

# Reinforcement Learning

Evan Russek

Max Planck UCL Centre for Computational Psychiatry and Ageing Research, UCL

email: [e.russek@ucl.ac.uk](mailto:e.russek@ucl.ac.uk)

# What is Reinforcement Learning?

- Computer science: algorithms for using experiences to make choices that maximize rewards
- Psychology / Neuroscience
  - How do we make choices given experiences?
  - What role do different neural structures / circuits / neuromodulators play in converting experience to choice?
  - Why is choice sometimes flexible / inflexible?
  - “Meta-reasoning”: What computations do we do and when? What should we think about?

# Reinforcement Learning in Psychiatry

- Why might certain experiences cause certain choices characteristic of psychiatric conditions?
- How might alterations to neuromodulators, perhaps by drugs, alter choices?
- Inflexible / maladaptive choices in psychiatric conditions
- Why might we think about things we don't want to think about?

# Plan

- Part 1: Dual systems for choice
  - Reward Prediction errors, Dopamine
  - Predicting rewards through time, Model-free RL, Habits
  - Flexible choice, Model-based RL
- Part 2: A more fine grained view - approximate planning
  - Tree-search, Pruning
  - Temporal Abstraction, Successor representations
  - DYNA, Prioritized Simulation

# Reward prediction error (RPE) updating

Choice Experienced rewards (r)



4,2,5,4,3

$R(\text{burrito})$



3,1,4,2,3

$R(\text{burrito in basket})$

Prediction:

$R(\text{burrito}) = 0$

Target:

$r = 4$

Prediction error: Target - Prediction

$4 - 0 = 4$

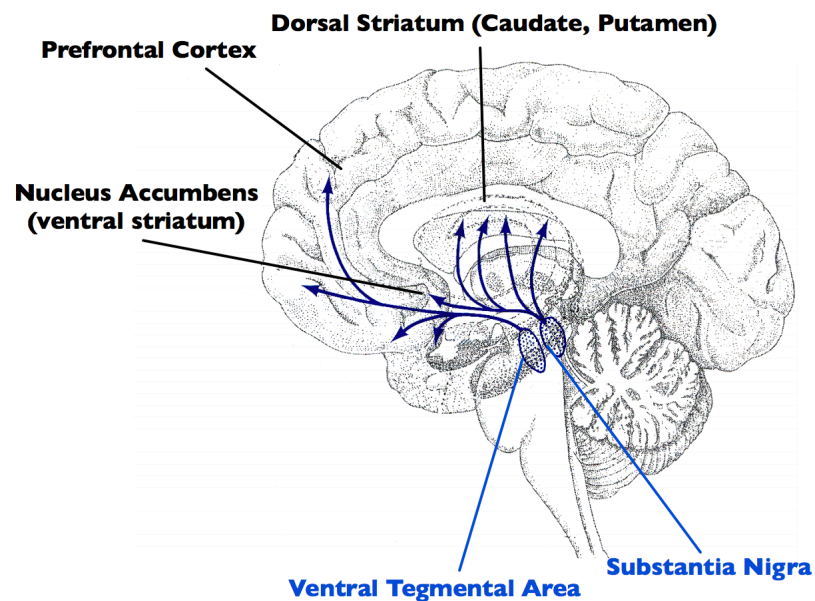
new prediction <-

old prediction +  $\alpha$  x prediction error

$R(\text{burrito}) \leftarrow 0 + .5 \times 4$   
 $= 2$

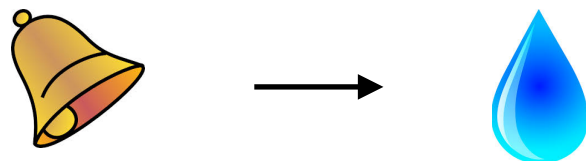
# Dopamine as RPE

## Dopamine (DA) system:



## Conditioning:

Cue (CS)      Reward (US):

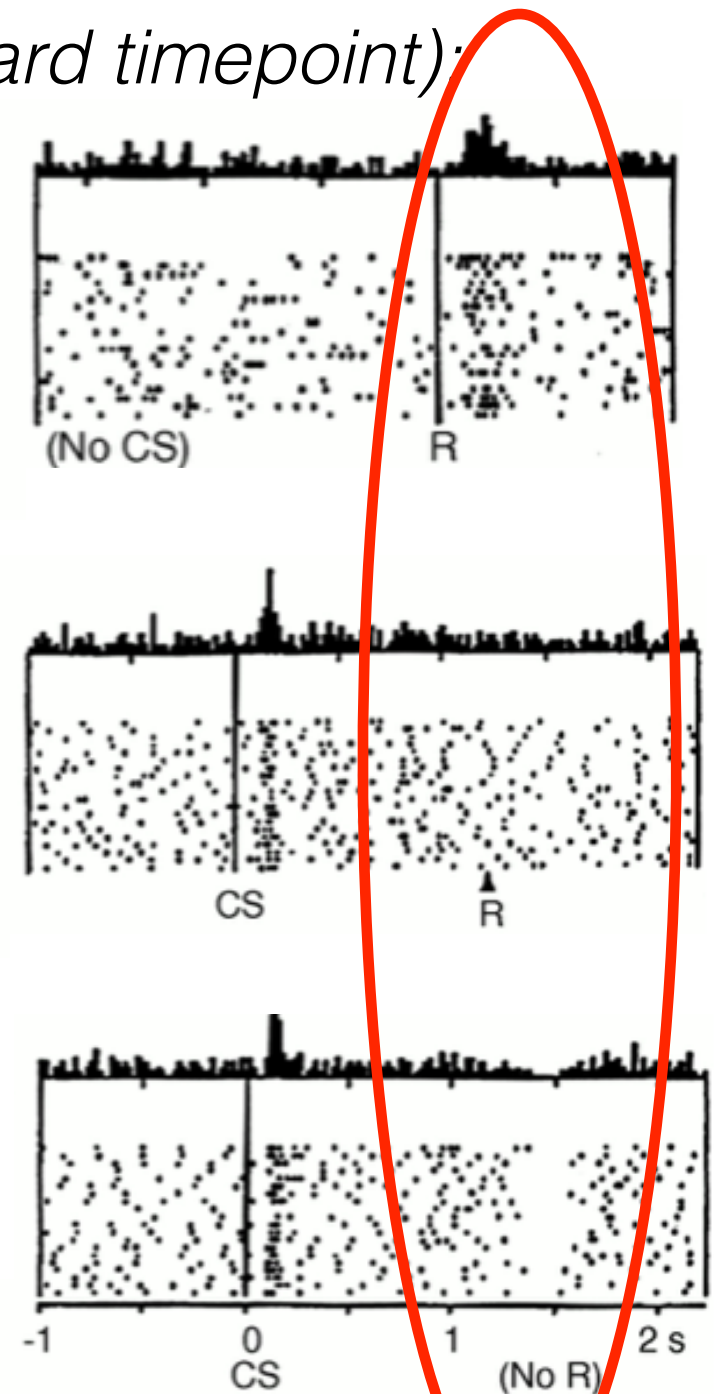


## DA Neurons (at reward timepoint):

reward > prediction

reward == prediction

reward < prediction

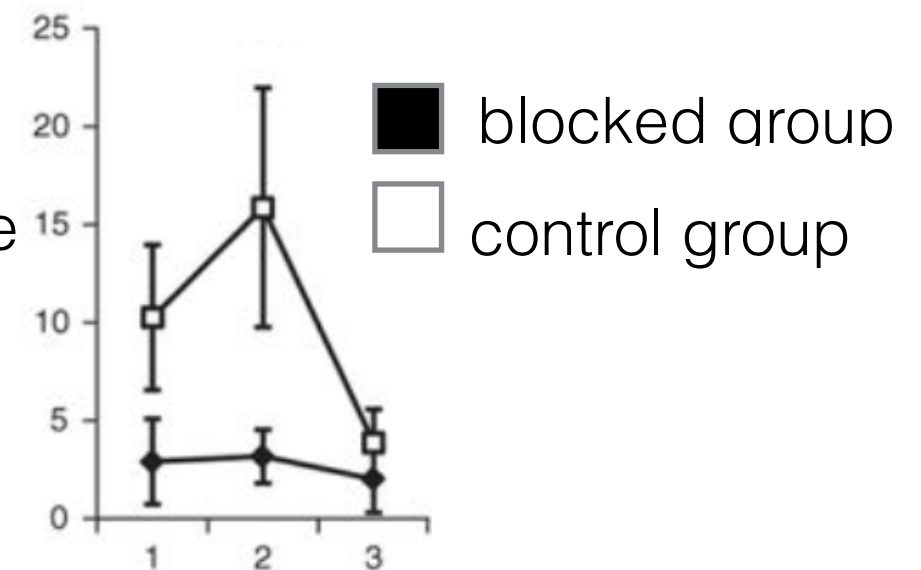


# Causal Evidence for DA as RPE

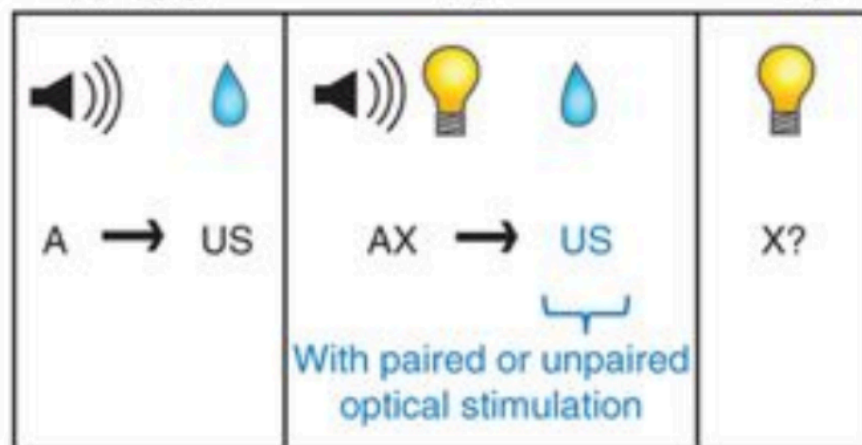
## Blocking Prevents New Reward Learning



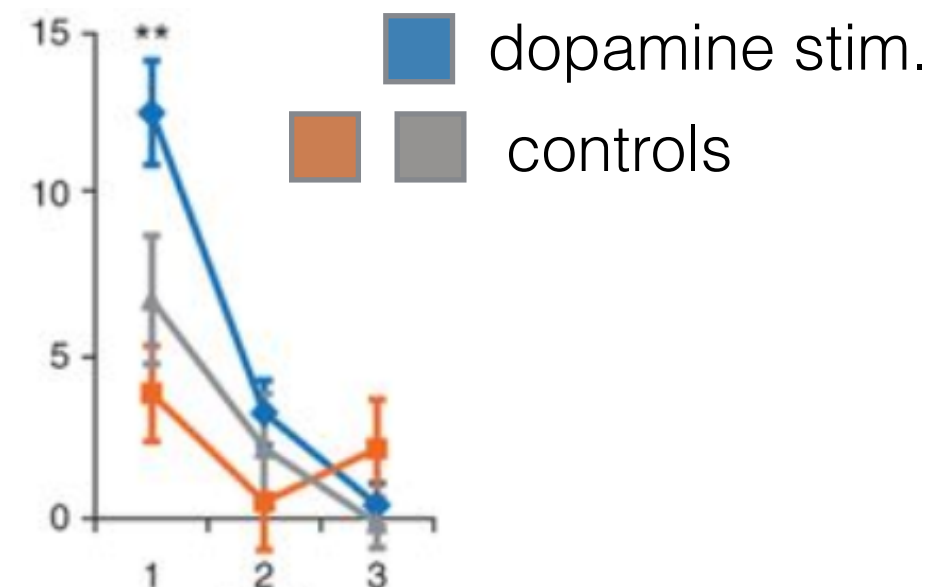
Response to X



## DA Stimulation Saves Learning



Response to X



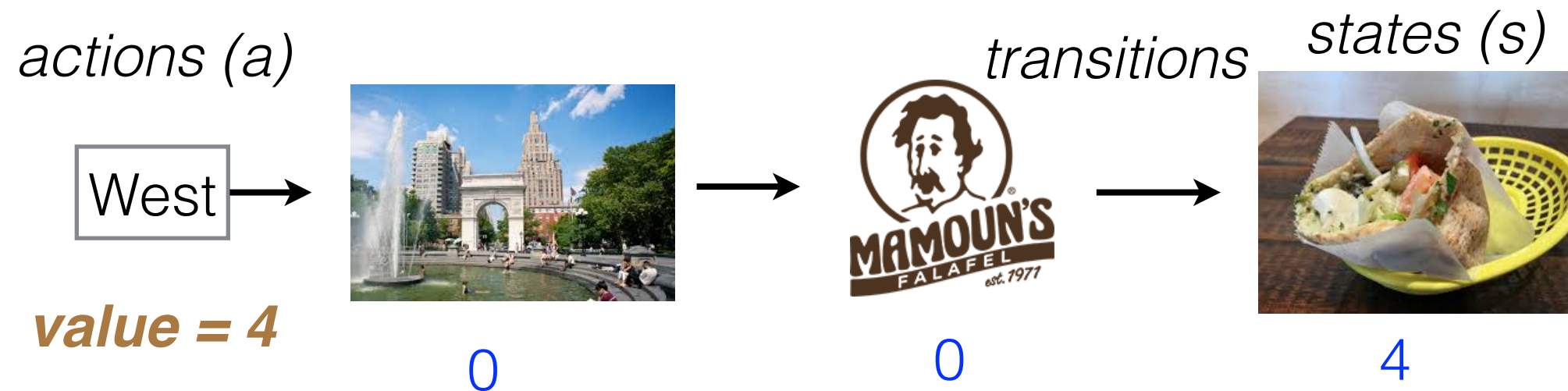
## RPE learning and Psychiatry

- Drug addiction:
  - Cocaine causes positive RPE? *(Redish, (2004), Science)*
  - But cocaine don't stop blocking *(Panillo, Thorndike, Schindler (2007), Pharm. Bio. Behav.)*
- Parkinsons
  - Impaired RPEs -> impaired movement values *(Niv, Rivlin-Etzion, (2007) J. Neuro)*
- Depression / Anhedonia
  - Altered reward learning in MDD *(Huys, Pizzagali, Bogdan, Dayan, (2013) Biol. Mood. Anx. Disorders)*
  - No differences in fMRI RPE response *(Rutledge, Moutoussis, ... ,Dolan (2017), JAMA Psych)*
  - Behavioral Activation Therapy *(poster w/ George Abitante\*, Jackie Gollan, Quentin Huys)*

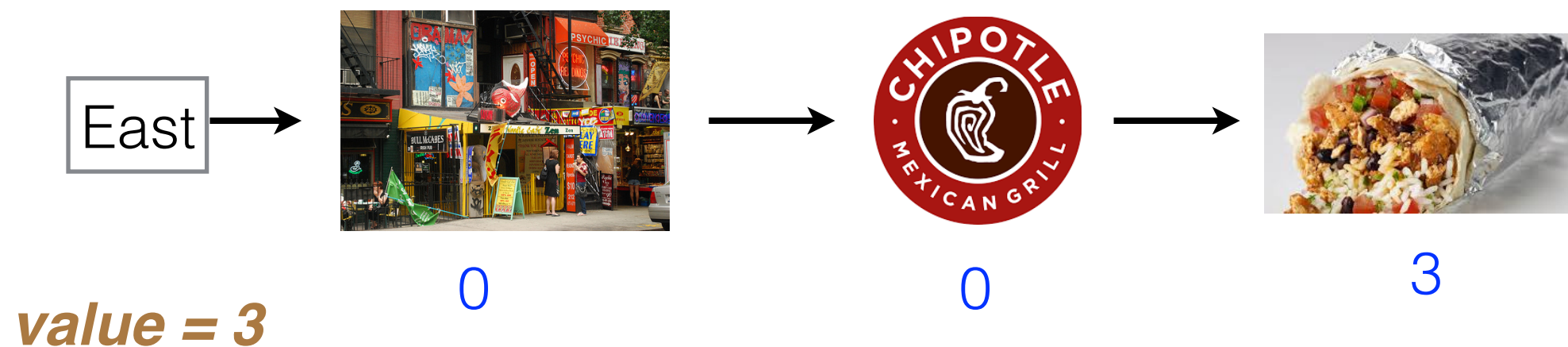


# The reinforcement learning problem

## ***The environment (MDP)***

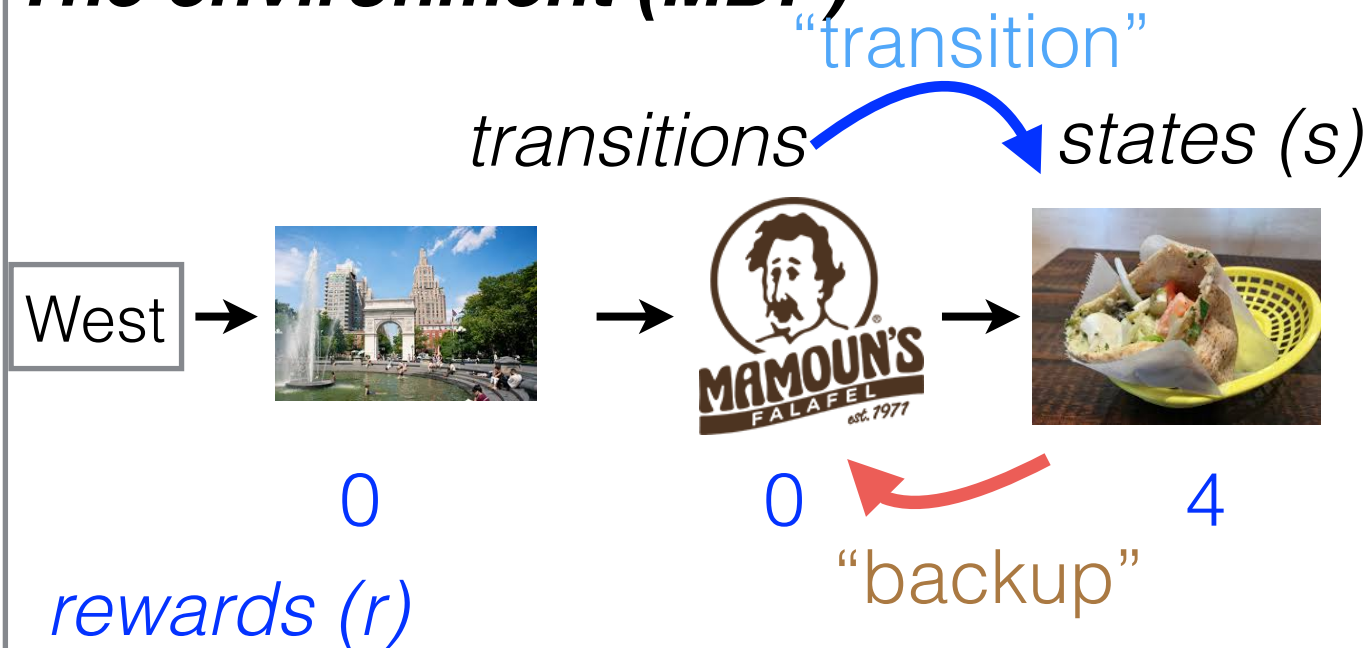


*value (V): cumulative rewards following a choice (or state)*      *rewards (r)*



# Model-free “Temporal Difference” Learning

## The environment (MDP)



## Bellman’s Equation:

$$V(s) = R(s) + V(s')$$

$$V(\text{MAMOUN'S FALAFEL}) = R(\text{MAMOUN'S FALAFEL}) + V(\text{Falafel Basket})$$

## TD update:

Prediction:  $V(\text{MAMOUN'S FALAFEL}) = 0$

Target:  $R(\text{MAMOUN'S FALAFEL}) + V(\text{Falafel Basket})$   
 $= 0 + 4$

prediction error: target - prediction

$$= 4 - 0$$

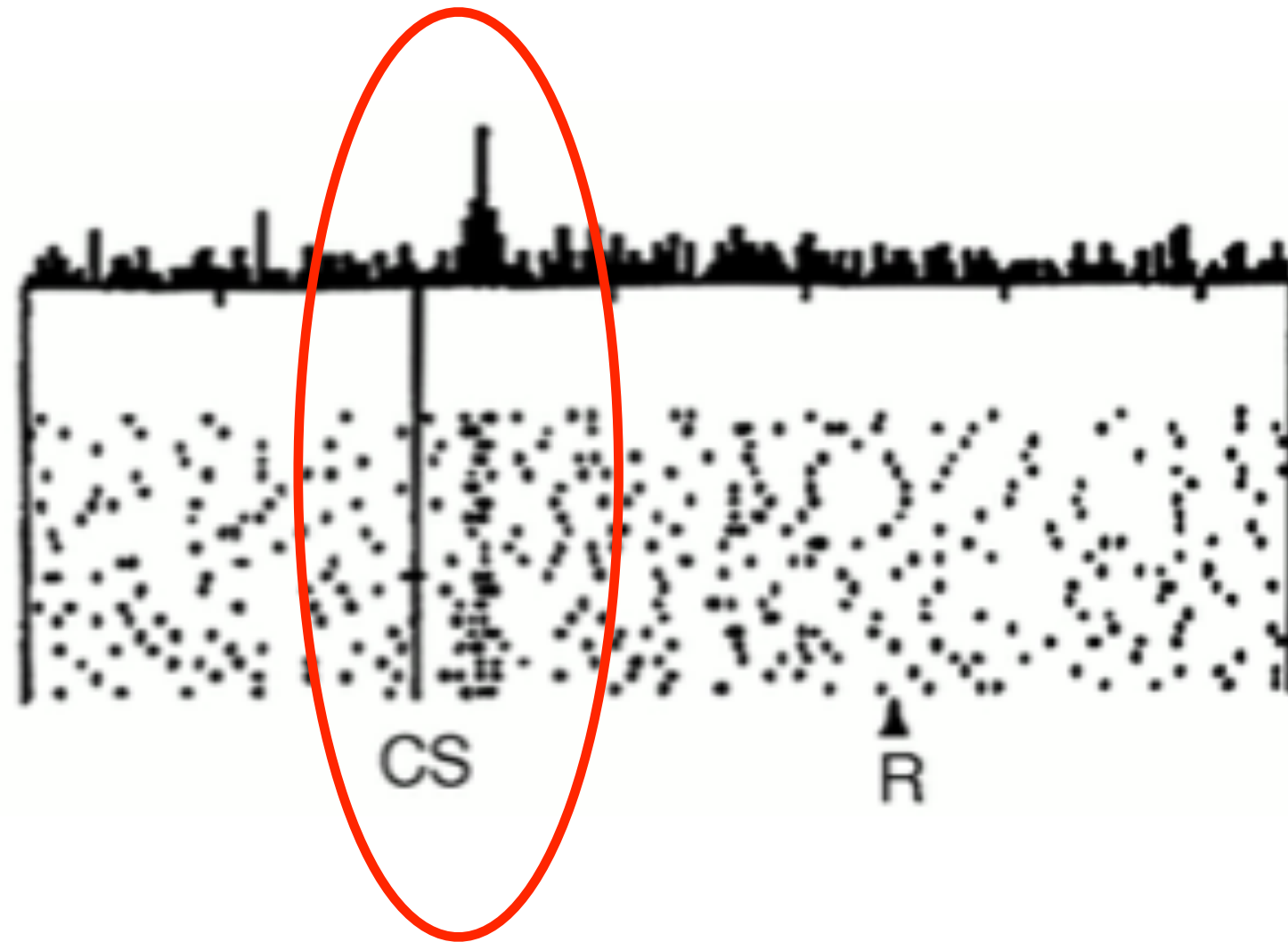
new prediction <-

old prediction +  $\alpha$  x prediction error

$$V(\text{MAMOUN'S FALAFEL}) <- 0 + .5 * 4$$

$$<- 2$$

# Dopamine as temporal difference prediction error



CS causes increase in  $V(s_t)$  relative to prior time-point

Also seen in humans (O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003), *Neuron*;

Mcclure, Burns, Montague (2003) *Neuron*), rodents (Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., & Uchida, N 2012; *Nature*)

# Model-based and Model-free RL

***model-free RL***

**Fast at decision time**

**Inflexible**

West

stored value

4

***model-based RL***

**Slow at decision time**

**Flexible**

West



0



0



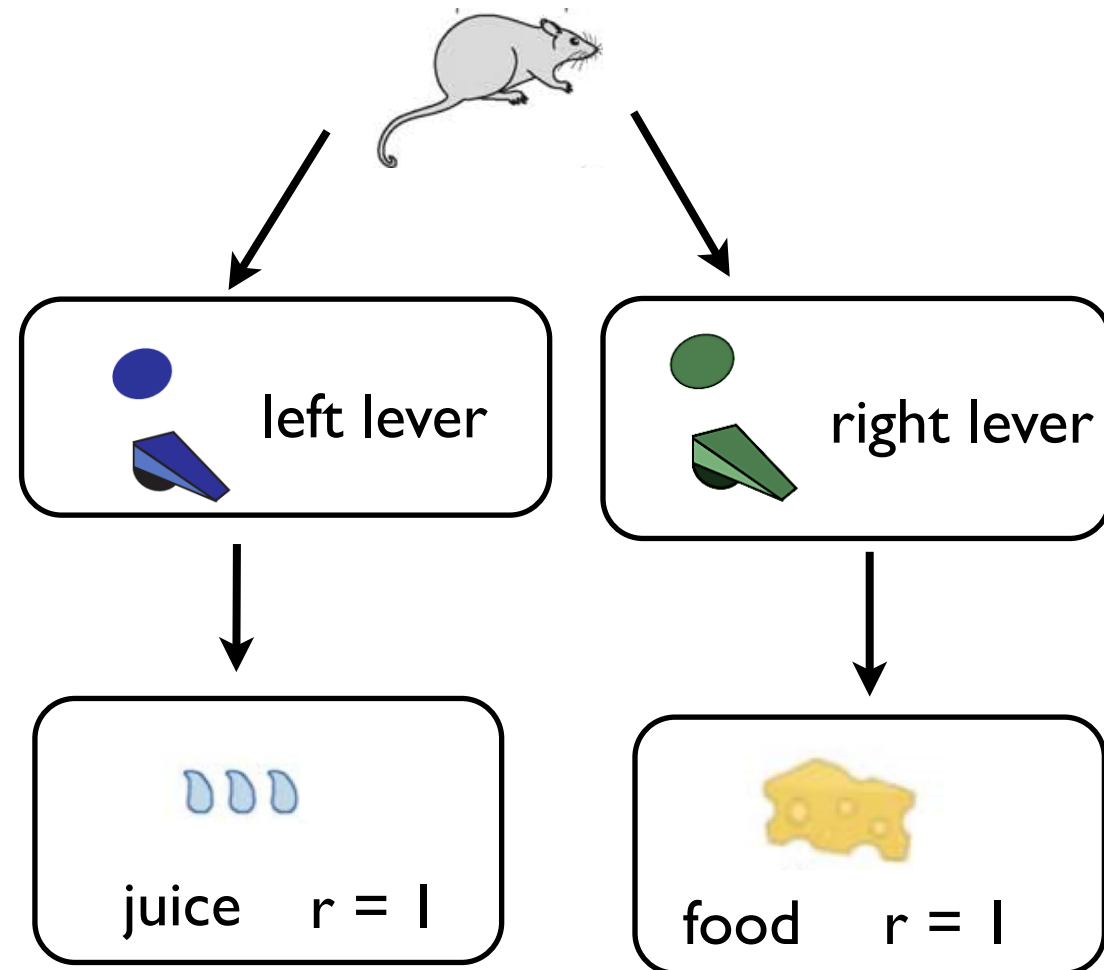
~~4~~ 0

*rewards (r)*

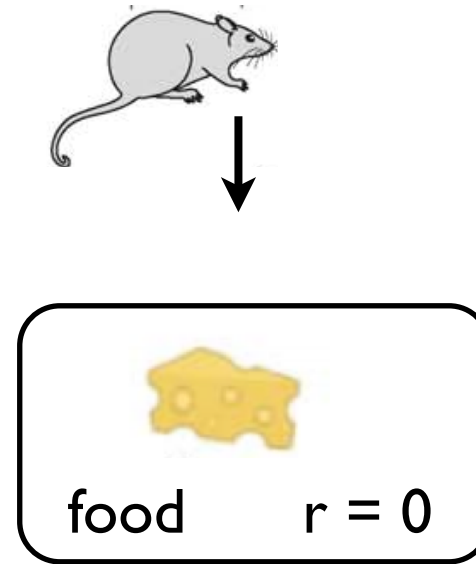
**reward  
change**

# Outcome Revaluation Paradigm

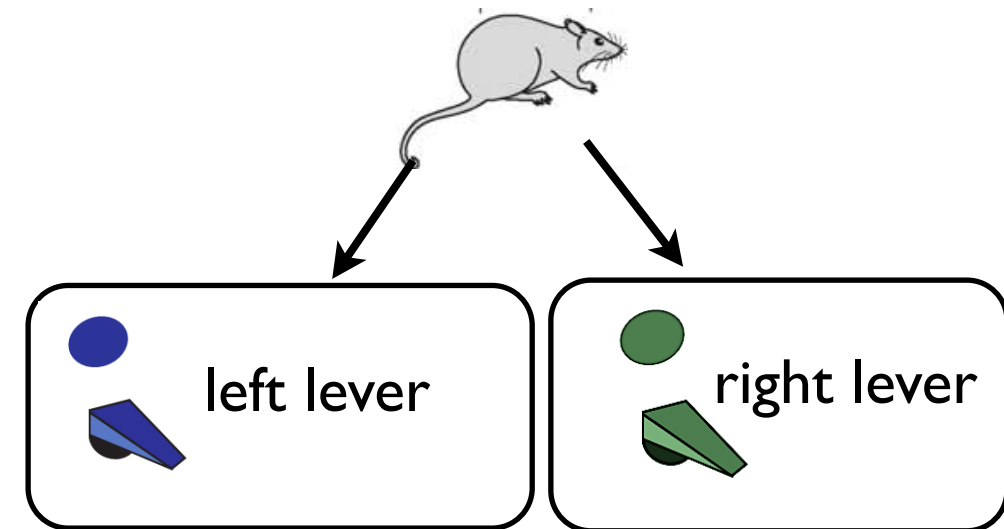
Training



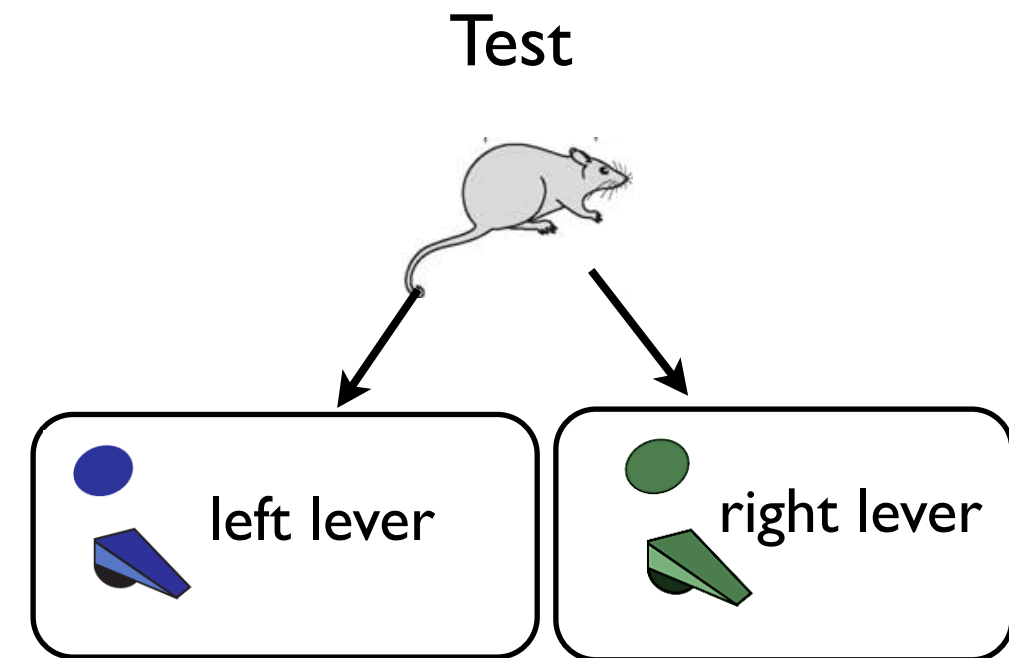
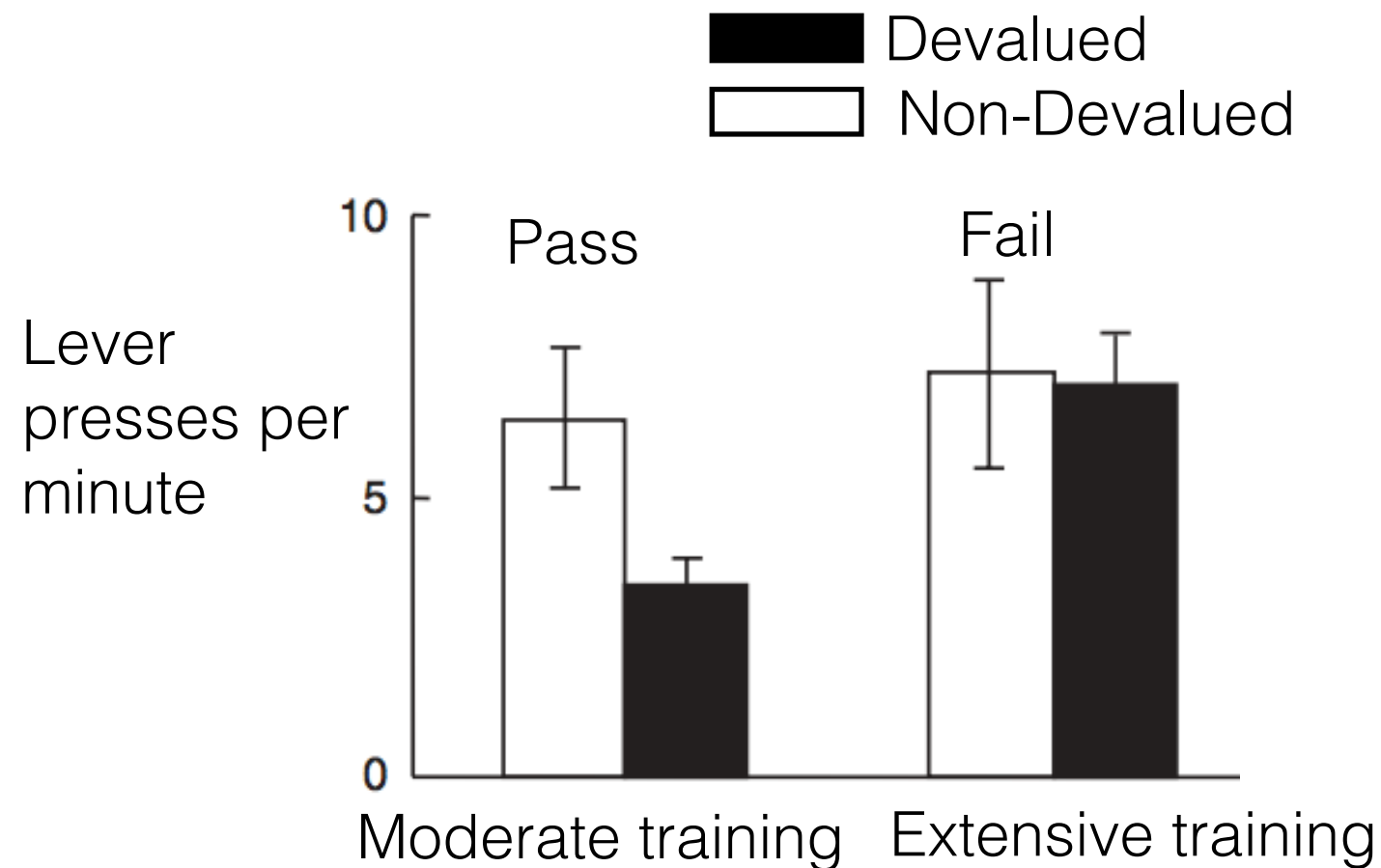
Re-training



Test



# Outcome Revaluation Paradigm



Idea: Brain has **model-free** system that fails revaluation  
+ **model-based** system that passes

*neural dissociation*  
Yin et al, 2004,2005, ESN  
*arbitration?*

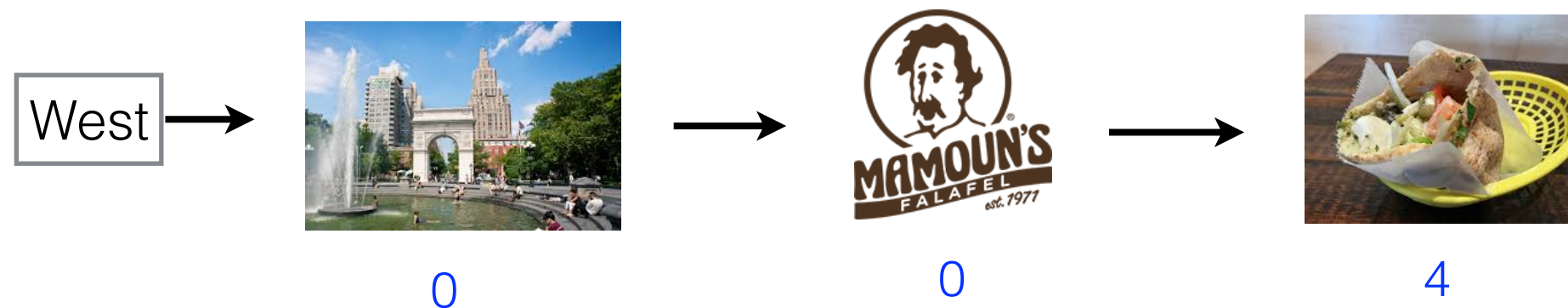
# Model-Based vs Model-Free in Psychiatry

- Habit-like behavior produced by model-free RL may be a model of compulsion in drug addiction (Everitt ,Robbins, 2005, Nat. Neuro).
- Tasks which measure model-based and model-free influences in humans find abnormal balance across many compulsive behaviors (Voon et al., 2014; Gillan et al., 2016 Elife)
- Increased model-based decisions in social anxiety symptoms? (Hunter, Meer, Gillan, Hsu, Daw, 2019, bioarxiv)



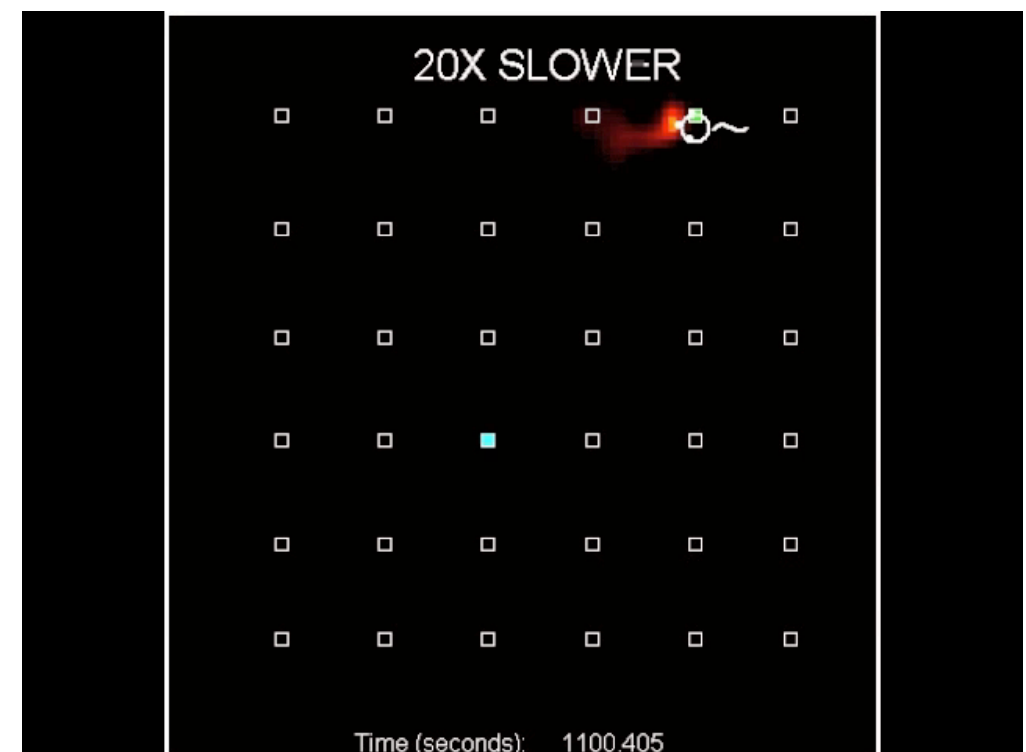
# How does model-based RL work?

## *model-based reinforcement learning (RL)*



If we have a model, how do we form values?

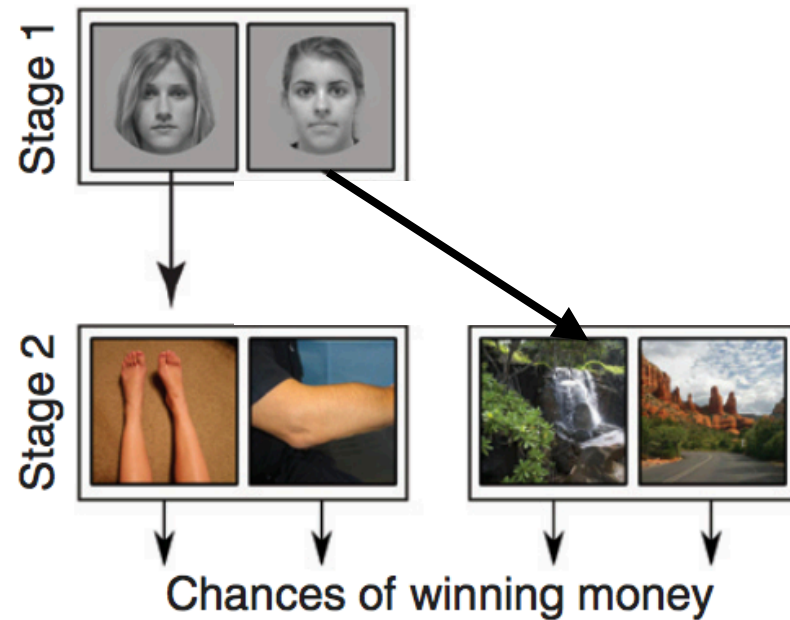
Mental simulation?



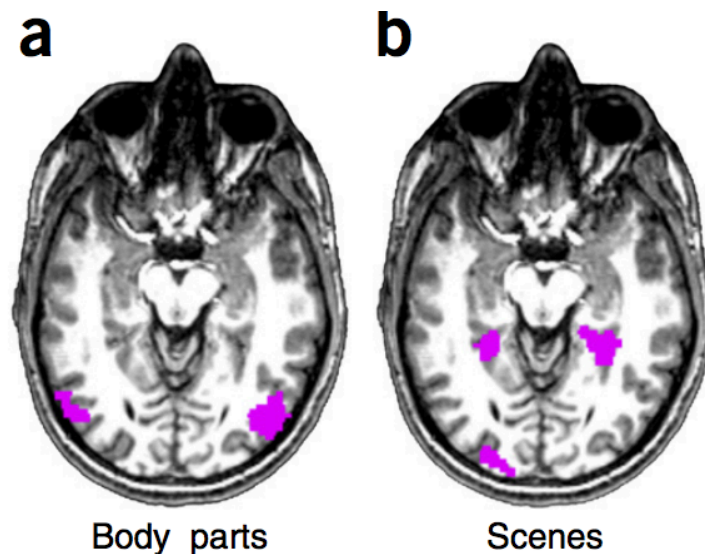


# Does mental simulation underlie model-based behavior?

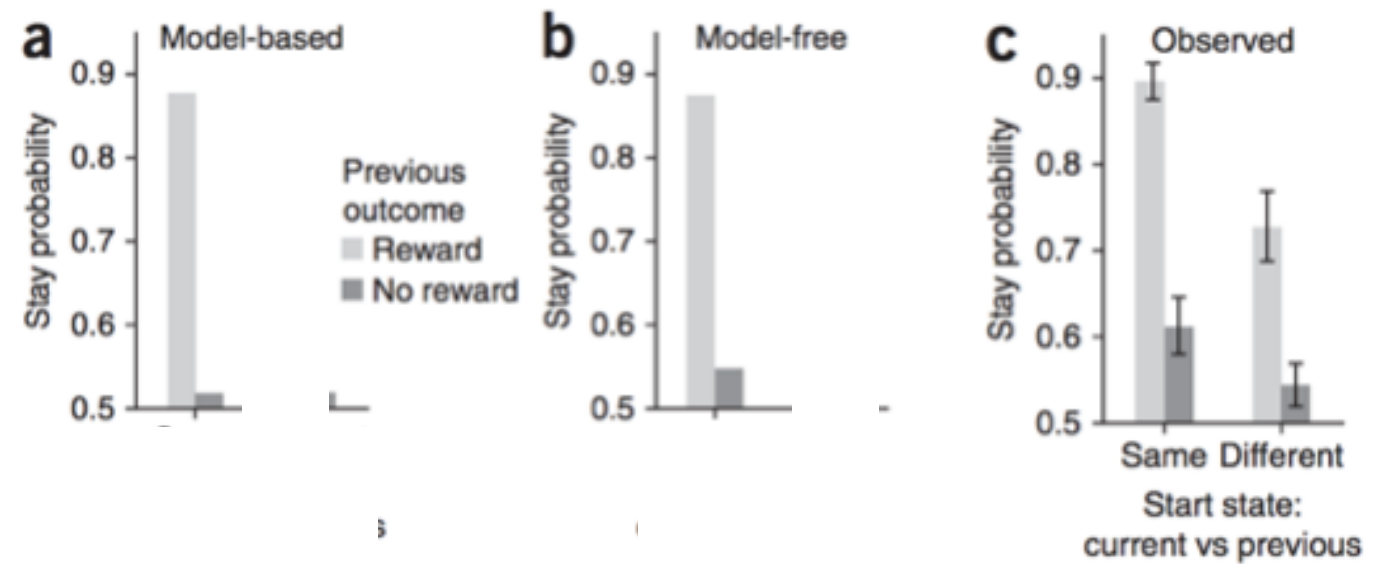
## Human Revaluation Task



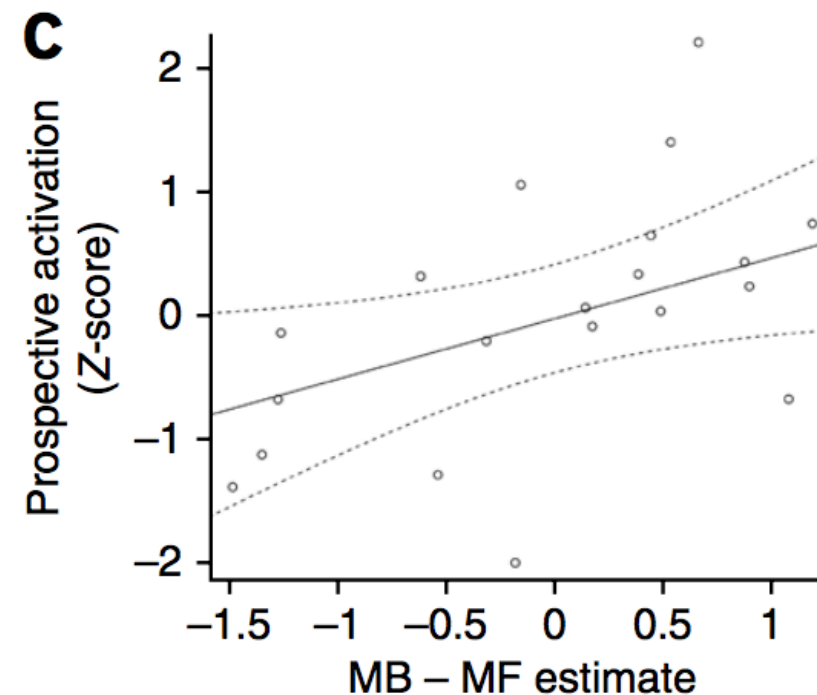
## Measuring Neural Simulation



## Behavior:





## Simulation underlies MB Behavior



**Mental health:**  
Simulation as what we think about.



How do we learn a model?

## More prediction error based updating

prediction:  $P(\text{  \rightarrow \text{  }): .4$

target: 1 if transition occurs, 0 otherwise

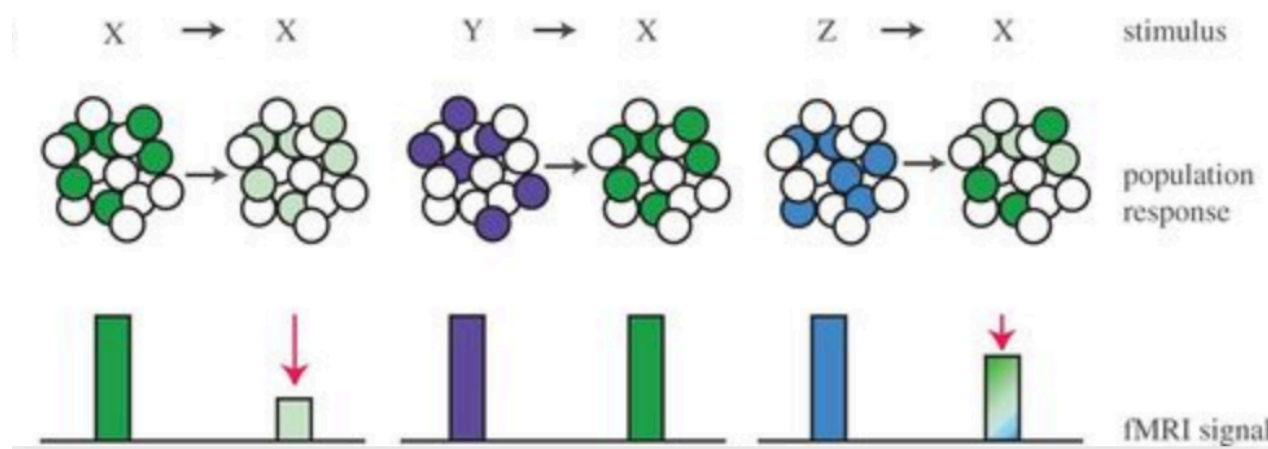
prediction error:  $1 - .4 : .6$

update:  $P(\text{  \rightarrow \text{  }) \leftarrow$   
 $.4 + \text{learning\_rate} \cdot .6$

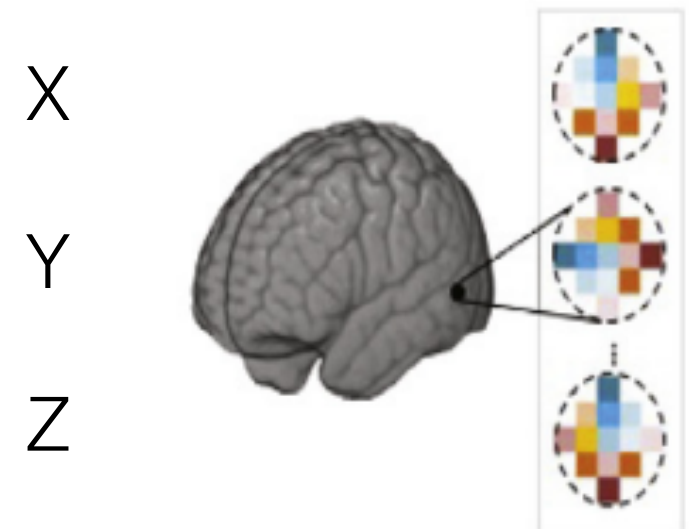
# How to measure transition models in the brain?

**Idea:** Transitions encoded as overlap in neural representations for states.  
(e.g. if  $Z \rightarrow X$  then some neurons fire for both  $Z$  and also for  $X$ ).

**Stimulus  
suppression**

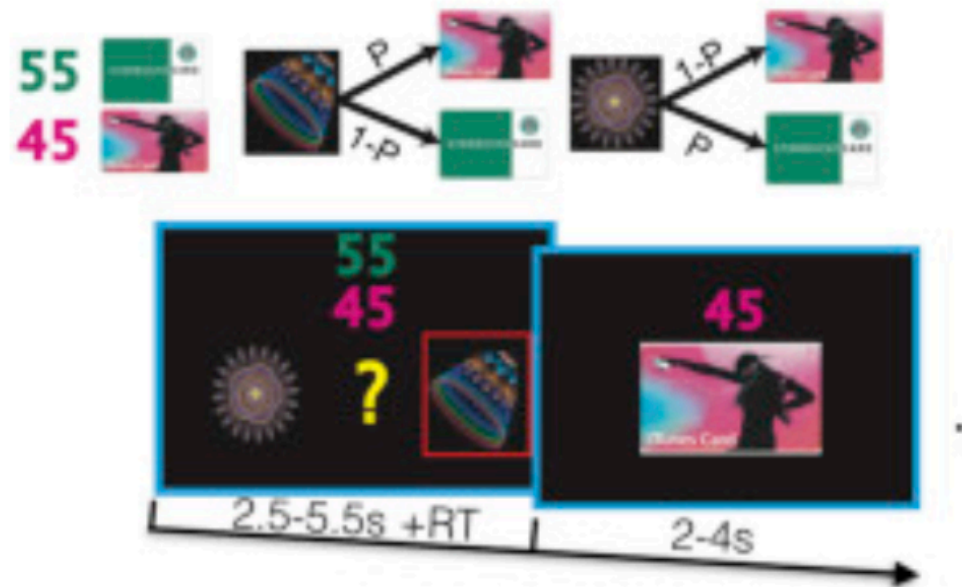


**Representational  
similarity analysis**

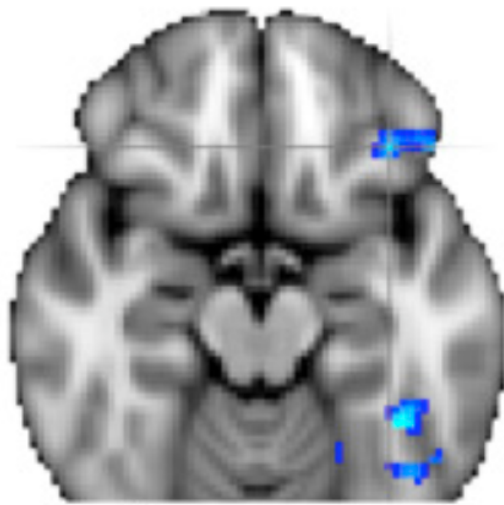


# Brain represents and updates transition expectancies

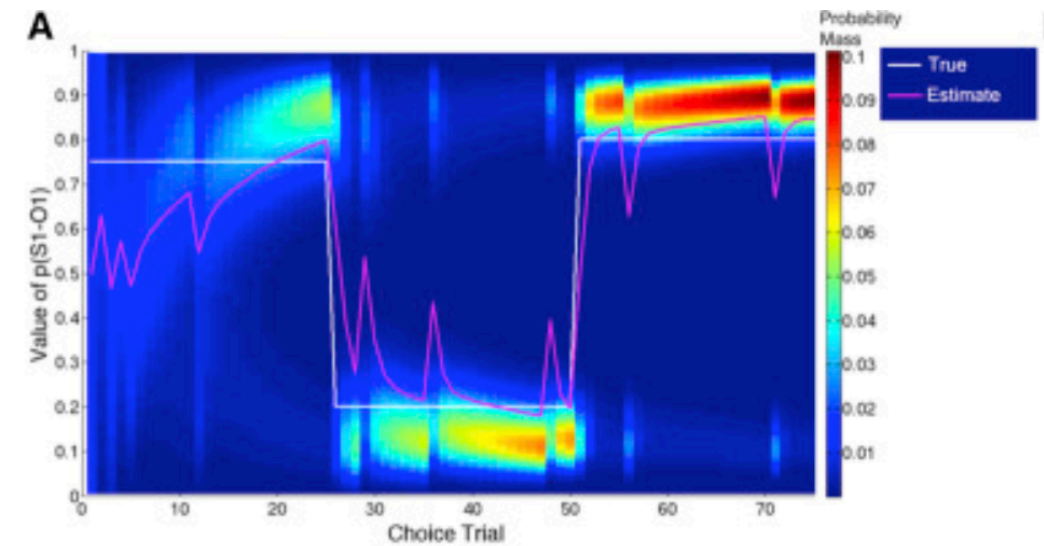
Task:



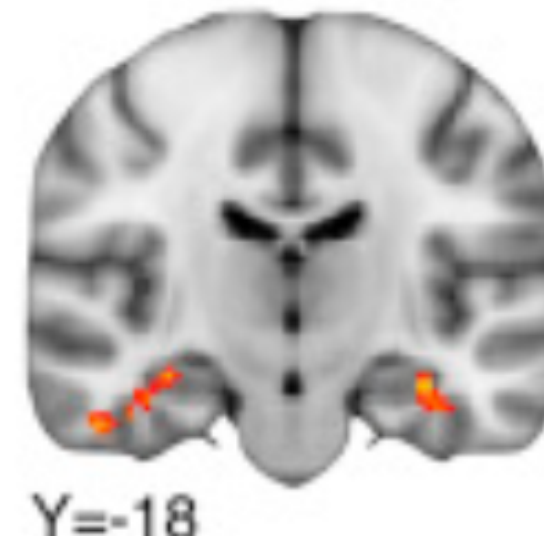
OFC: Update/Prediction Error



Modeled transition predictions :



Hippocampus: Transition Model

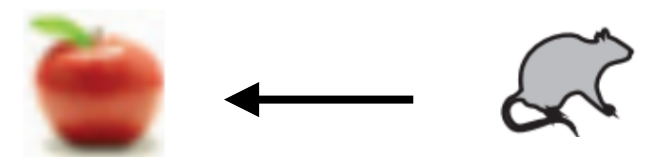
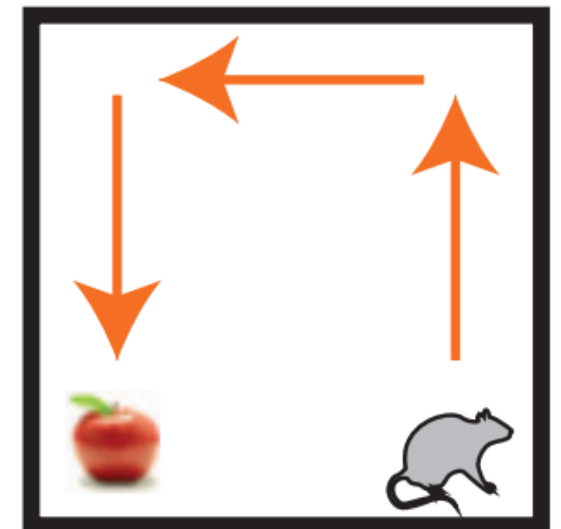


# Representing and updating models in psychiatry

- Hippocampal lesion patients deficient in model-based decisions (*Vikbladh, Maeger,..., Shohamy, Burgess, Daw 2019 Neuron*)
- Reduction in hippocampus in ageing and Alzheimers
- Anxiety: Difficulty in updating causal models in aversive situations (*Browning, Behrens, Jocham, O'Reilly, Bishop 2015 Nat. Neuro*)

# Structural generalization

- We don't learn new transition functions from scratch.
- Structural knowledge places strong constraints on what transitions are possible.
- Helps us learn new transition functions for new environments with less experience.



# Structural Generalization in Depression

- Control and depression (*Abramson, Seligman, & Teasdale, 1978; Willner 1985; Williams, 1992; Alloy, 1999; Maier & Watkins, 2005*)
- Animal model: learned helplessness (*Mair and Seligman 1976*)
  - Trained in environment with no control to avoid shocks
  - enter environment where they can avoid shocks, animals choose to not take avoidance actions
- Computational model (*Huys and Dayan, 2009, Cognition*)
  - Structure Learning: desirable outcomes are not reachable.
  - Generalization constrains learning of transitions in new environments

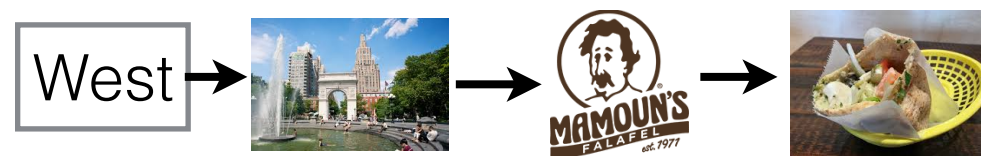
# Plan

- Part 1: Multiple systems for choice
  - Reward Prediction errors, Dopamine
  - Predicting rewards through time, Model-free RL, Habits
  - Flexible choice, Model-based RL
- Part 2: Approximate Planning
  - Tree-search, Pruning
  - Temporal Abstraction, Successor representations
  - DYNA, Prioritized Simulation

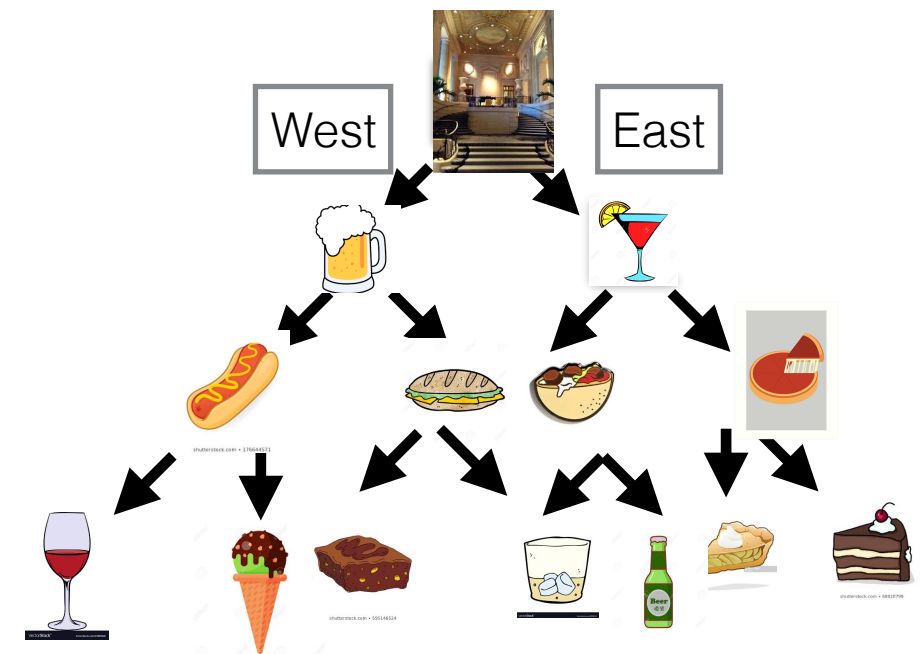


# Trees are more challenging

How do i know which rewards will follow west?



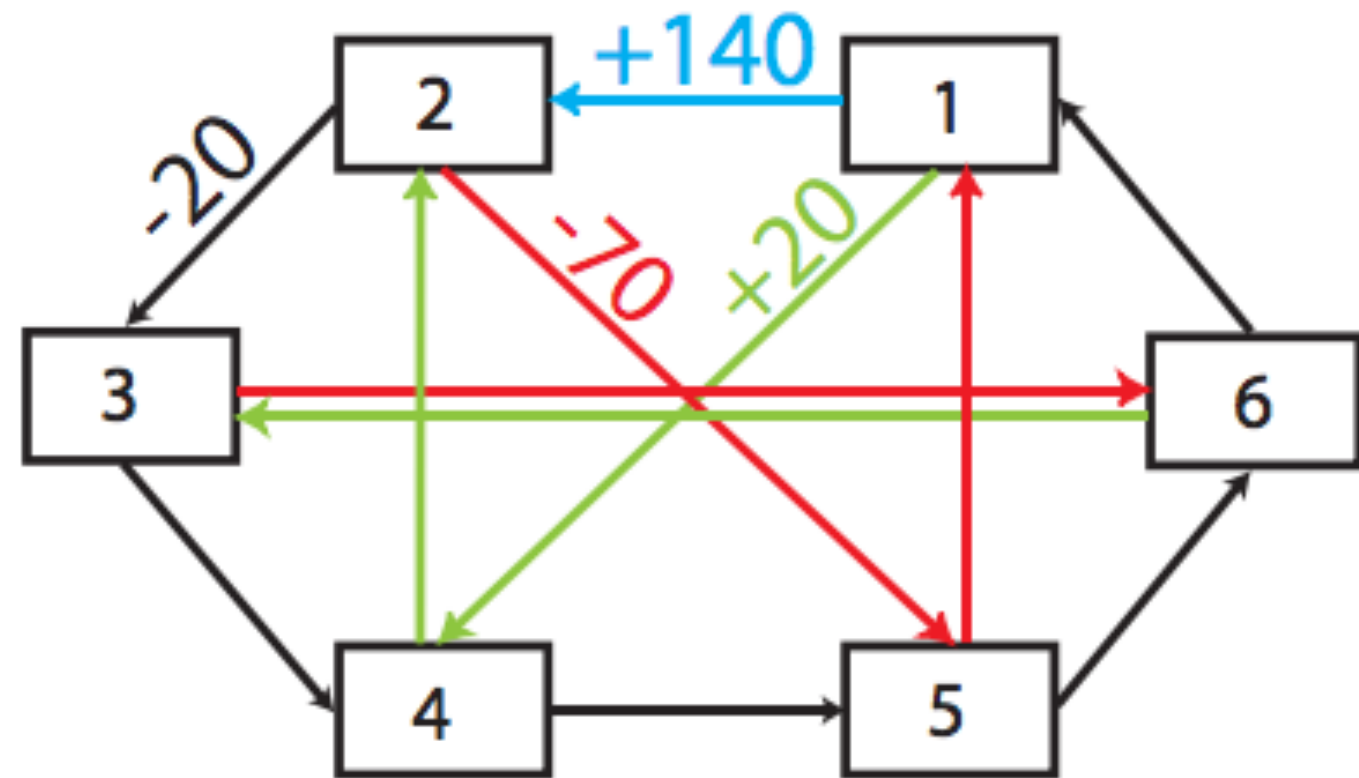
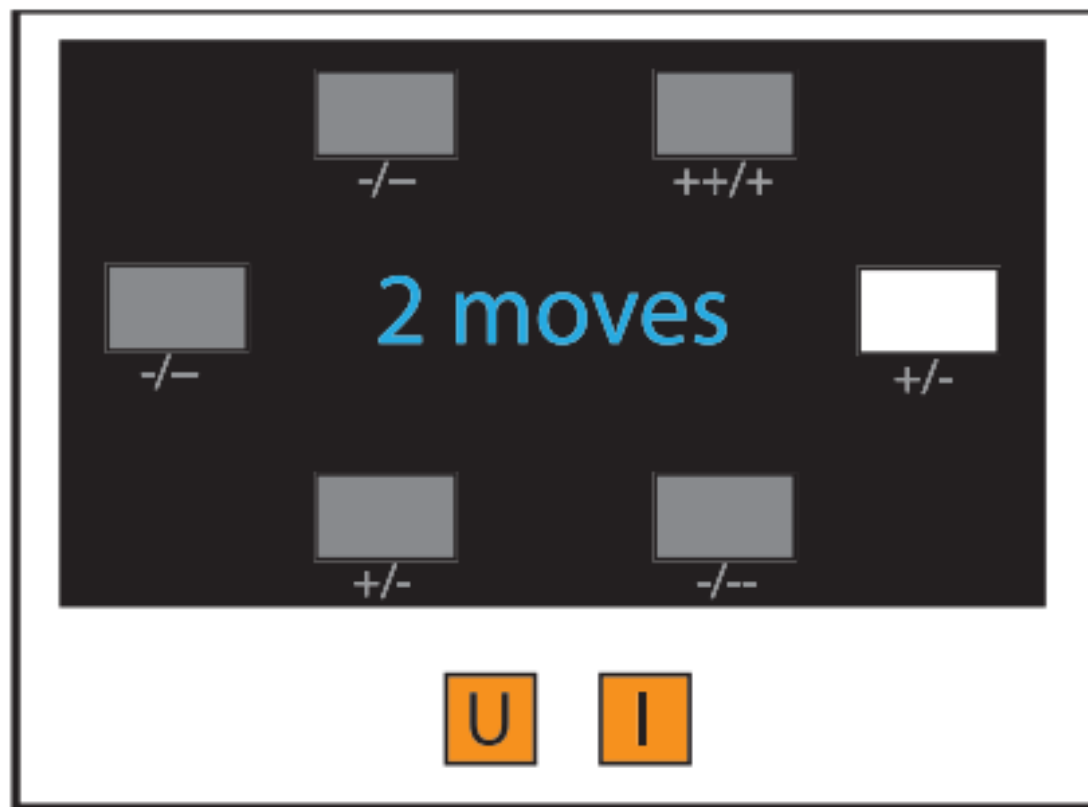
vs.



**Tree search:** simulate every possible sequence of states that might follow west to find most rewarding sequence.

**Pruning:** stop search along certain paths if they're unlikely to be the best

# Pruning Task

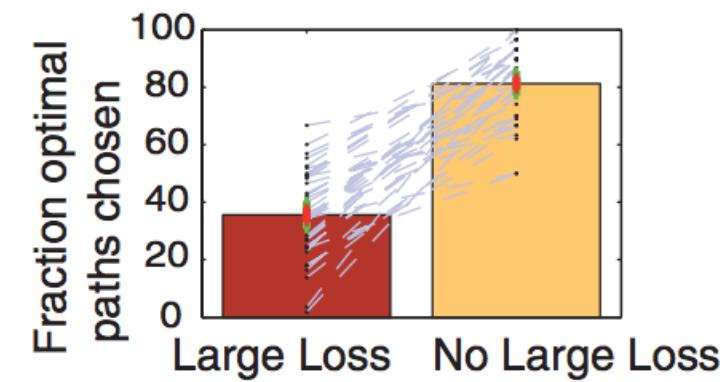
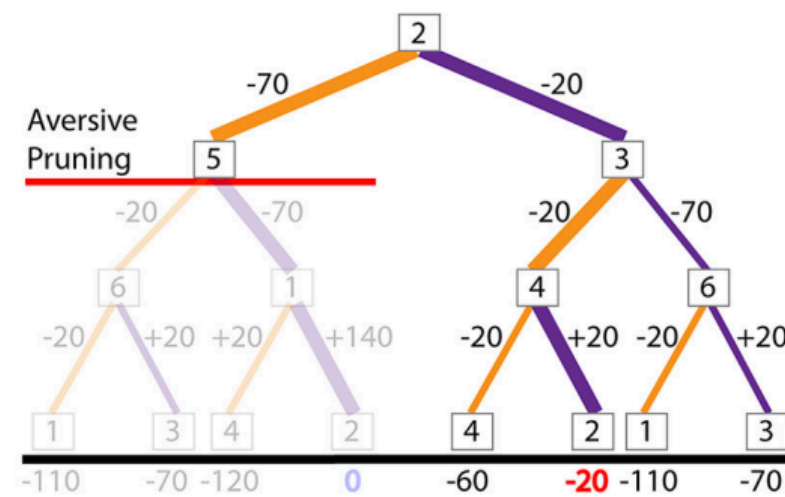
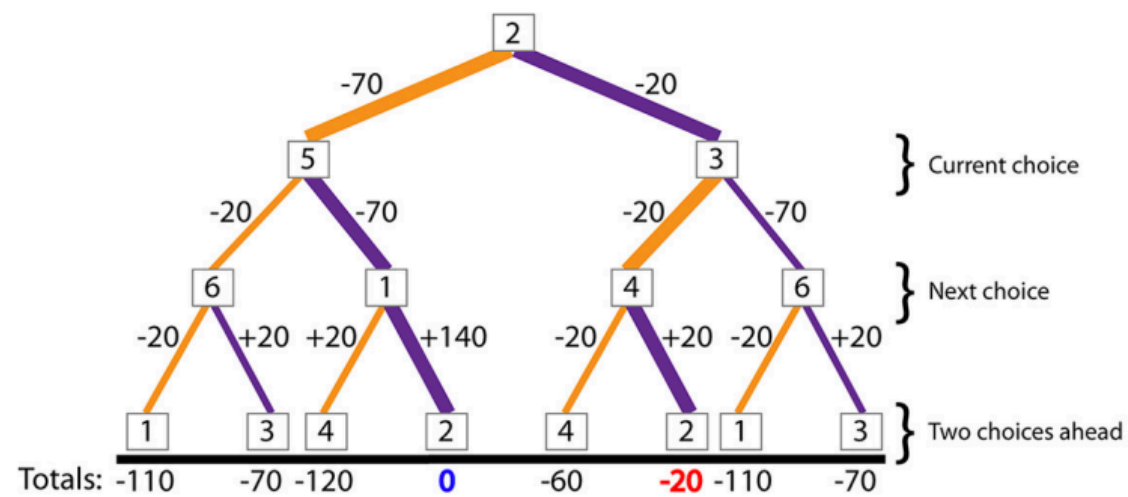


# Aversive Pruning in Humans

Best choice is left

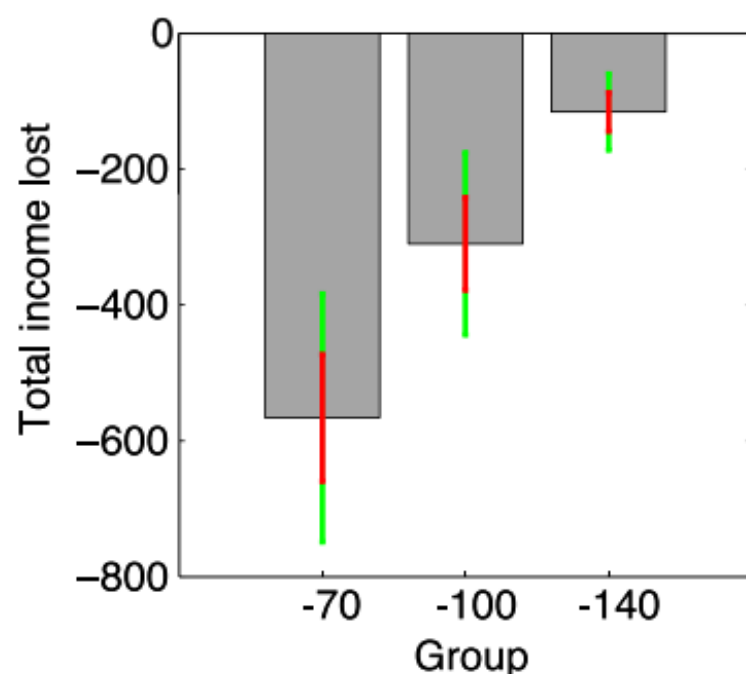
Not searching past large loss -> mistakes

Mistakes on these trials

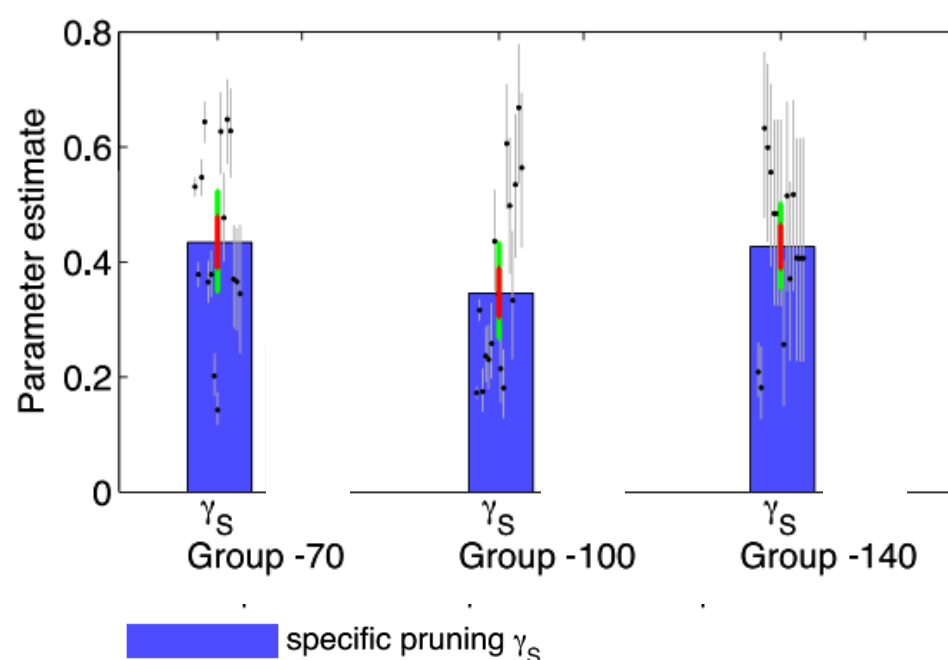


# Aversive Pruning as 'Emotional' Mental Action

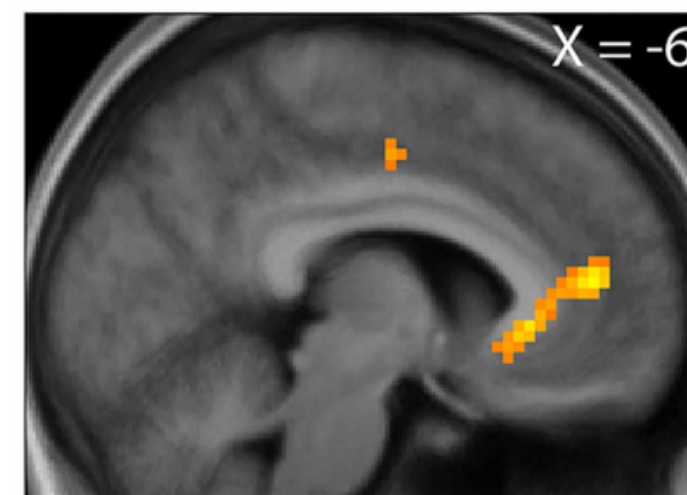
value of pruning  
changes by condition



behavior insensitive  
to this



sgACC higher on  
pruning trials



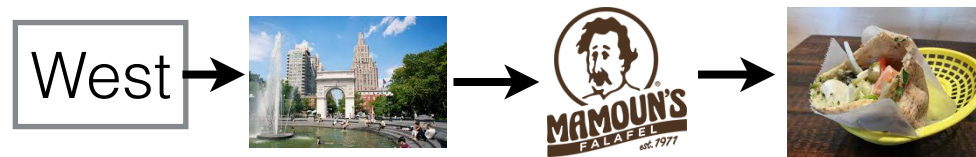
sgACC: represents negative motivational value (Amemori and Graybiel, 2012 Nat. Neuro.), overactive in mood disorders (Drevets WC, Price JL, Simpson JR Jr., Todd RD, Reich T, Vannier M, Raichle ME, 1997, Nature)

# Pruning and psychiatry

- Pruning is often the ‘resource-rational’ optimal computational strategy *(Callaway, Lieder, Dan, Gul, Krueger, Griffiths , 2018, Proc. Cog Sci Soc.)*
- Emotion-based cognitive actions? *(Huys Renz, 2017, Biological Psychiatry)*
- What if variation in pruning?
  - Depression/Anxiety? *(Huys, Eschel, O’Nions, Sheridan, Dayan, Rosier, (2012) PLOS CB; Lally\*/Huys\*, Eschel, Faulkner, Dayan, Rosier (2017) J Neuro)*
  - OCD? / Compulsion?

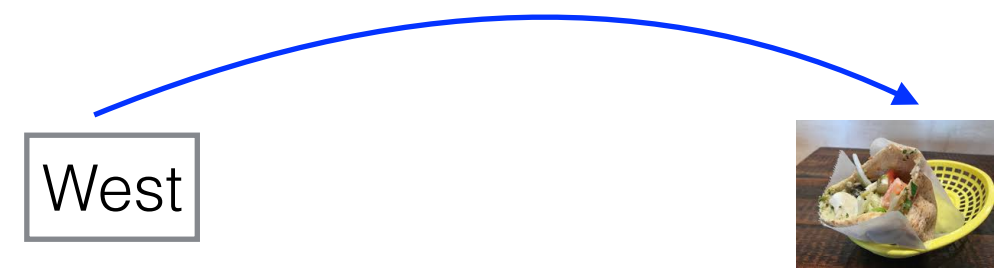
# Temporal Abstraction

Standard model-based: simulate 1-step of transition at a time.



Alternative: store multi-step transition predictions

“3-step” transition



3-step  
successor state

Multi-step models: allow multiple steps of state prediction at once.

Successor representation: use multi-step successor states to represent states

# Successor Representation

**Represent states/actions as sum of  $n$ -step successors** *Dayan 1993 Neural Computation;*

Environment:



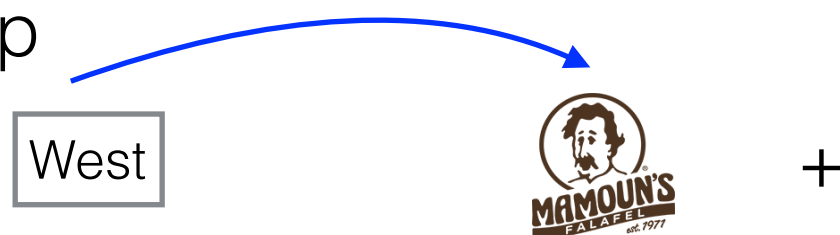
Representation for West :



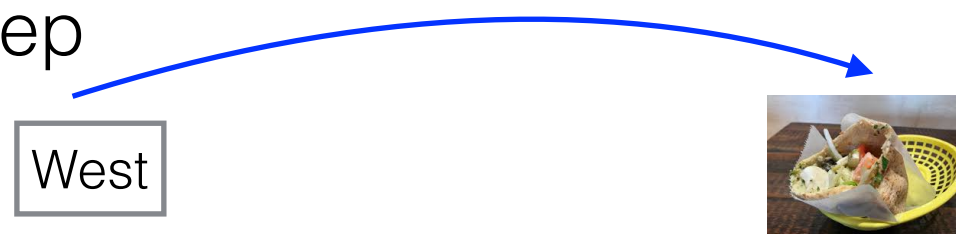
1-step



2-step

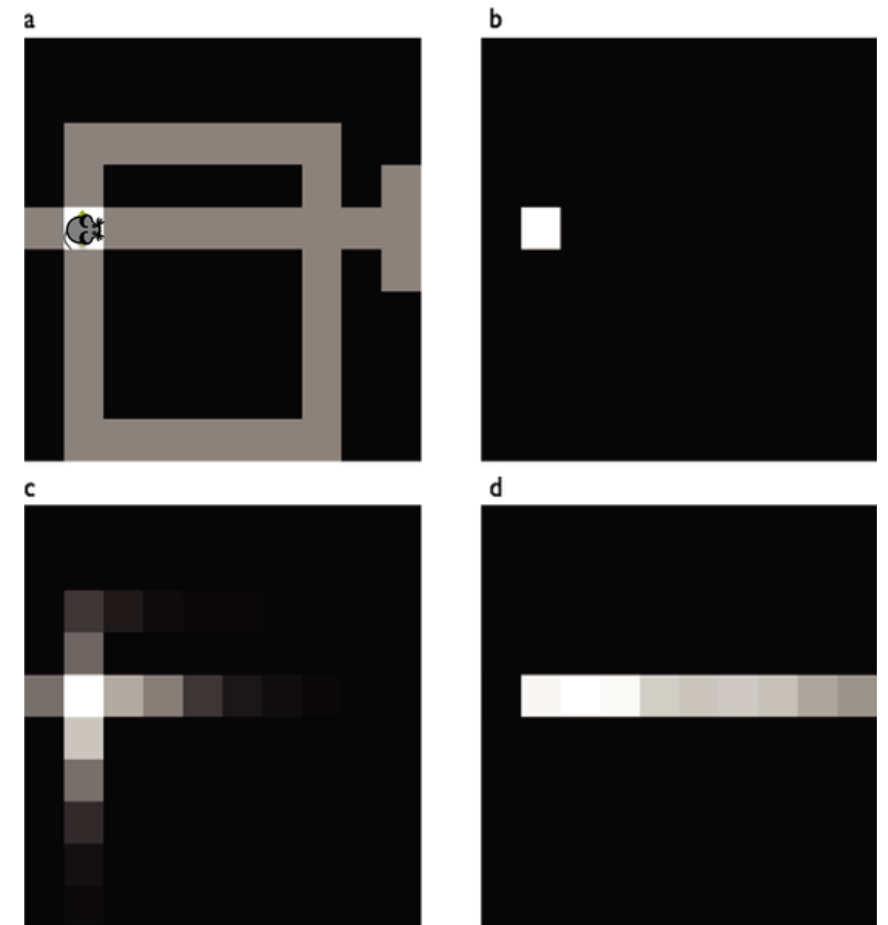


3-step



Spatial Example:

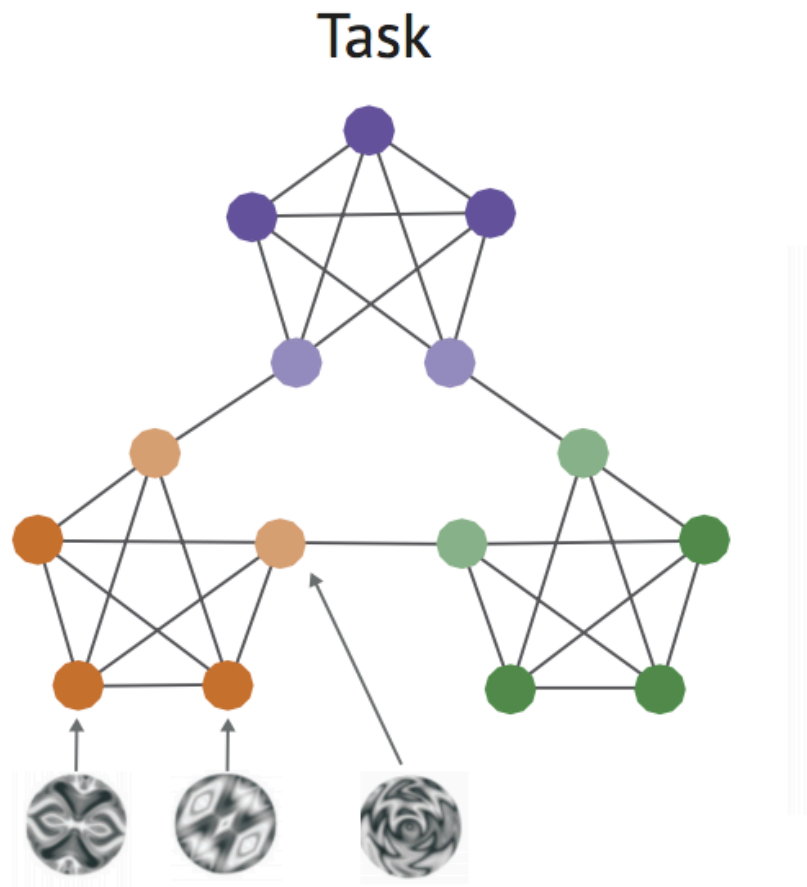
Punctate Rep.



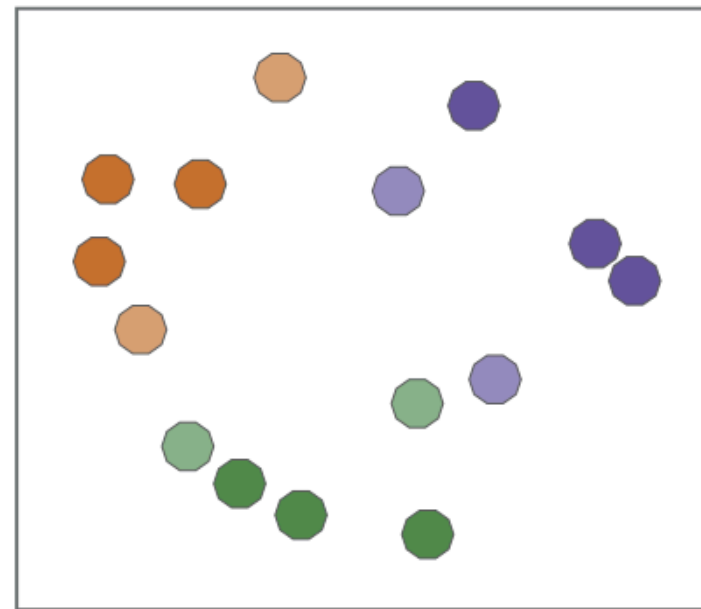
Successor Rep.

*Russek\*/Momennejad\*, Botivinick,  
Gershman Daw 2017 PLOS CB*

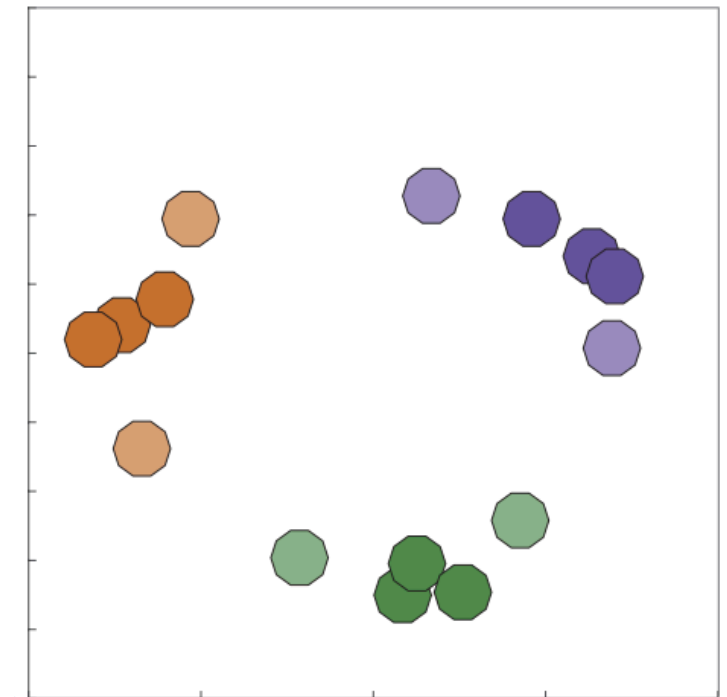
# Successor Representation in Hippocampus



Bilateral hippocampus  
MDS



Successor representation  
MDS



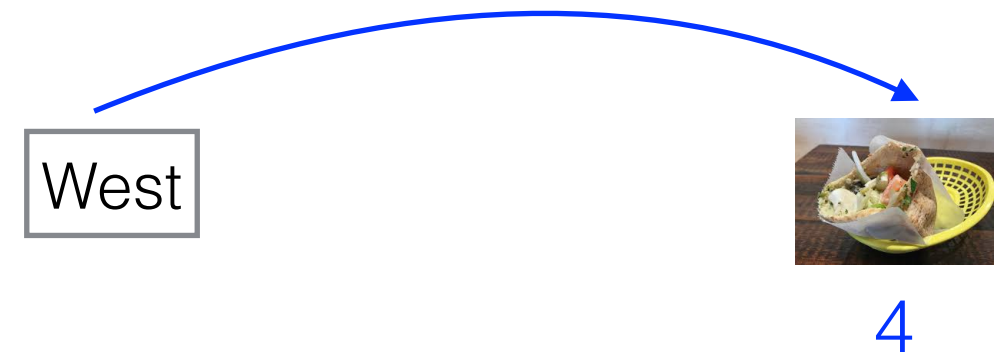


# Inflexibilities of successor representation in choice

Full model-based:



Successor representation:

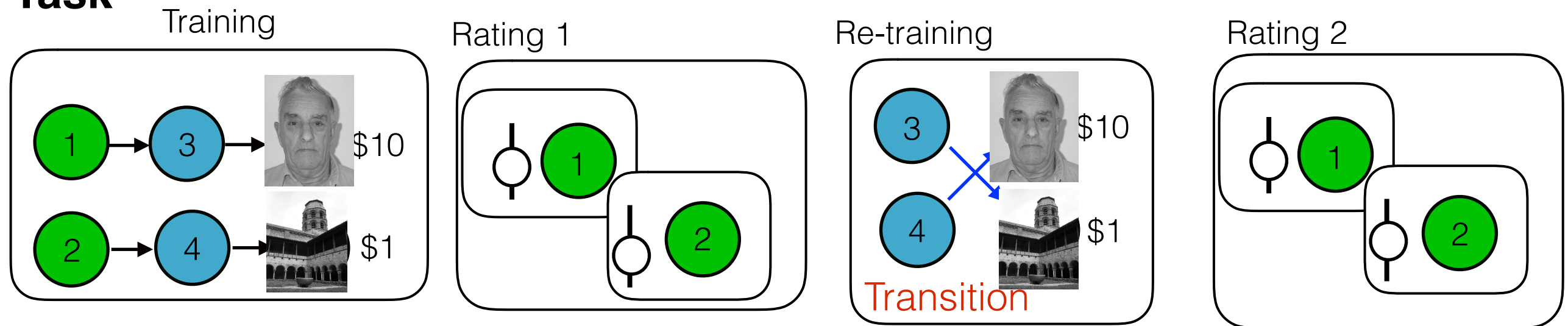


Following changes to (one-step) transitions, SR needs new experience to re-learn correct multi-step transition predictions.

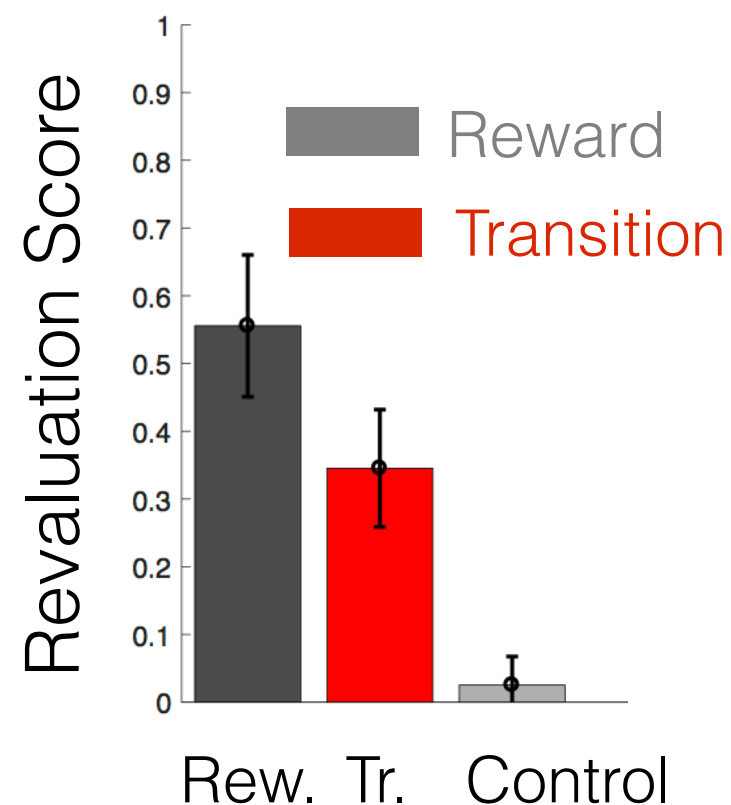


# Do brains/people make same errors as the successor representation?

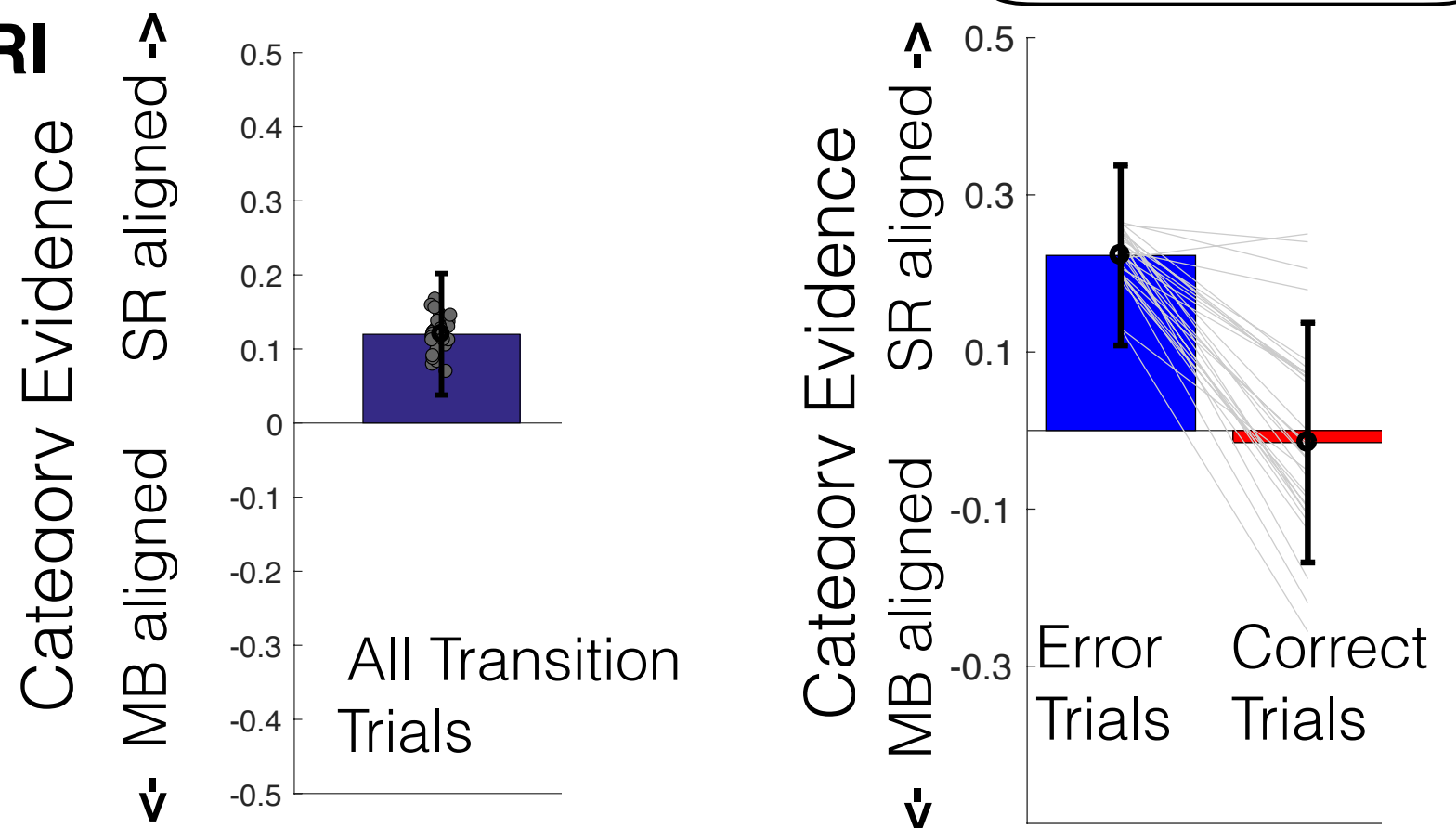
## Task



## Behavior



## FMRI



# Successor Representation and Psychiatry

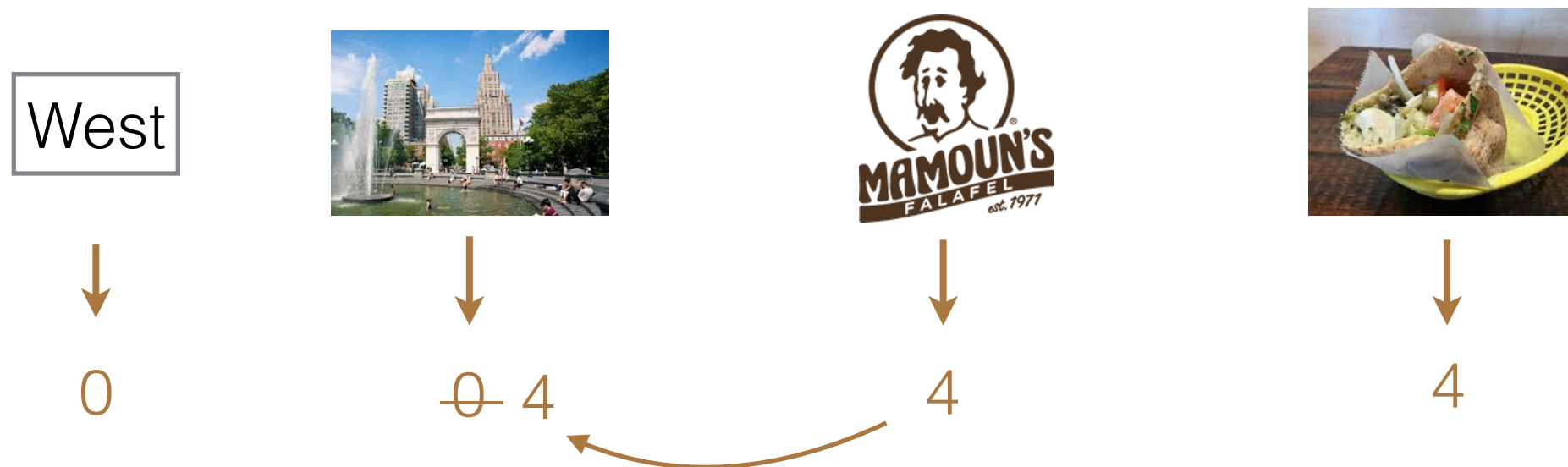
- SR causes behavioral habits similar to model-free learning: compulsion
- Difficult to change predictions of long-run future states
- Aberrant generalization of rewards to value

# Dyna: Mixing Model-Based and Model-Free RL

store **model** (like **MB**): use to **simulate** transitions

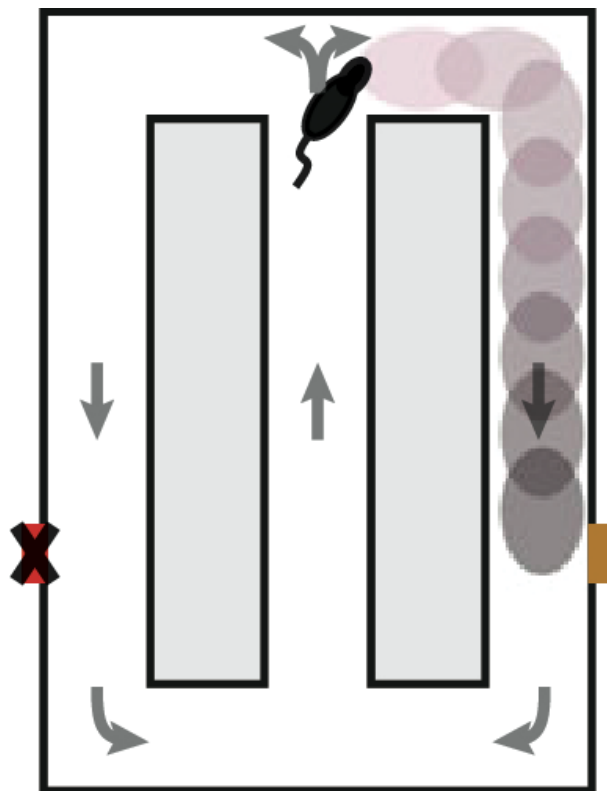


also store **value estimates** (like **MF**): **What to simulate and when?**

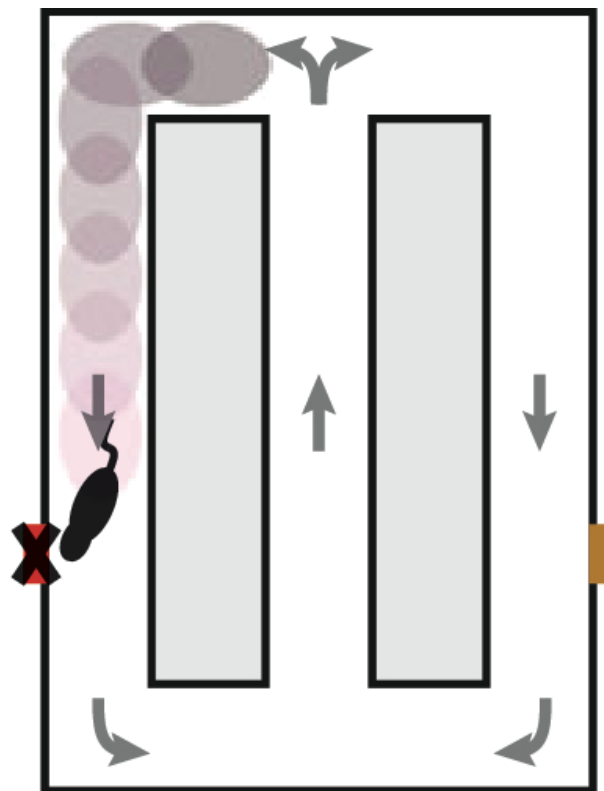


update from both **experienced** and **simulated** transitions (TD rule)

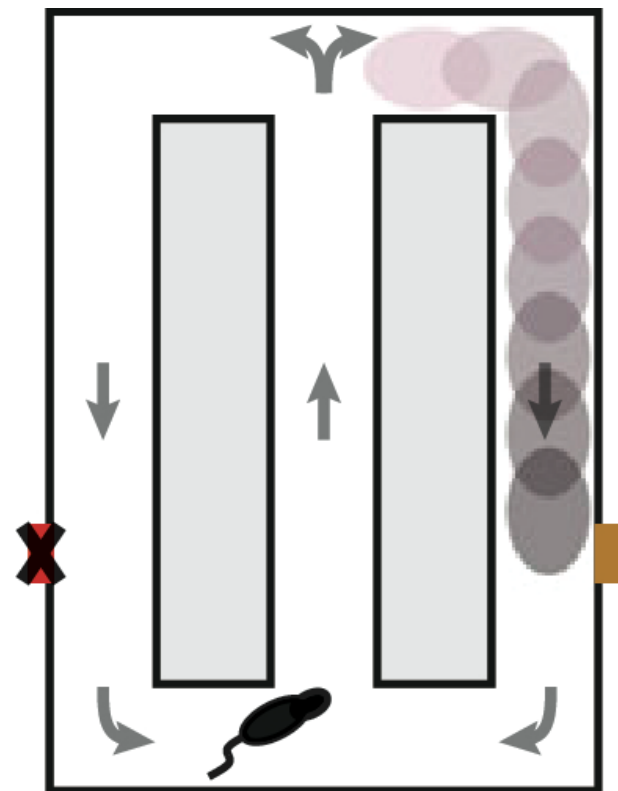
What to simulate? Replay as a window.



Forward sequence



Reverse sequence



Remote sequence

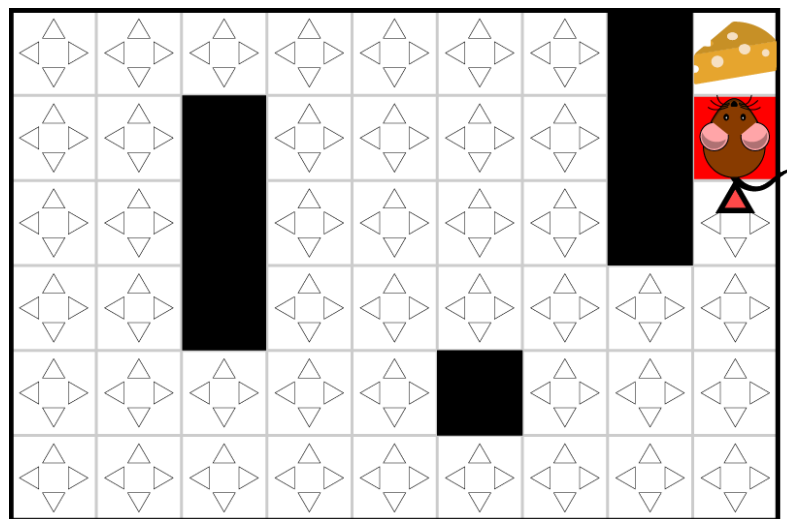
# What state/action should one simulate?

Define value of simulation

$$\text{VOS}(s,a) = \mathbf{E}[\text{return}|s,a] - \mathbf{E}[\text{return}|\sim s,a] = \text{Gain}(s,a) \times \text{Need}(s)$$

## Gain

Additional reward expected due to what is learned from value update



► Gain

Value

Drives backward replay

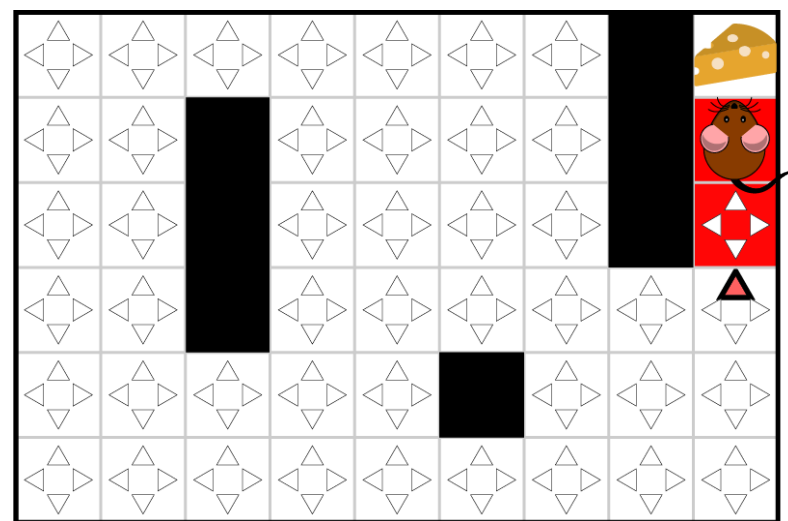
# What state/action should to simulate?

Define value of simulation

$$\text{VOS}(s,a) = \mathbf{E}[\text{return}|s,a] - \mathbf{E}[\text{return}|\sim s,a] = \text{Gain}(s,a) \times \text{Need}(s)$$

## Gain

Additional reward expected due to what is learned from value update



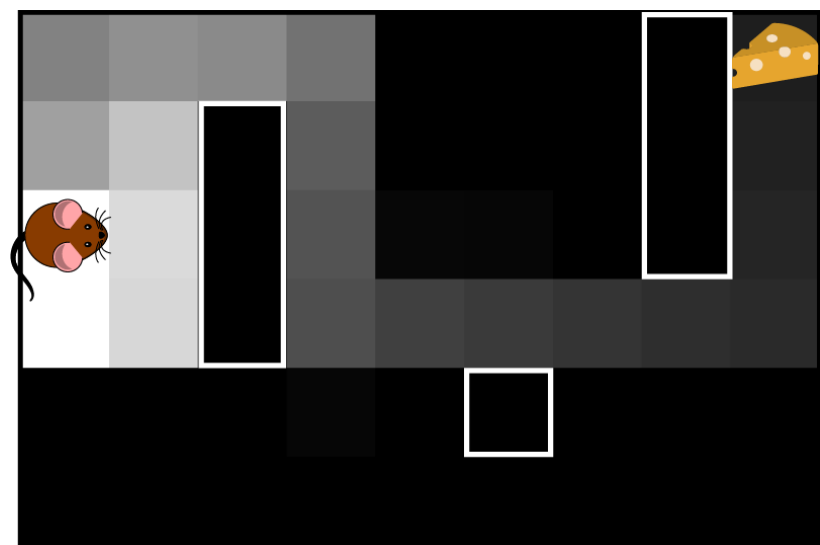
► Gain

Value

Drives backward replay

## Need

Number of times agent expects to visit the target state (SR).



Need

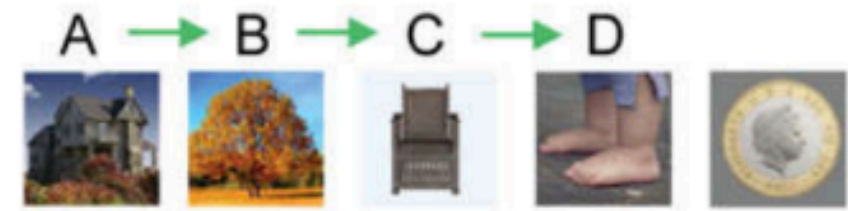
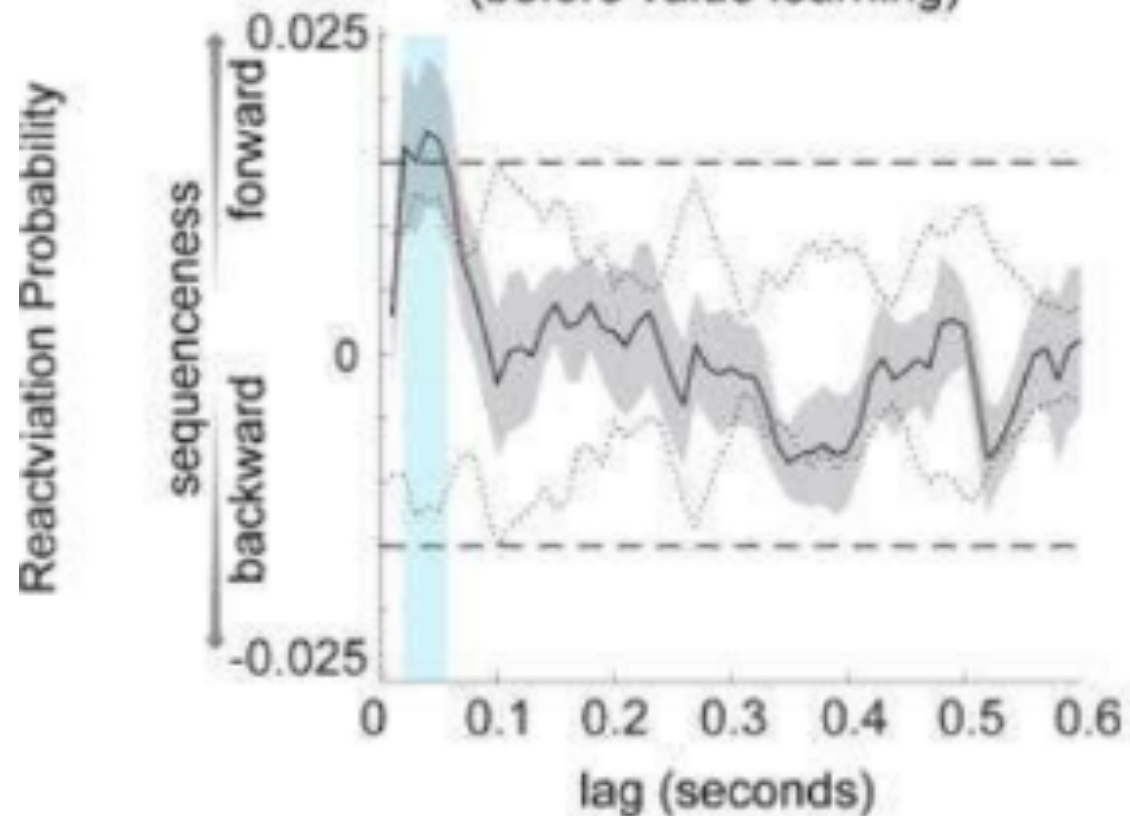
Drives forward replay



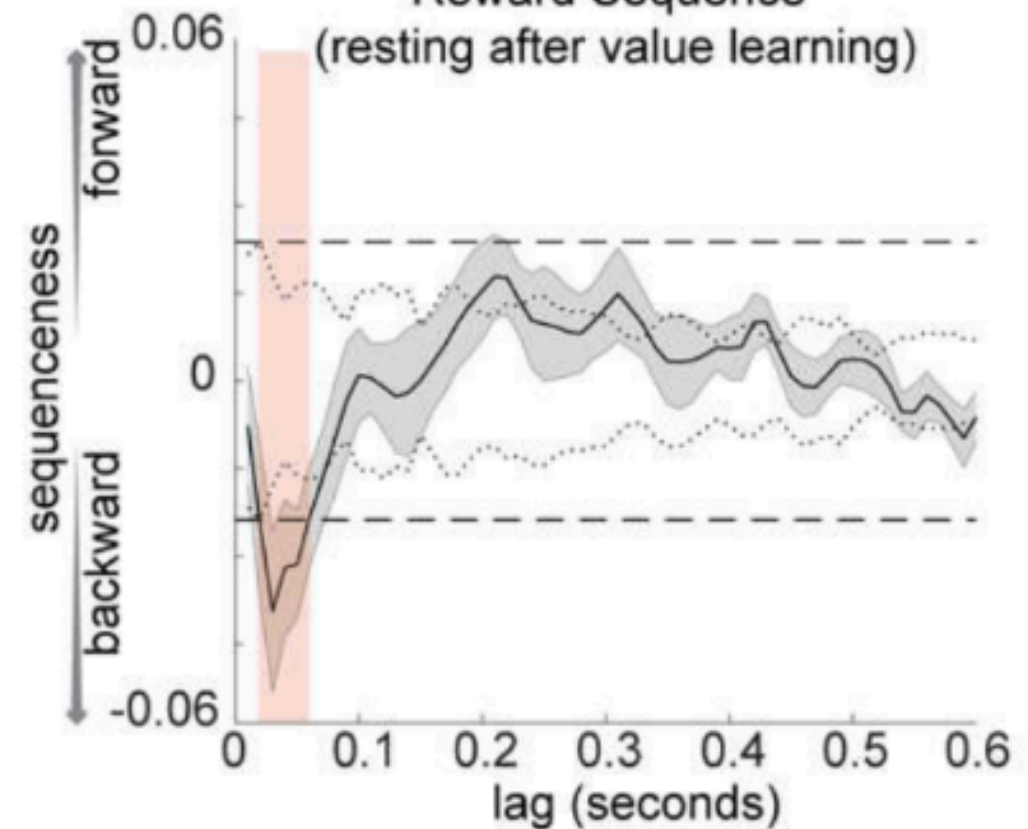
# Forward and Reverse Replay in Humans



True Sequence  
(before value learning)



Reward Sequence  
(resting after value learning)





# Prioritized Simulation and Psychiatry

- A model of what to think about and when
- Anxiety - biased prediction?
- PTSD - high gain?
- Depression, automatic thoughts

# RL in Psychiatry

- Reinforcement learning theory relates experiences to choices
- How neural processes contribute to choice.
- How choice might be flexible or inflexible
- How experience drives what we think about.

# Thanks

- Contributing helpful feedback and figures:
  - Quentin Huys, Jolanda Malamud, Fatima Chowdhury, Daniel Mcnamee, Rachel Bedder, Yunzhe Liu, Marcelo Mattar, Oliver Vikbladh