

Acknowledgements

Slides:

Anne Collins

Kim Stachenfeld

Collaborators at GDM:

Zeb Kurth-Nelson

Kevin Miller

Nathaniel Daw

Chris Summerfield

Slide credit:
Maria Eckstein
(mariaeckstein@deepmind.com)