

# model-based planning II: mixing memory and desire

ccnss

2018.07.07

slides and references available at

<http://aaron.bornstein.org/ccnss/>

# outline

- I. computational role(s) of memories
- II. experimental evidence
- III. if time: open questions

# outline

- I. computational role(s) of memories
- II. experimental evidence
- III. if time: open questions

# uncertainty-based arbitration

- **model-free**: policies well-learned, cached values highly certain
- **model-based**: state spaces learned, policies and values require evaluation
- ... but what about when the state space isn't yet well-learned?

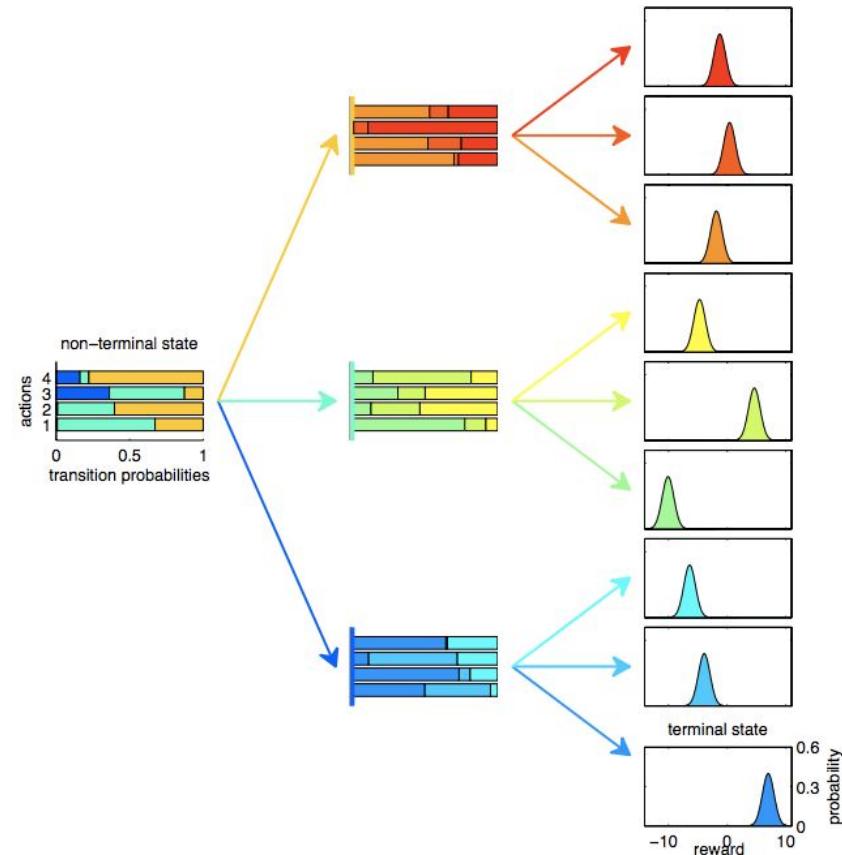
# the third way

“tree” MDP (tMDP)

- branching factor  $B$
- depth  $D$

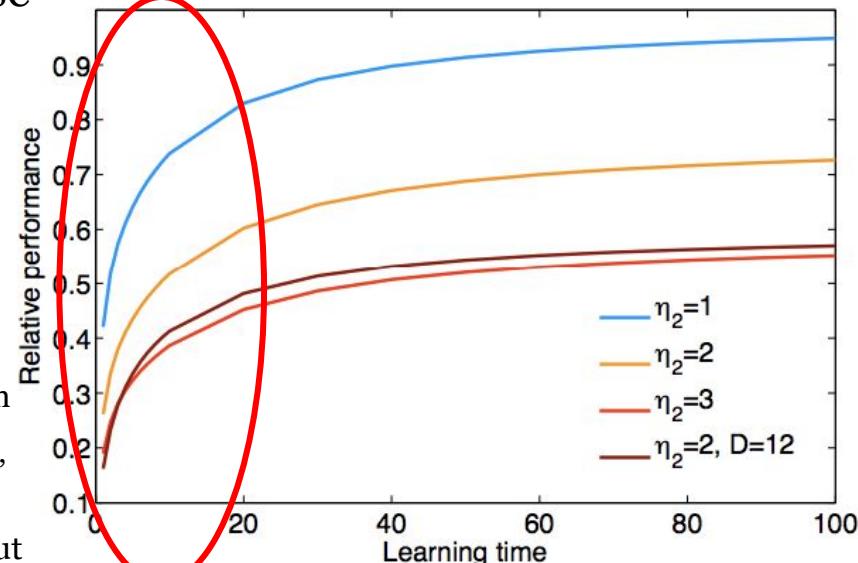
“episode”

- a single trajectory from  $S_o$  to terminal



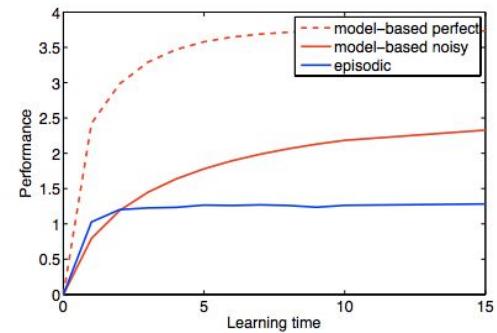
# model-based performance at noise and complexity

- model-based learning slows with greater noise and/or complexity
  - (*and* is asymptotically lower, hence model-free can come to dominate)
- proposal: “episodic control” - sample one previous trajectory, do that.
  - “each time the subject experiences a reward that is ... larger than expected ... it stores the specific sequence of state-action pairs leading up to this reward, and tries to follow such a sequence whenever it **stumbles upon** a state included in it”
  - “advantages will be ultimately counteracted by the haphazardness of using single samples that are ‘adequate’, but by that time the other controllers can take over.”

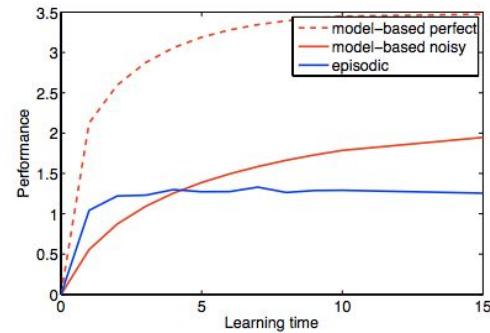


( $\eta$  = noise,  $D$  = depth; relative to theoretical perfect performance)

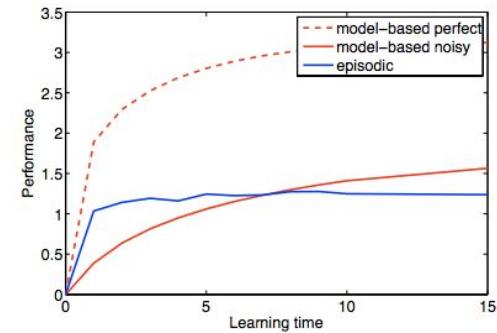
# “episodic control” useful in early learning



(B=2)



(B=3)



(B=4)

although “asymptotic performance of [episodic] is rather poor...”

# but the world is not a tree mdp

- generally speaking, we may not be very certain about our current state
- so to use our value function  $Q(\mathbf{S}, A)$  it helps to perform *state inference*
- episodic memories can be useful in inferring state when experience is sparse
  - “pattern completion” from “partial input”
  - not just repeating past actions

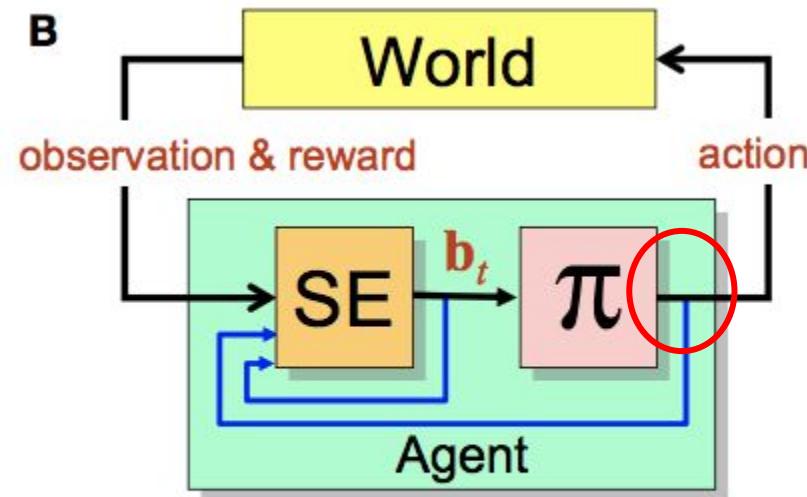
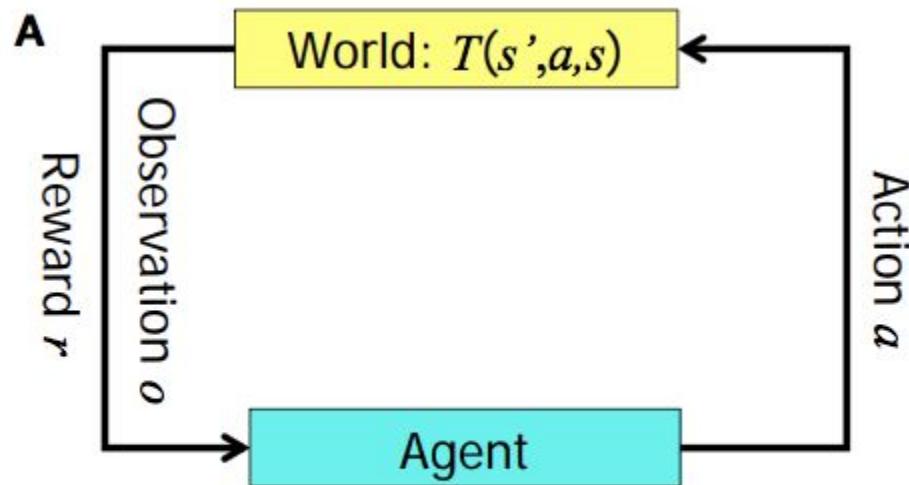
# partially observable mdp

- new state-space formalism: *partially observable* MDP (POMDP)
- rather than fixed  $S$ , distribution of *beliefs*  $\mathbf{b} = b(S)$ , and now  $Q$  can be an expectation

$$Q(\mathbf{b}, A) = \mathbb{E}_s Q(S, A)b(S)$$

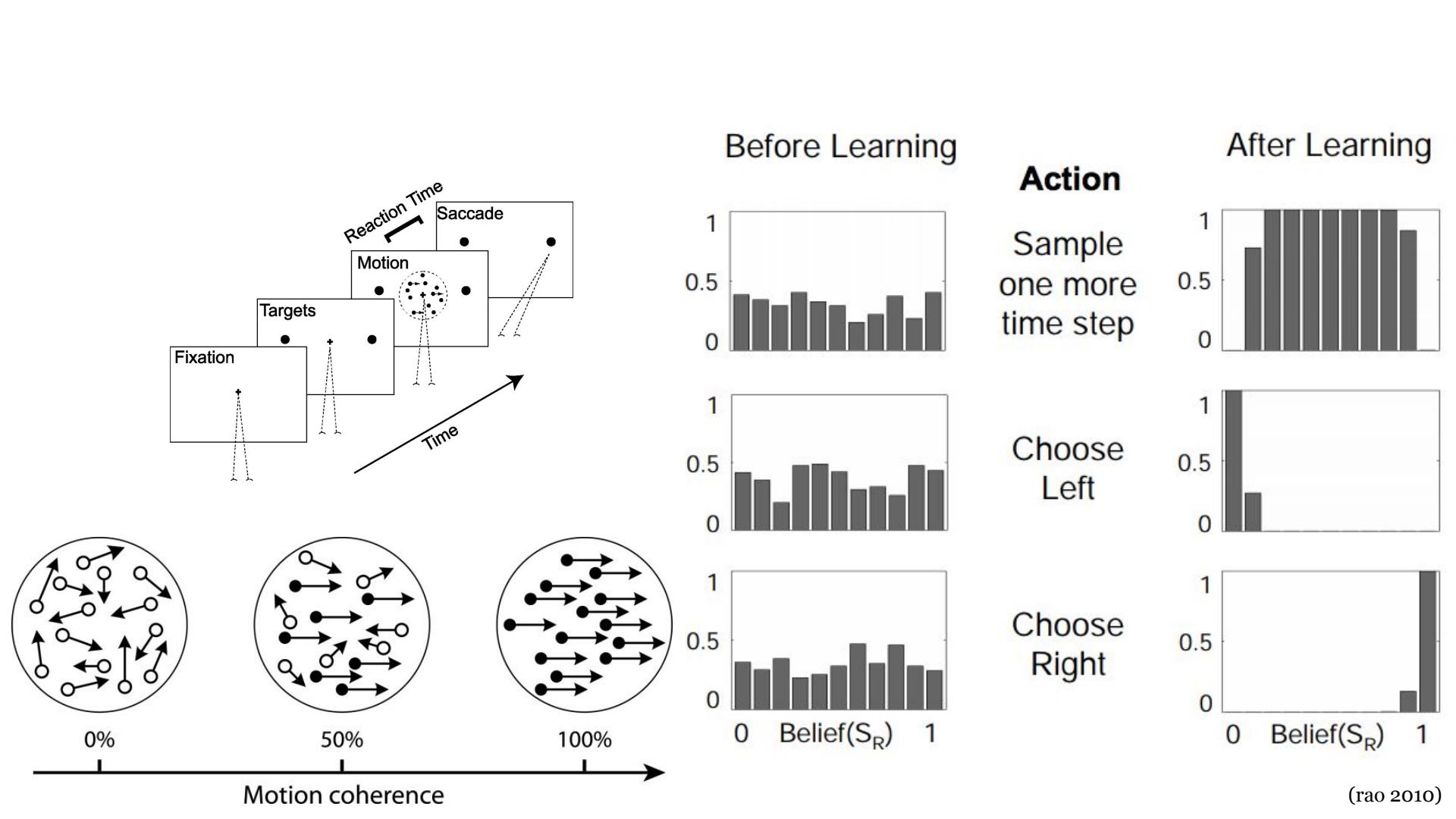
- this carries some degree of uncertainty
- the agent has the option to reduce this uncertainty by *sampling*
  - samples can be drawn from sensory information — or from memory
  - value function is convex in belief — there is always value to reducing uncertainty, modulo the cost of sampling (e.g. spikes, delaying reward, etc)
  - therefore, sample until the uncertainty is low enough that the cost of sampling (or not-acting) is overcome

# sampling as action

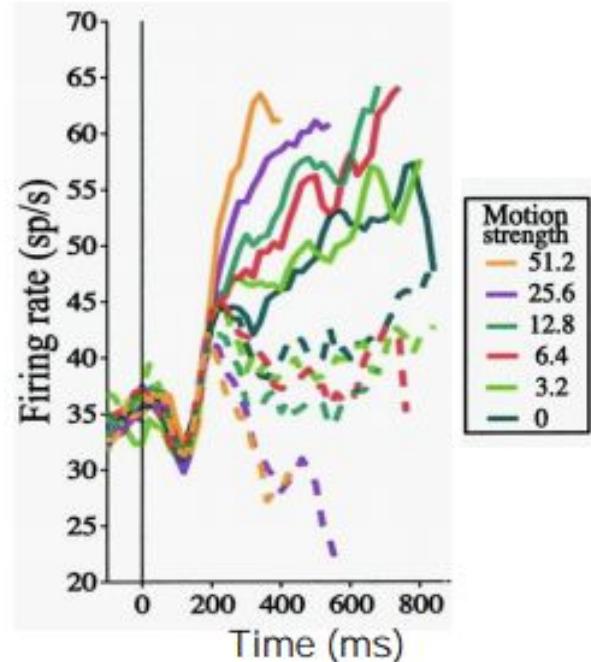
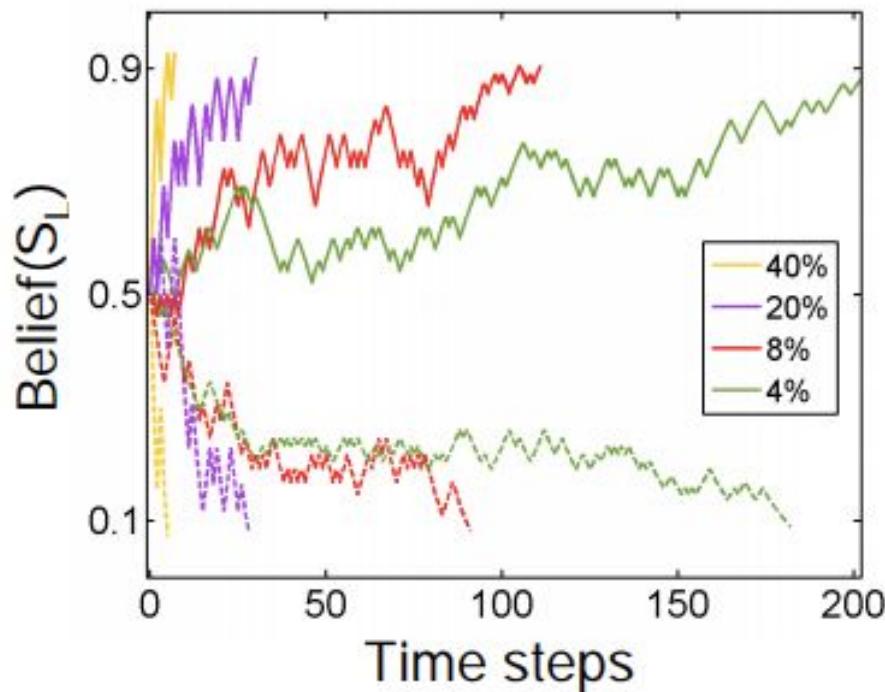


SE : belief state estimator

$b_t$  : belief state



# belief computation matches neural accumulation



# the past conditions the present

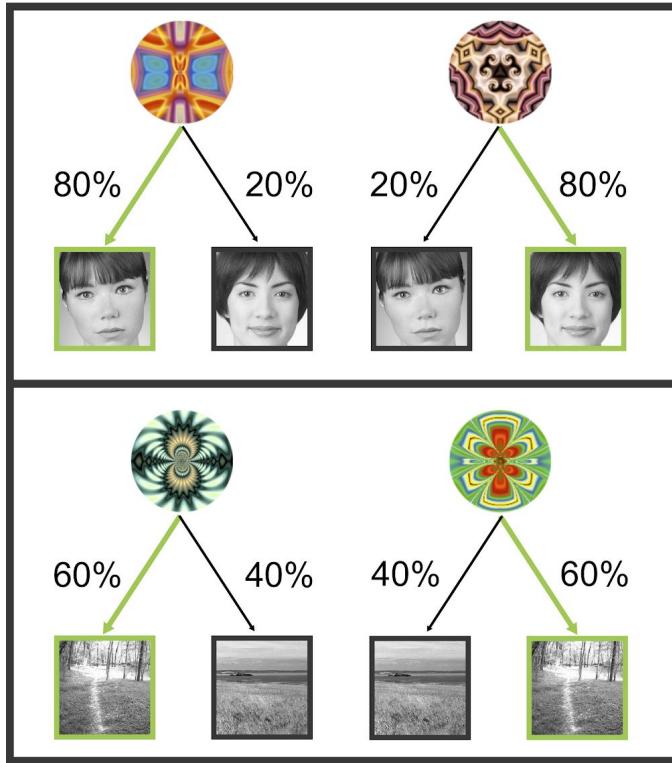
- sampling can resolve two sources of uncertainty
- 1. **estimating** the outcome of an action taken in the current state
- 2. **inferring** the current state

(note: these can be done simultaneously)

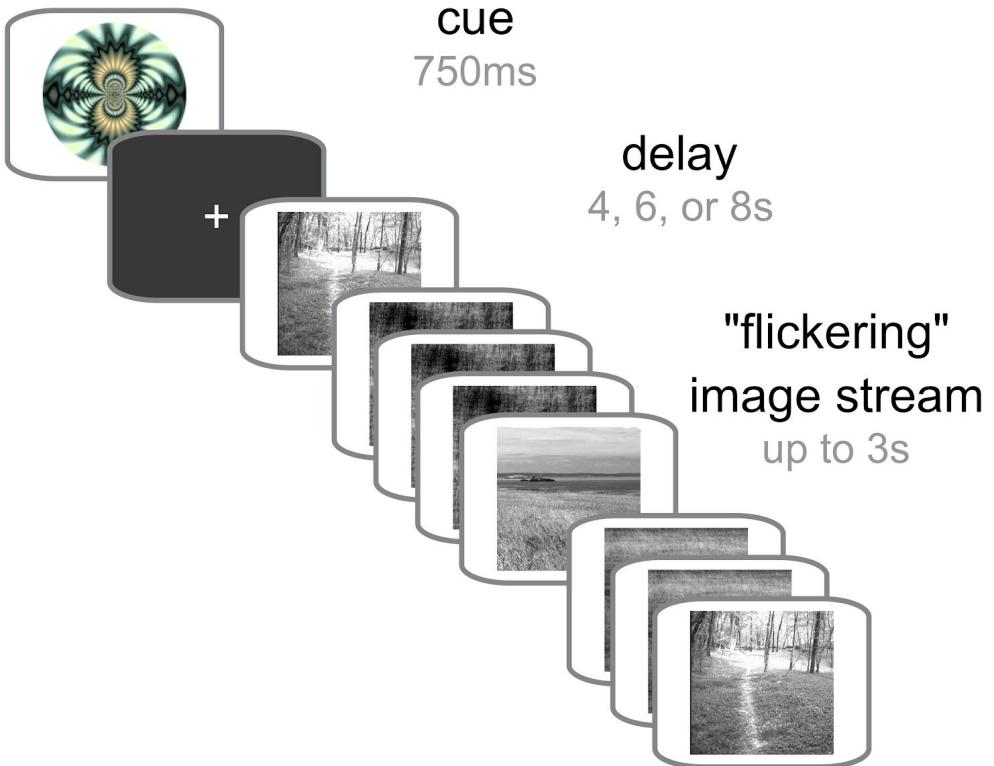
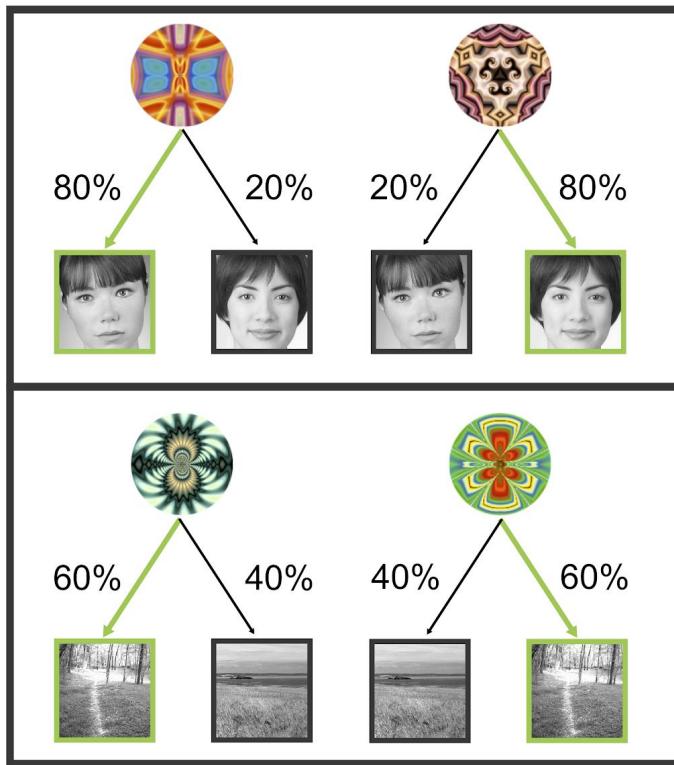
# outline

- I. computational role(s) of memories
- II. experimental evidence
- III. if time: open questions

# samples from memory resolve state uncertainty

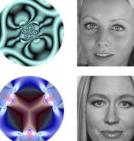
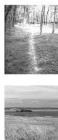


# samples from memory resolve state uncertainty

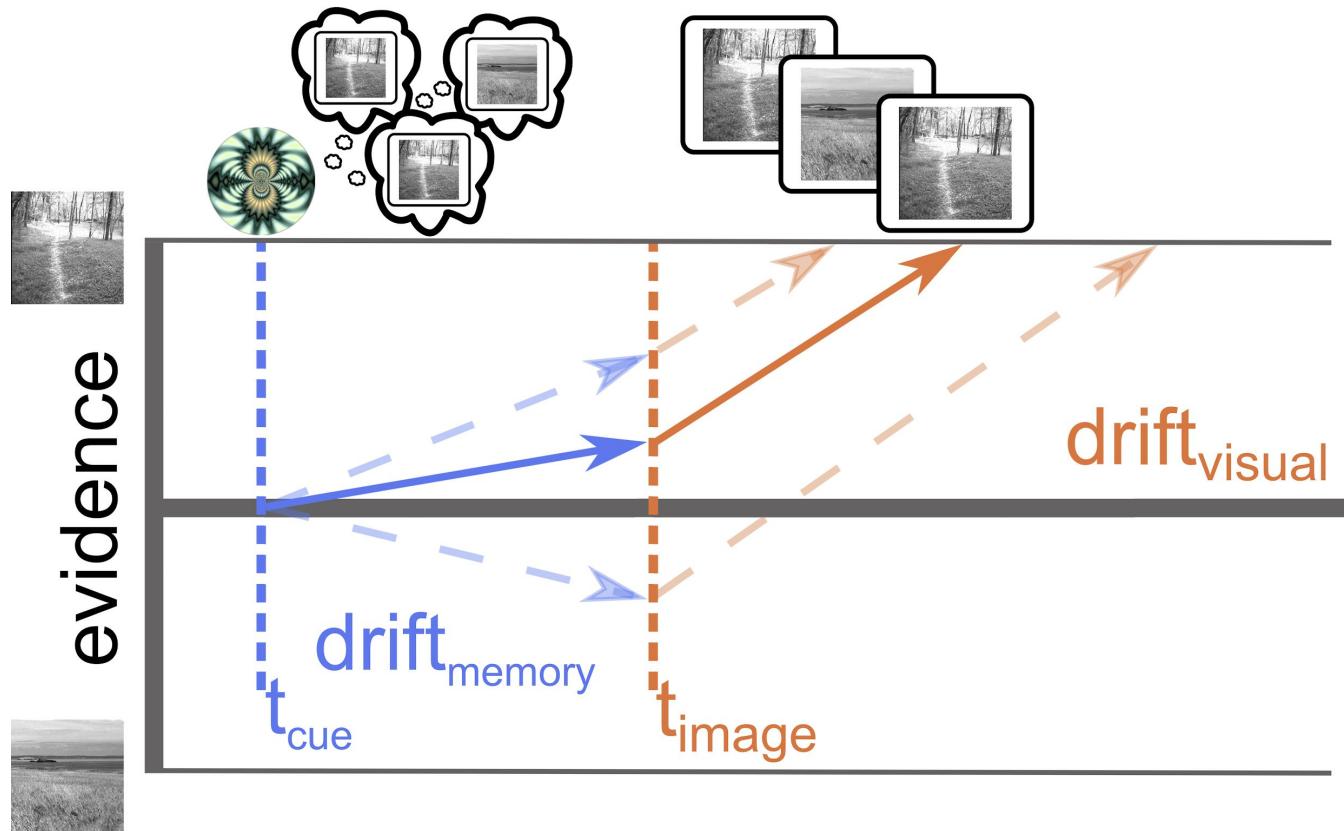


# samples from memory resolve state uncertainty

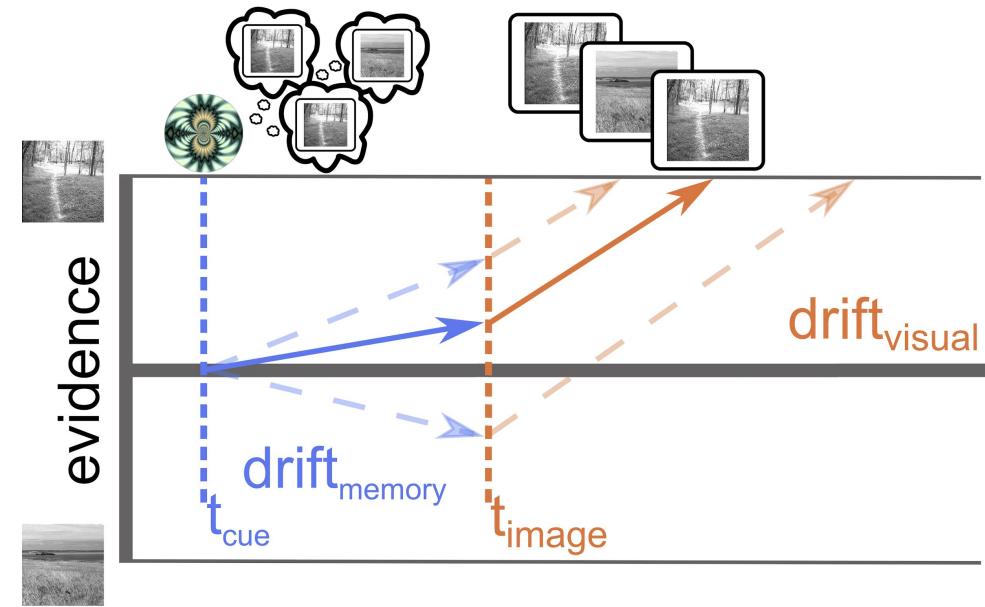
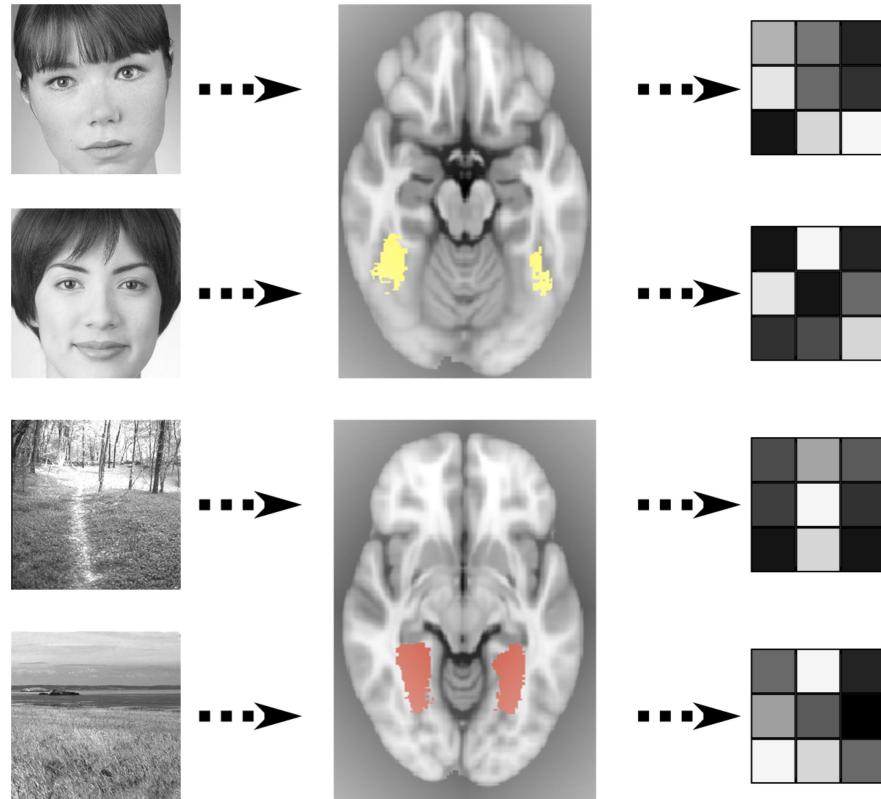


Stimuli	Cue validity	Sensory Evidence	Stimuli	Cue validity	time Sensory Evidence
 	50%	Strong	 	80%	Weak
 	70%	Weak	 	60%	Strong

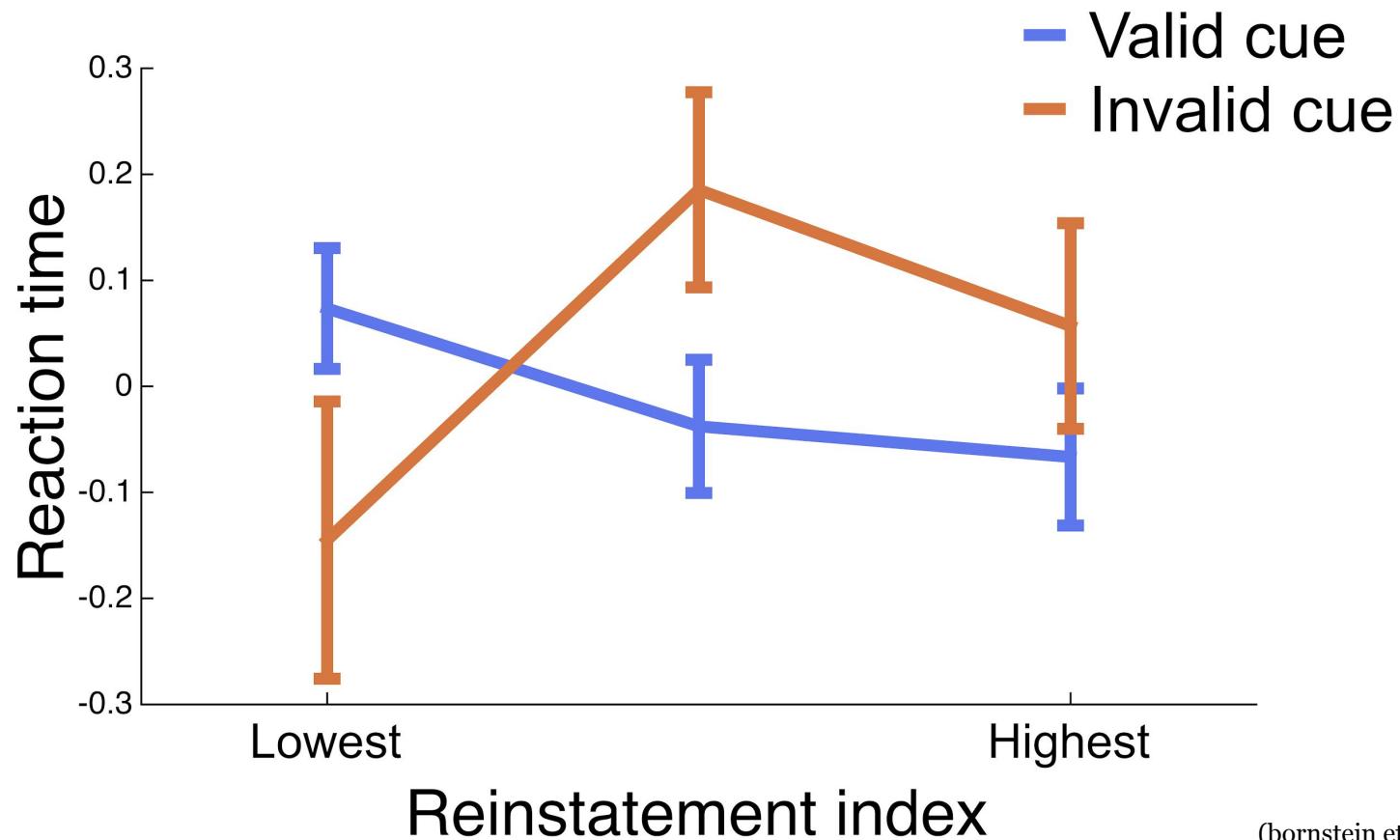
# two-stage uncertainty reduction



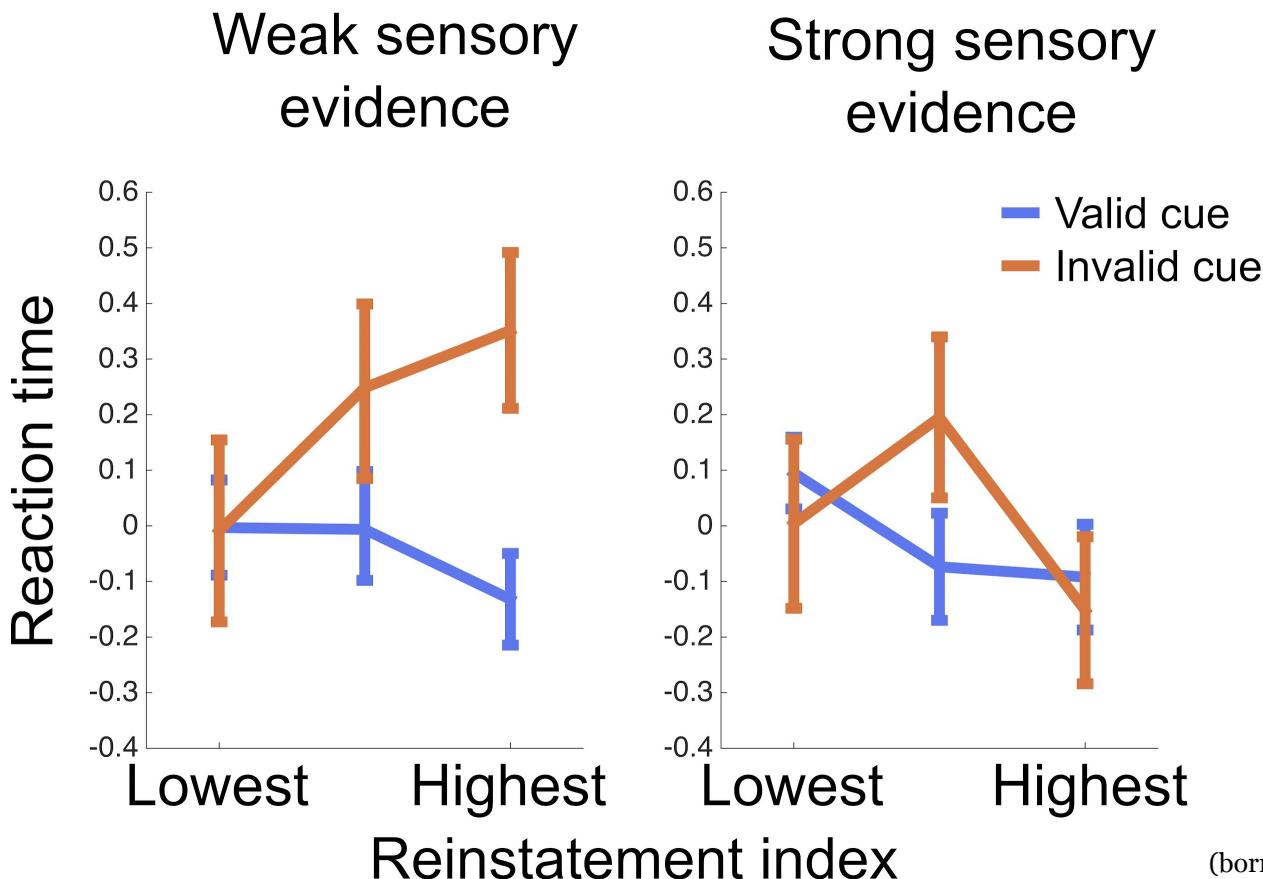
# two-stage uncertainty reduction



memory sampling reduces uncertainty



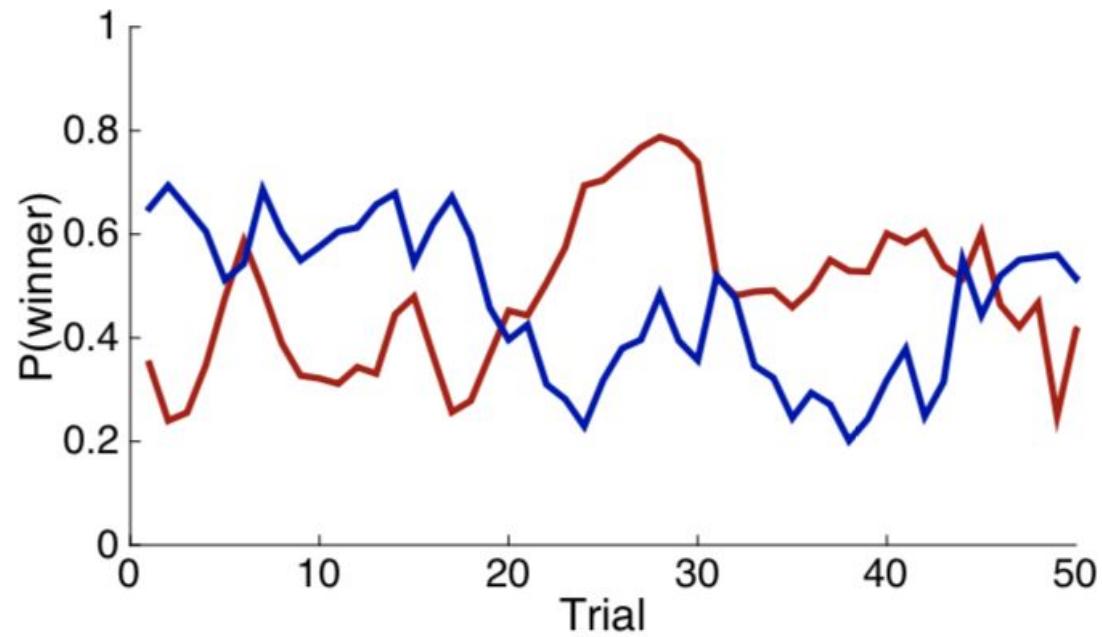
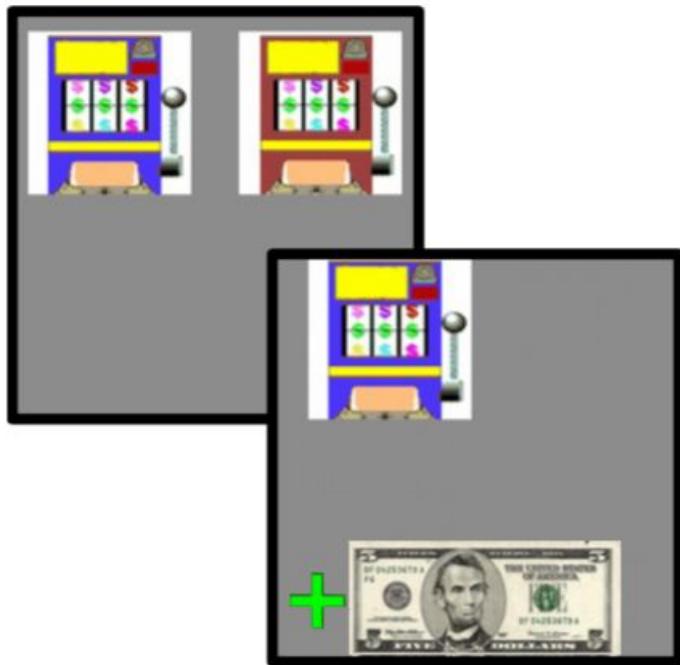
# memory sampling reduces uncertainty



# interim summary

- memory samples can reduce uncertainty about state
- pre-emptive memory-guided uncertainty reduction can reduce the need for sensory-guided uncertainty reduction
- multiple streams of evidence can be sampled to perform state inference
  - and are weighed by their informativeness

# memory sampling in a two-armed bandit task





\$0

$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$



\$0



$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$



\$2.50

$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$



-\$1.25

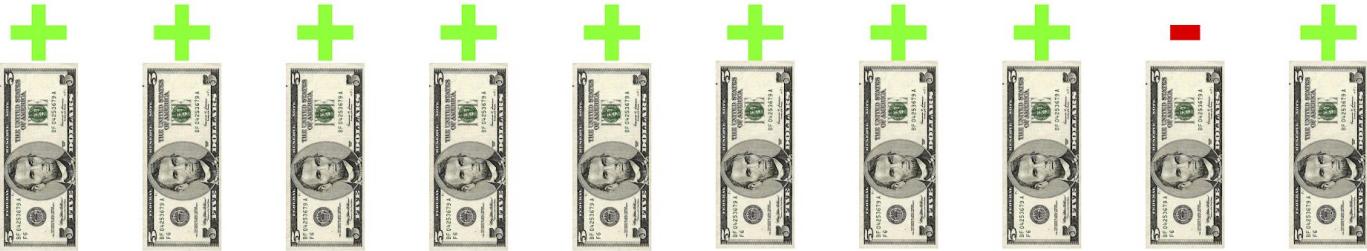
$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$



\$1.88

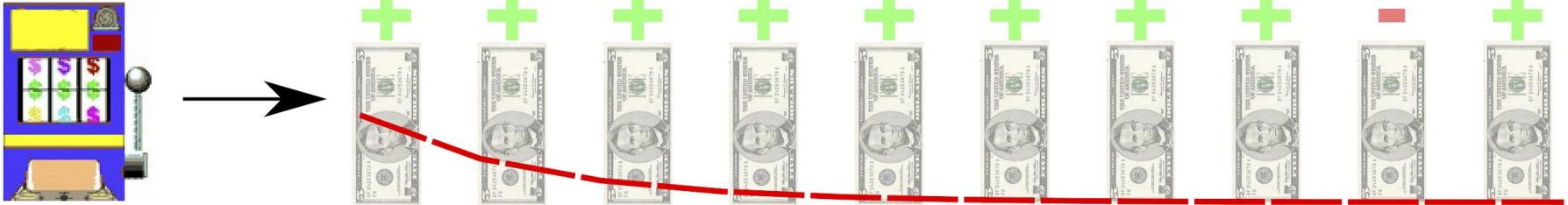


$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$



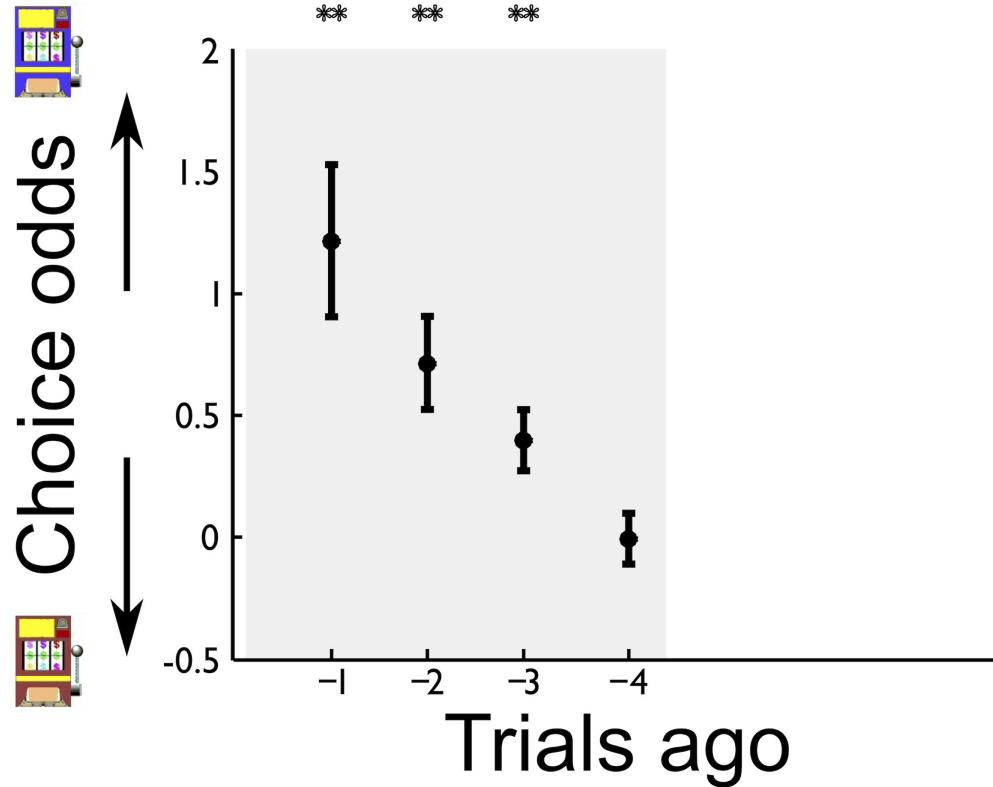
\$5

$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$

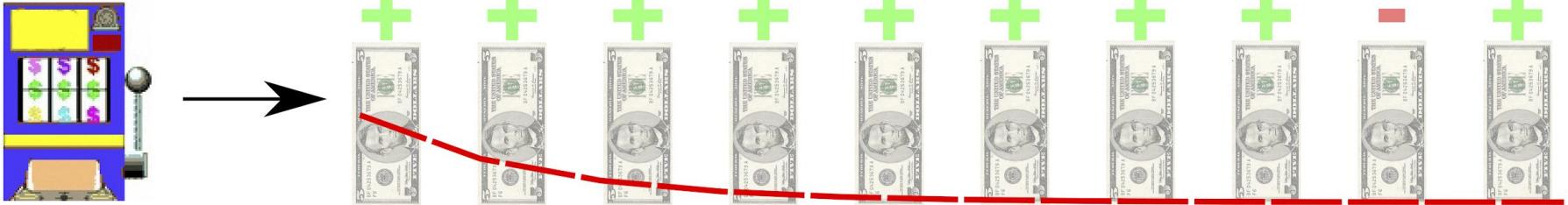


$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$

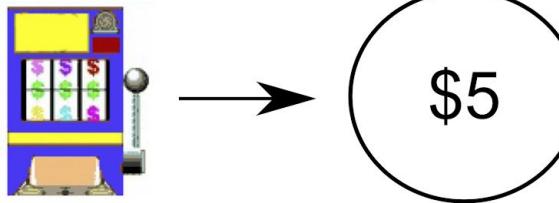
# Choices



$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$

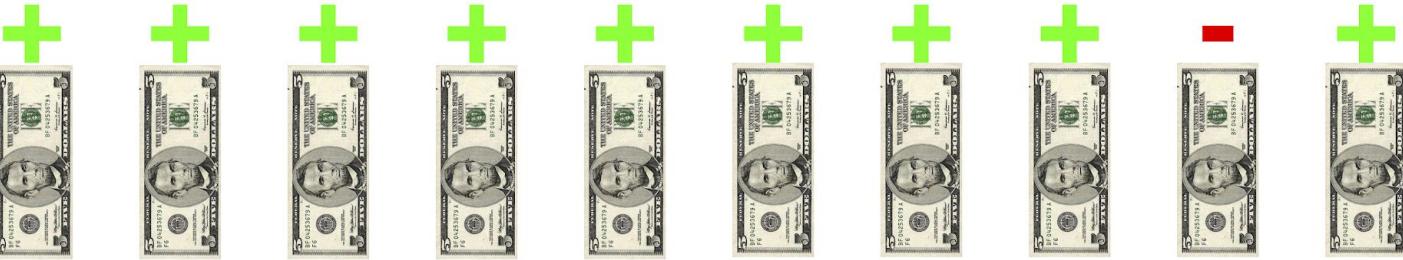


$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$

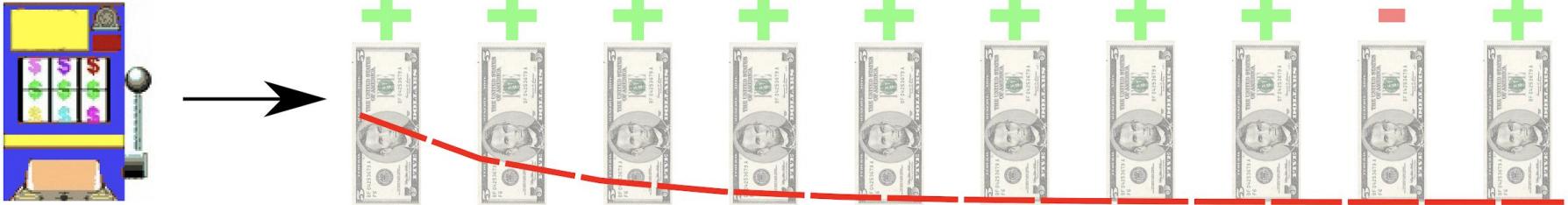


$$Q_t(a) = Q_{t-1}(a) + \alpha[R_t - Q_{t-1}(a)]$$

 $+$  $+$  $+$  $+$  $+$  $+$  $+$  $+$  $-$  $+$ 



$$P(Q_t(a) = R_i) = \alpha(1 - \alpha)^{t-i}$$

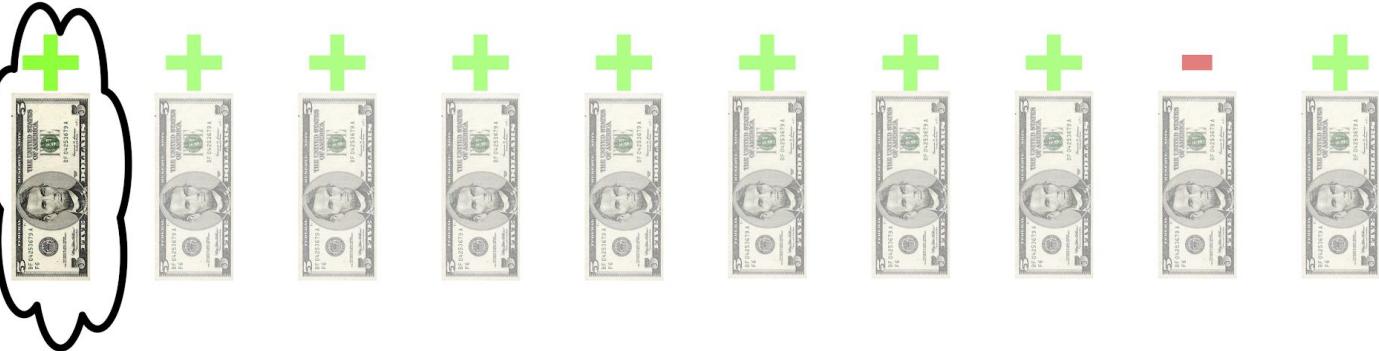


\$5

$$P(Q_t(a) = R_i) = \alpha(1 - \alpha)^{t-i}$$



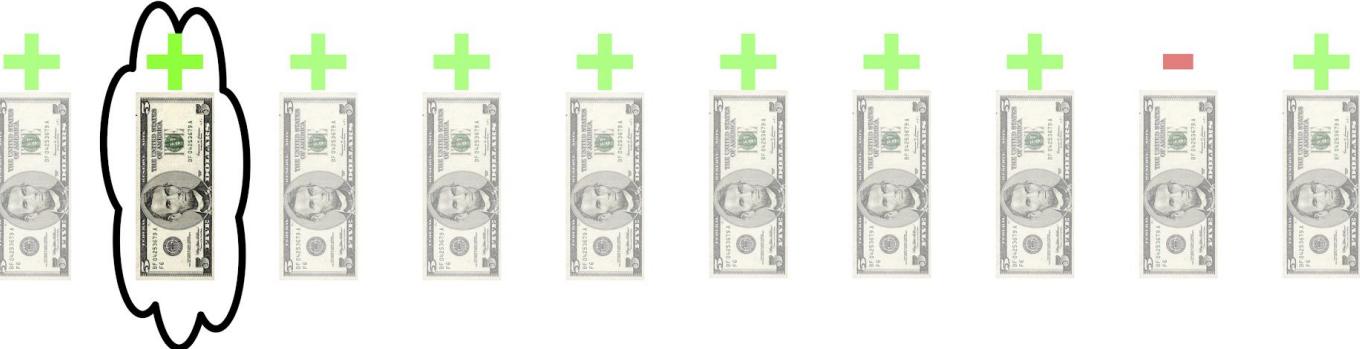
\$5



$$P(Q_t(a) = R_i) = \alpha(1 - \alpha)^{t-i}$$



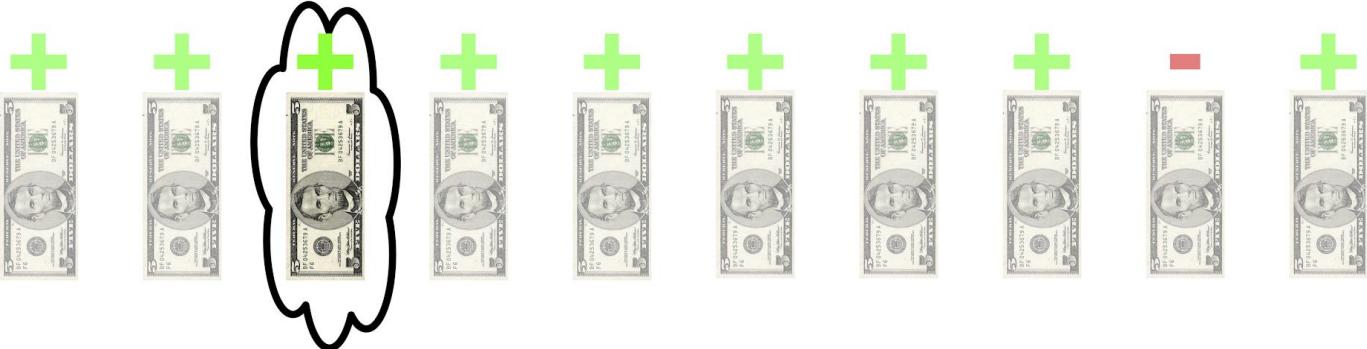
\$5



$$P(Q_t(a) = R_i) = \alpha(1 - \alpha)^{t-i}$$

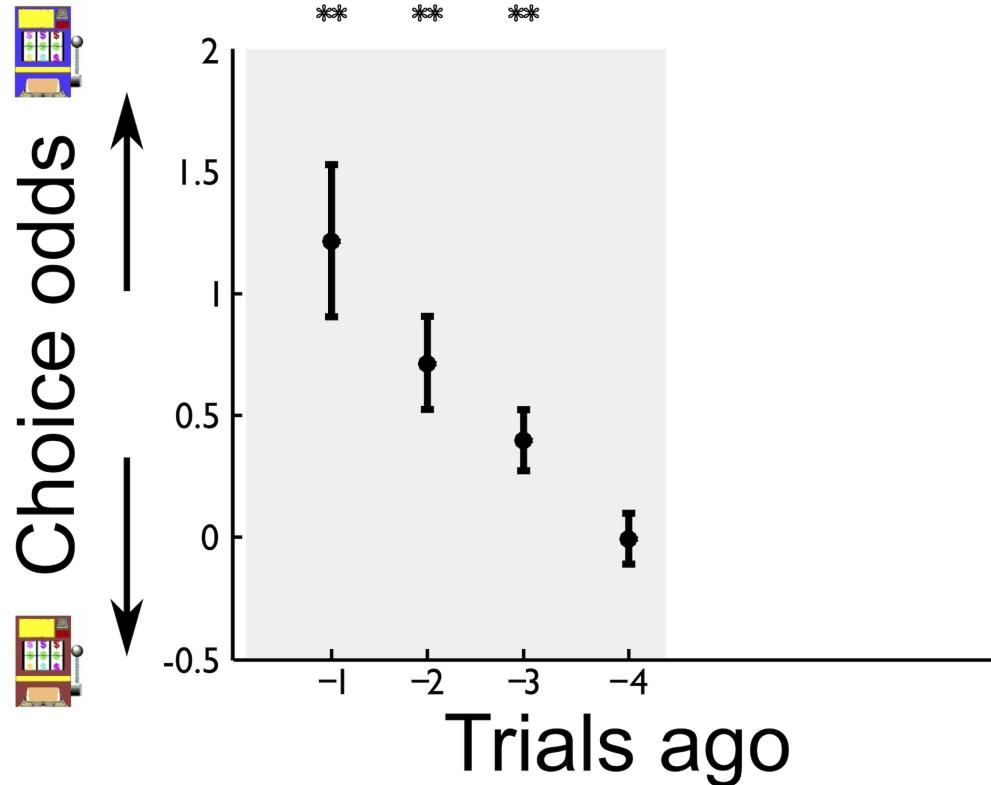


\$5

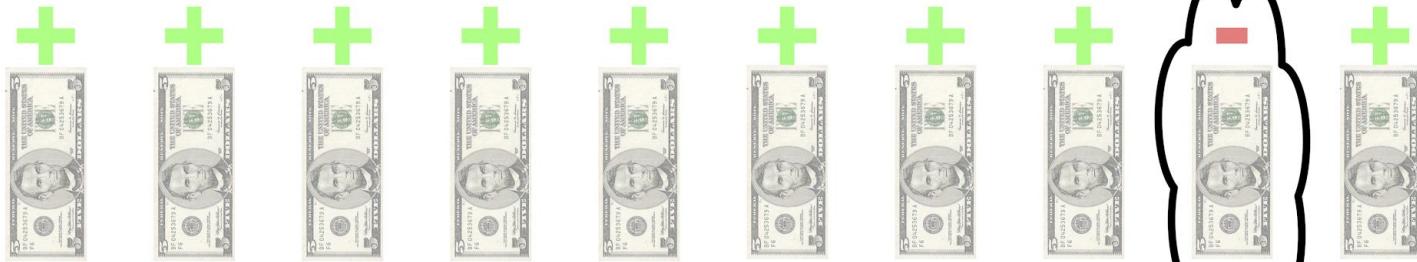


$$P(Q_t(a) = R_i) = \alpha(1 - \alpha)^{t-i}$$

# Choices

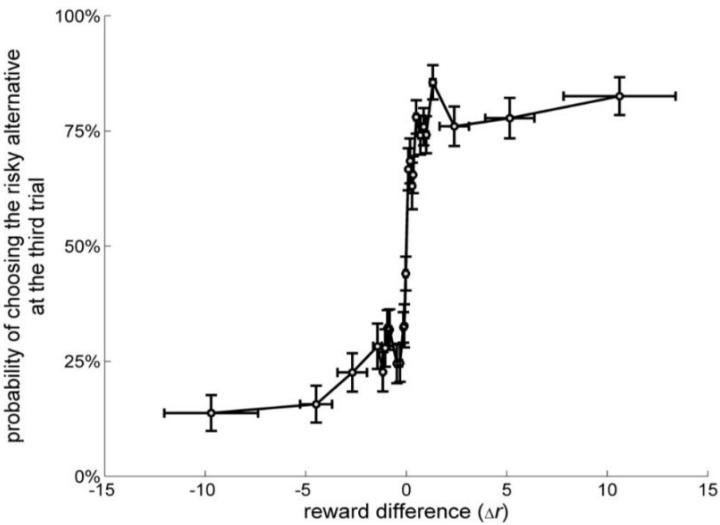
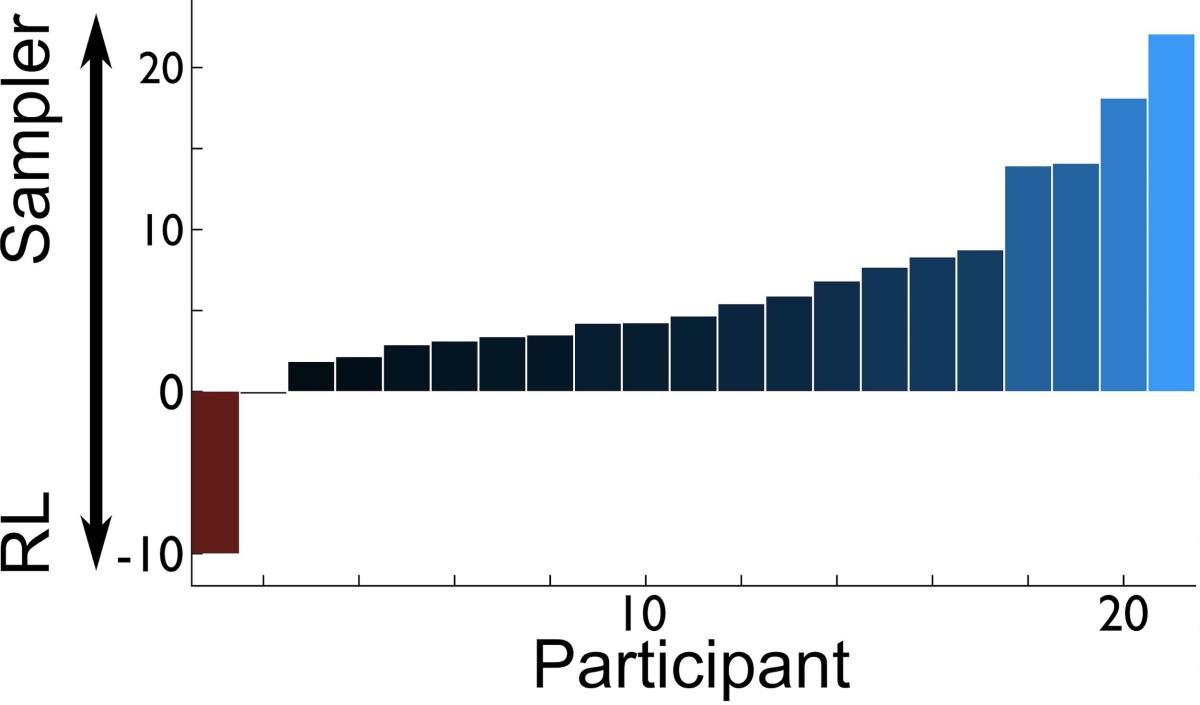


$$P(Q_t(a) = R_i) = \alpha(1 - \alpha)^{t-i}$$



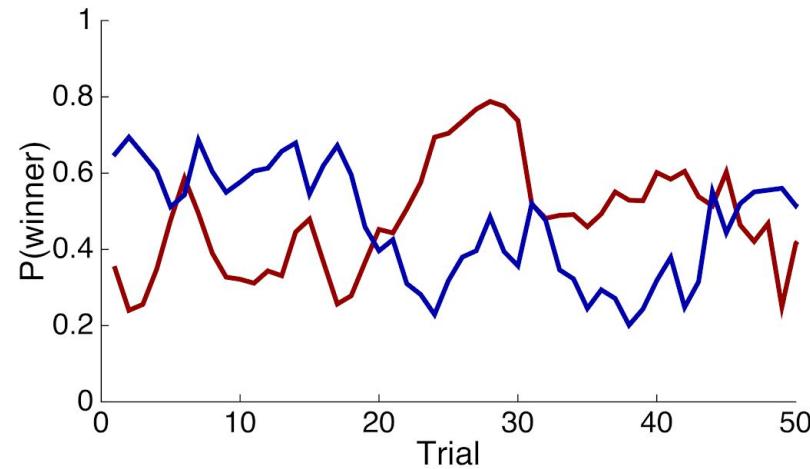
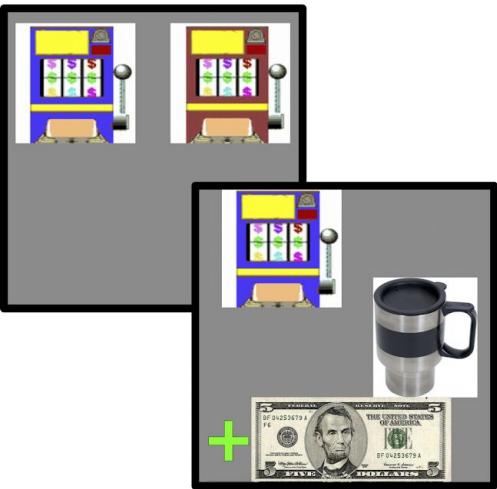
**-\$5**

$$P(Q_t(a) = R_i) = \alpha(1 - \alpha)^{t-i}$$

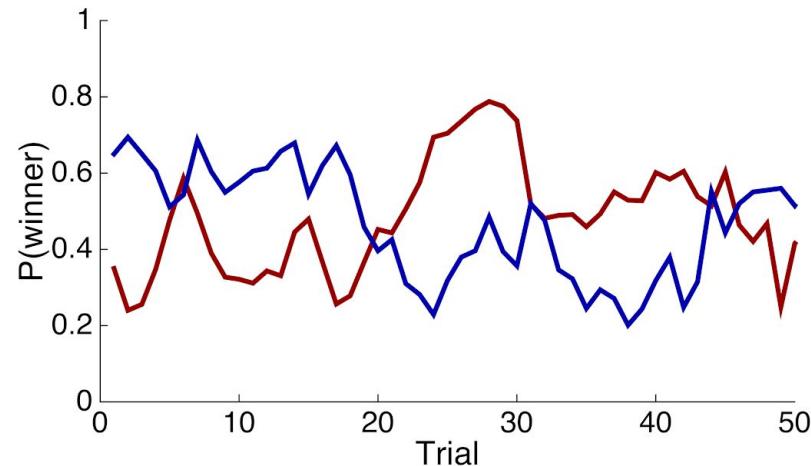
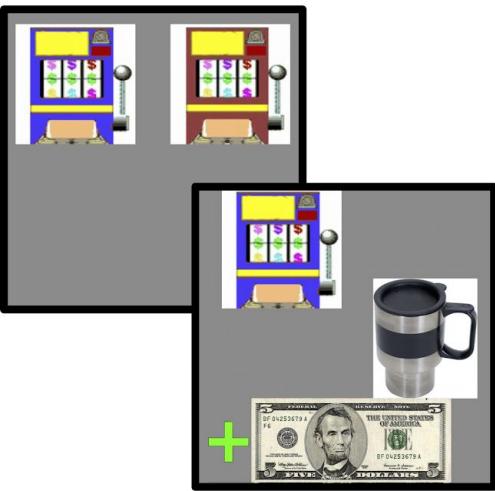


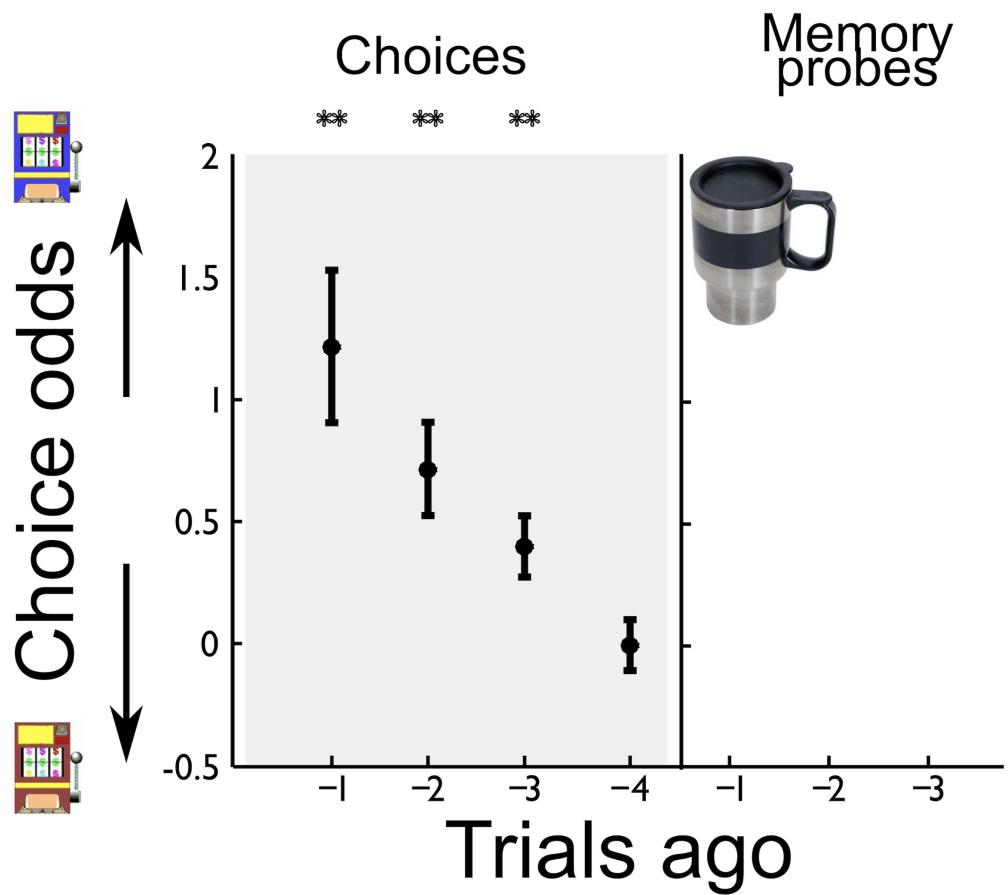
(bornstein et al 2017)

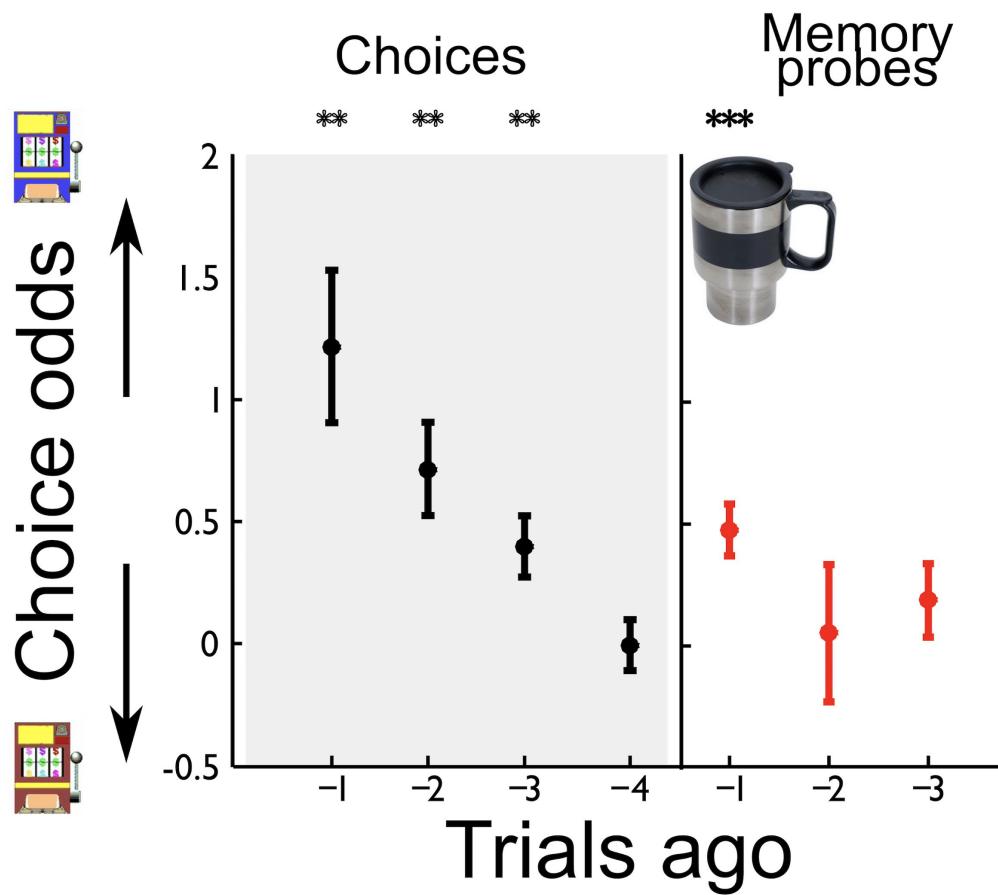
# “ticket bandit” task

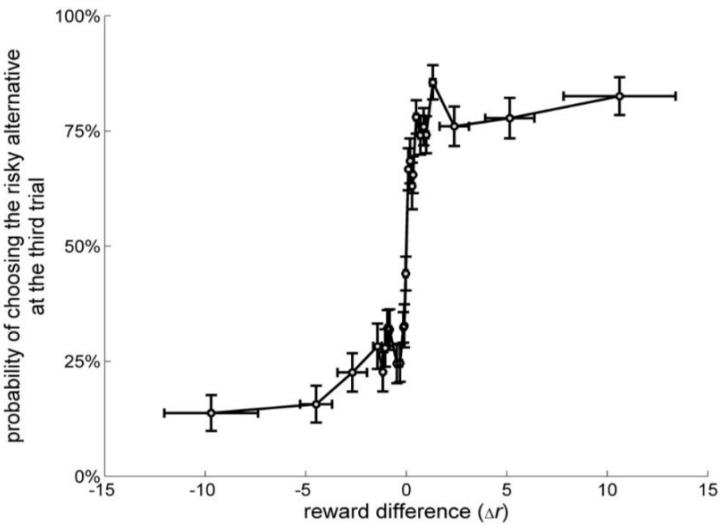
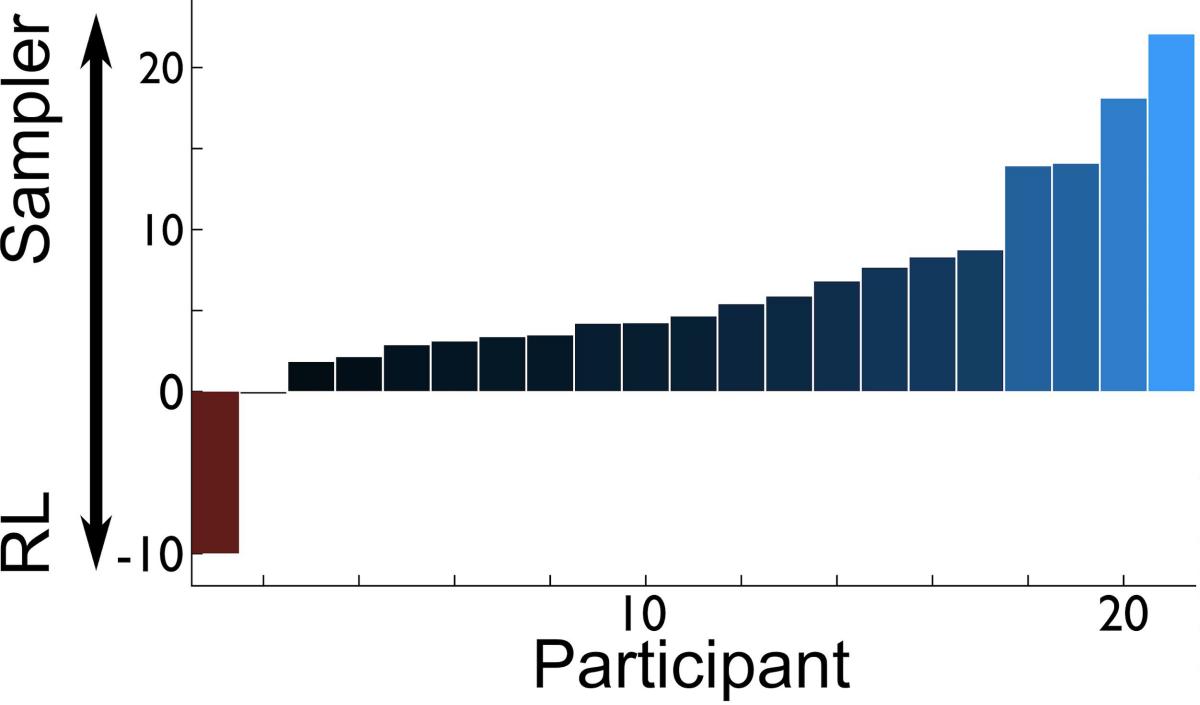


# “ticket bandit” task

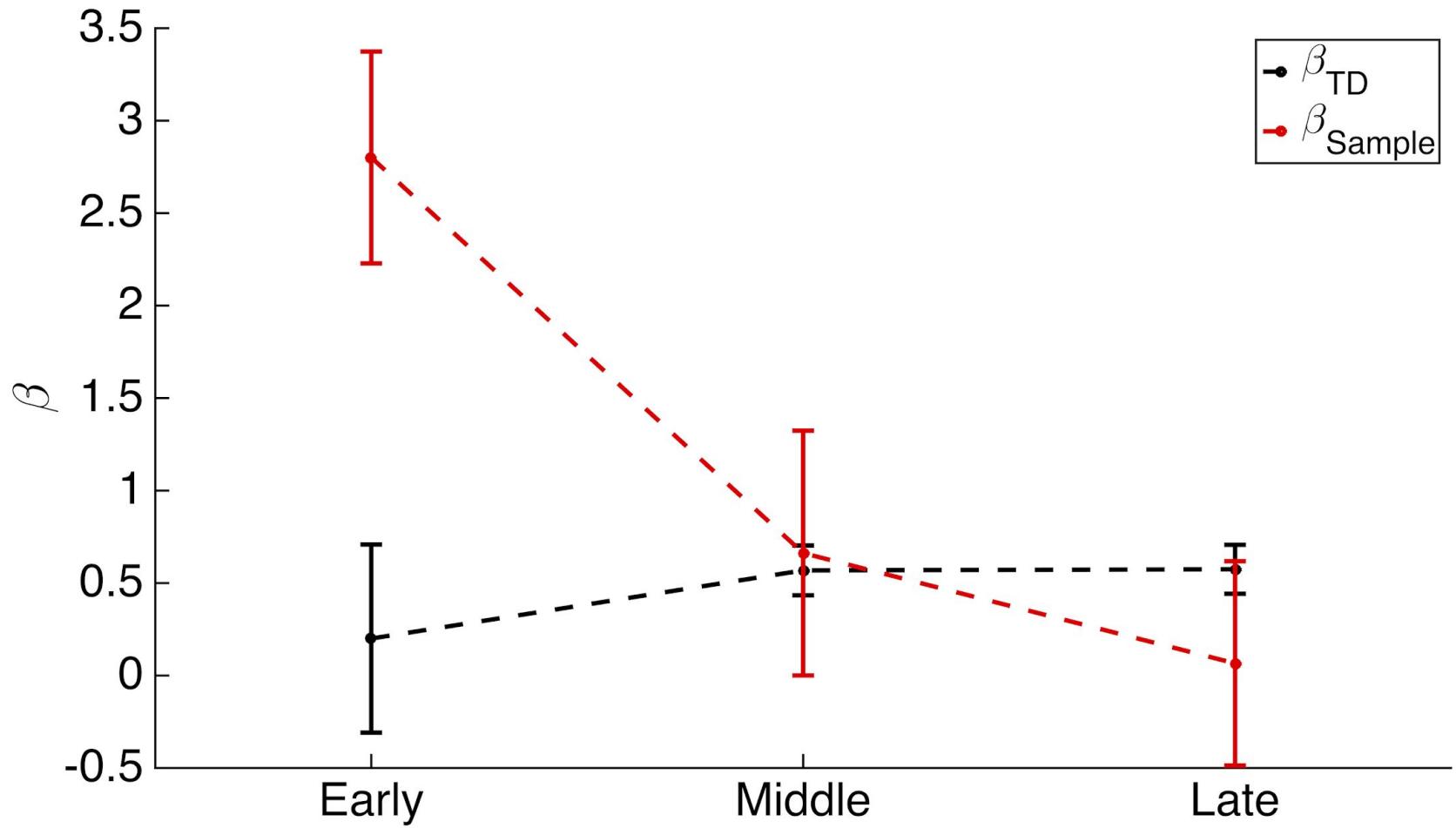








(bornstein et al 2017)



# interim summary

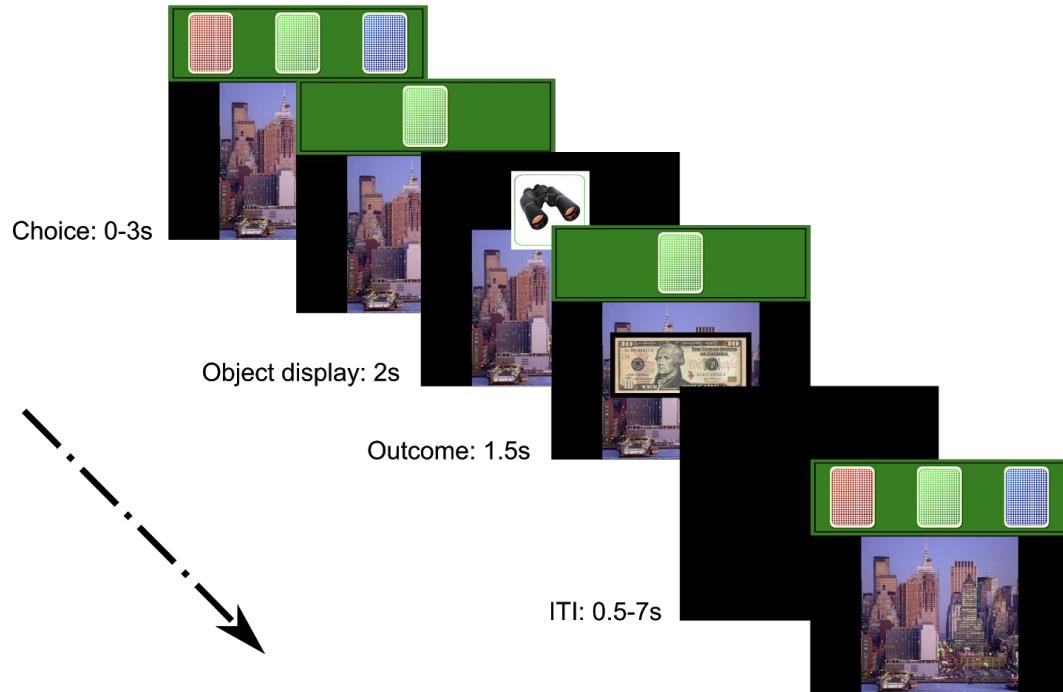
- memory sampling can be biased by incidental reminders of individual experiences
- suggests role for *episodic* memory
- less influence of episodic memory as task becomes well-learned

## next: more to a memory

- episodic memories are more than just actions and values
- also carry *context*: associations with other, nearby, experiences (howard & kahana 2002)
- does this other episodic information *also* affect decisions?

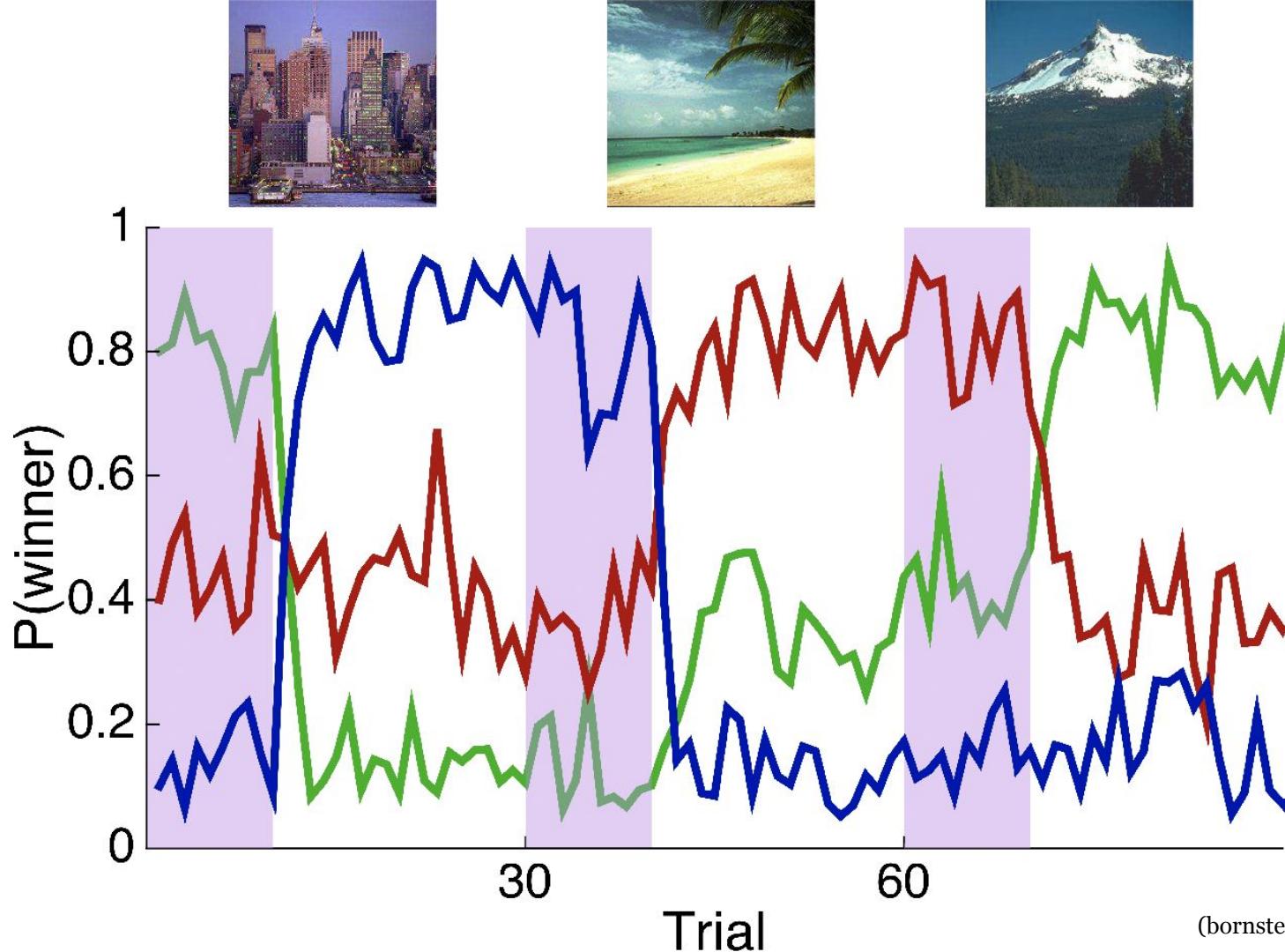
# “context bandits” task

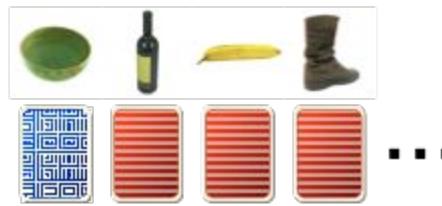
## Context room choices



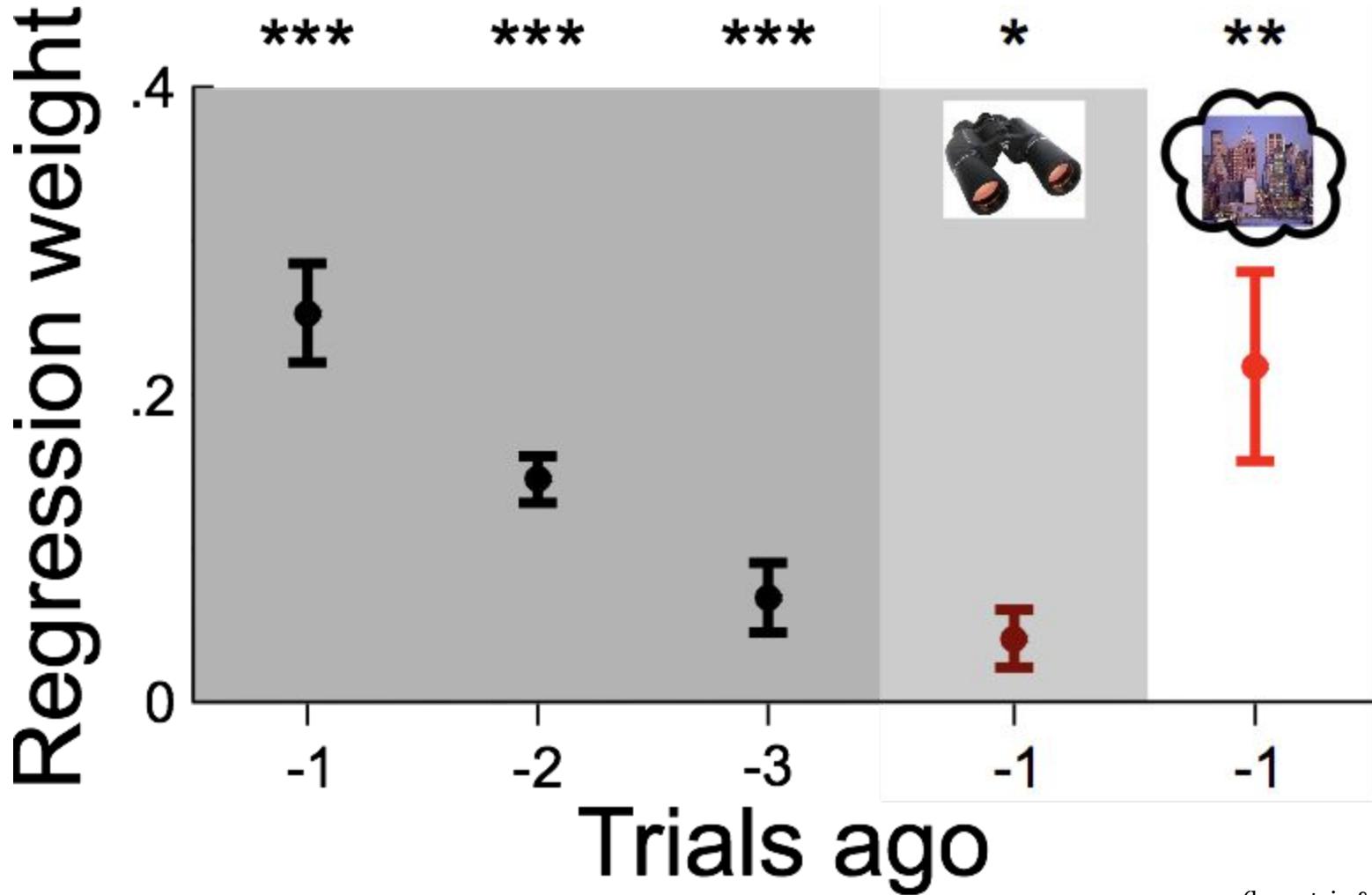
## Memory probes

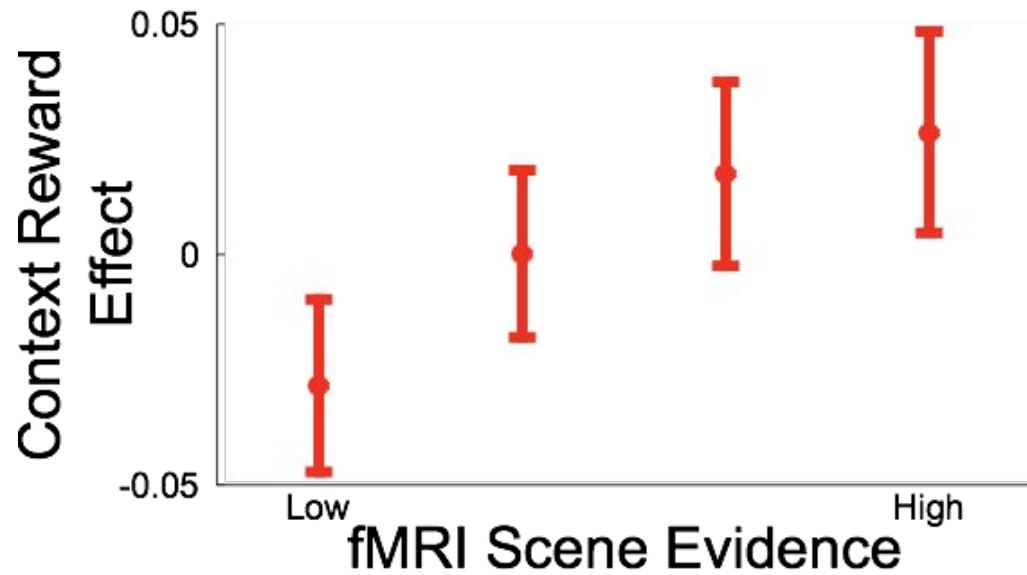
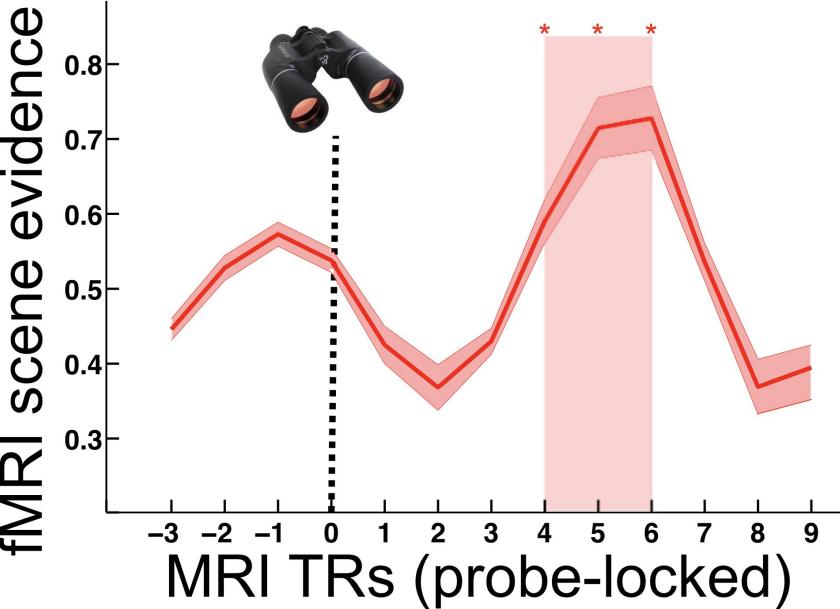


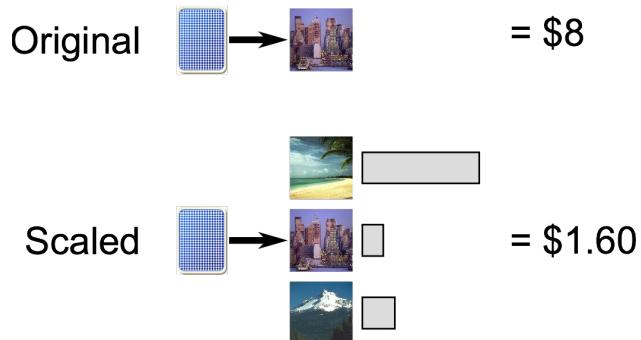
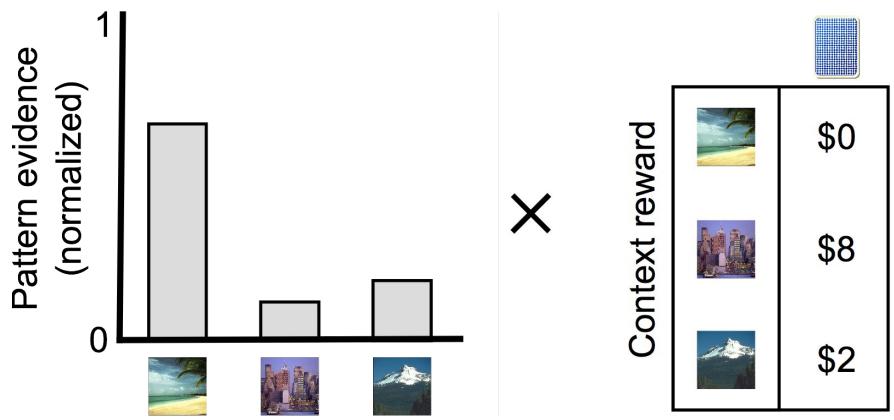
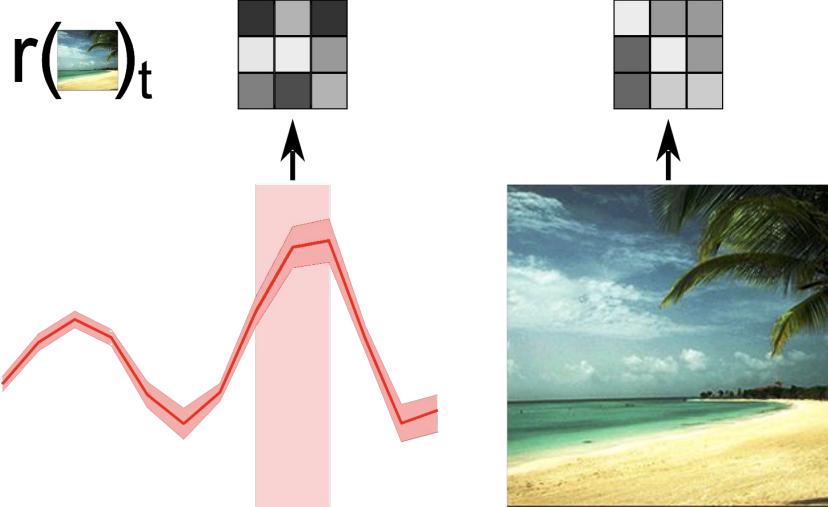
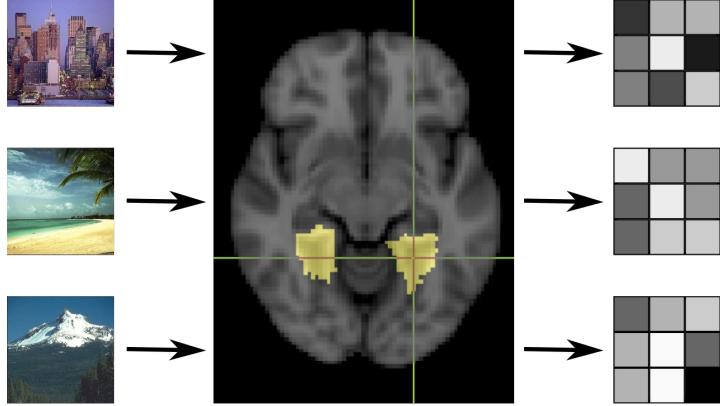




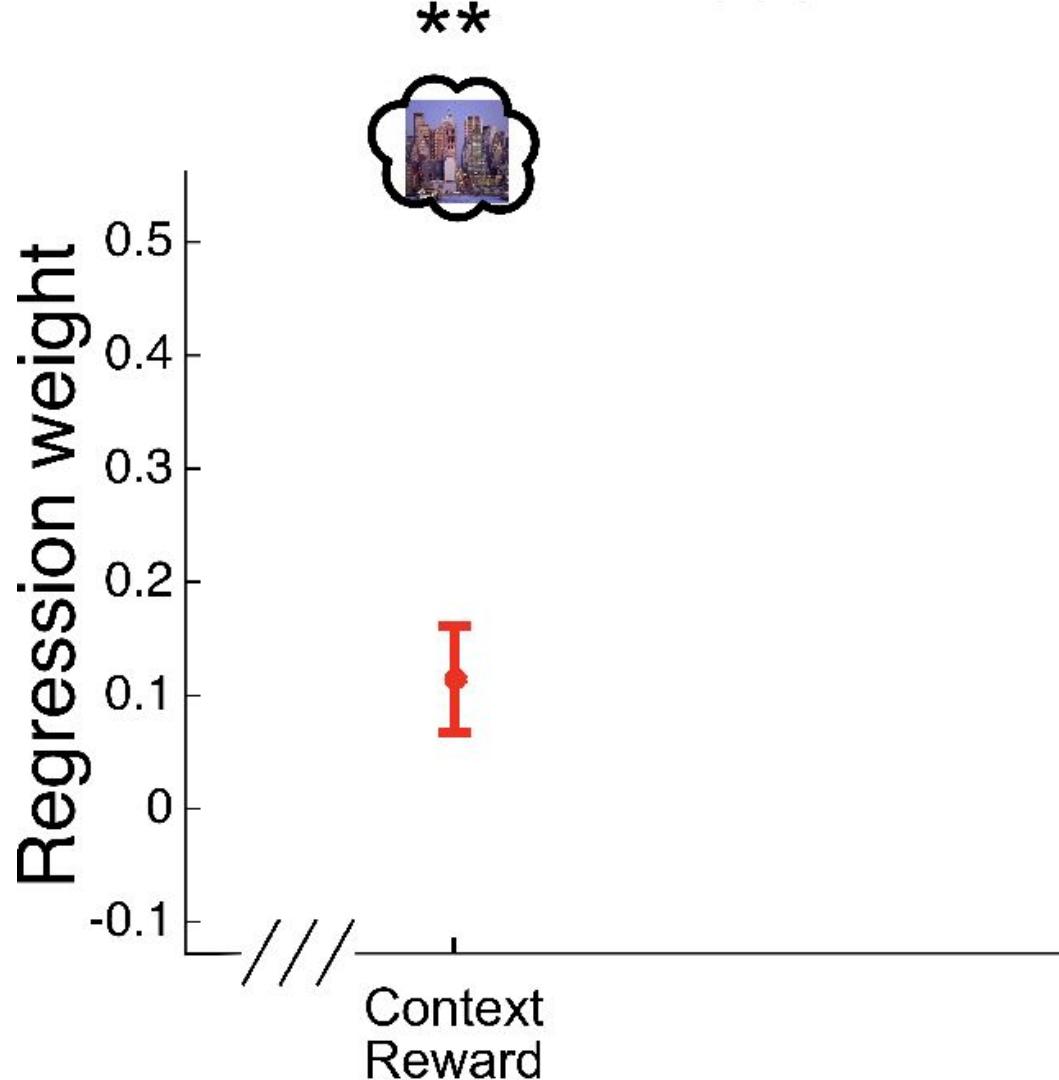
(bornstein & norman 2017)

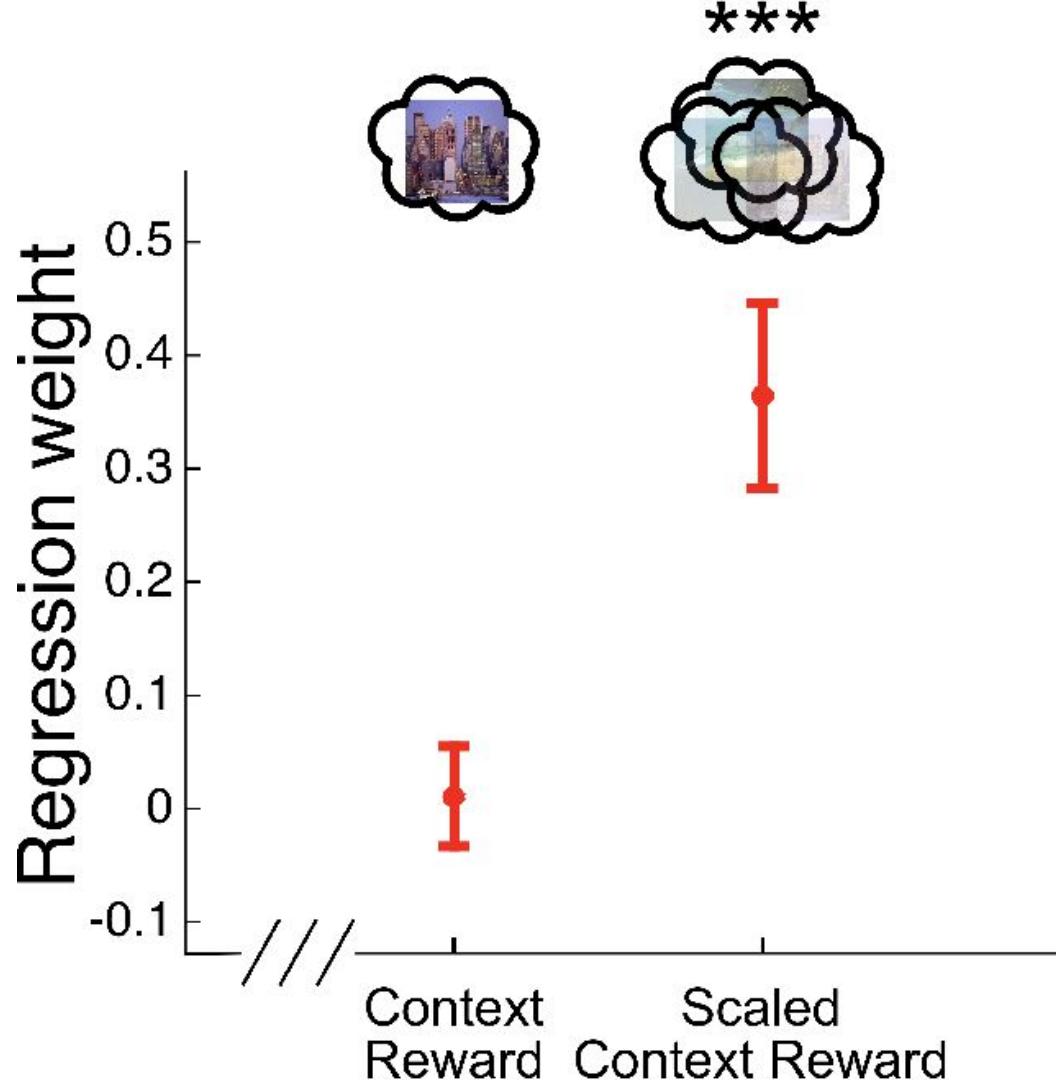






(bornstein & norman 2017)





# more to a memory

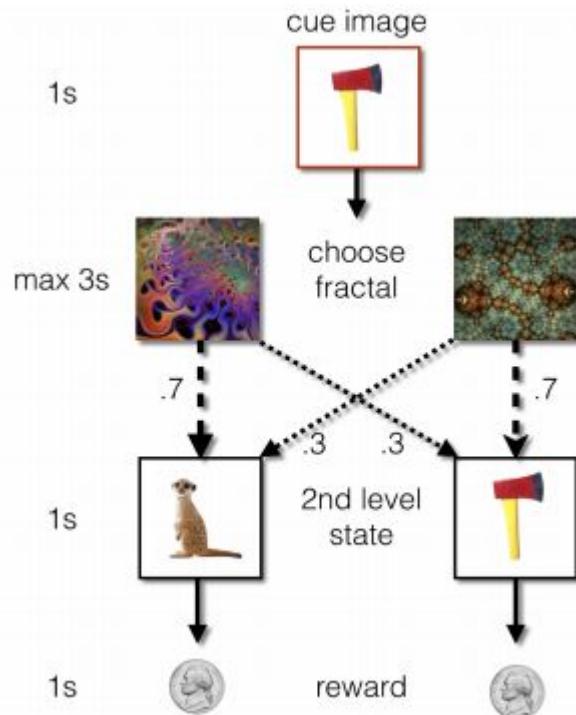
- episodic memories are more than just actions and values
- also carry *context*: associations with other, nearby, experiences (howard & kahana 2002)
- the actions and values in these other experiences can, when they are remembered, also bias decisions
- memory is also not veridical
- mistaken memories also bias decisions

# last: episodic model-free, or episodic model-based?

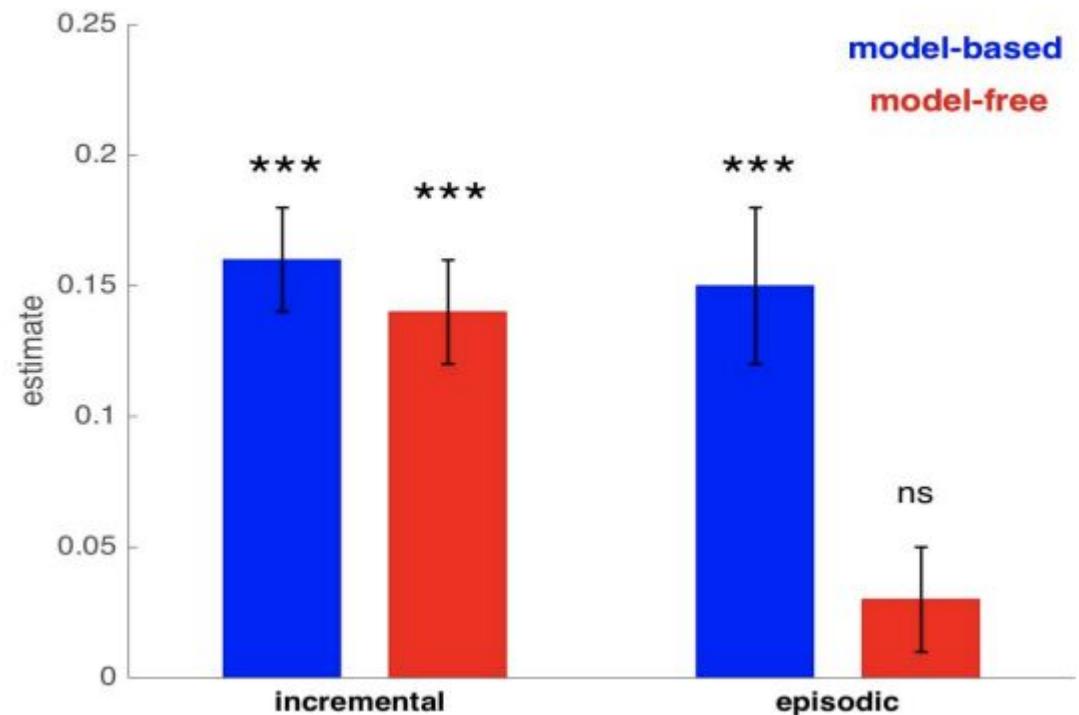
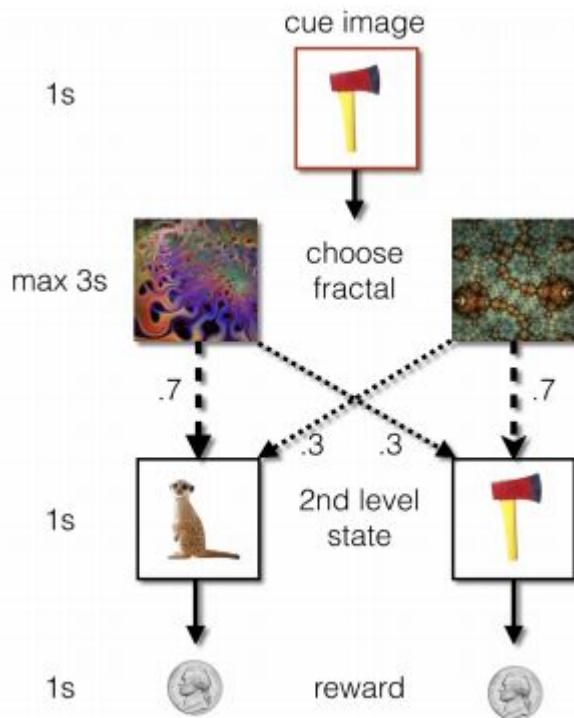
is episodic memory sampling “model-free” — guiding towards a specific past action-value pair

or is it “model-based” — flexibly combined with current transition structure

# episodic two-step task



# episodic two-step task



see ritter, wang et al 2018 for artificial RNN “discovering” e-mb, but not e-mf

(vikbladh et al 2017)

# **outline**

- I. computational role(s) of memories
- II. experimental evidence
- III. if time: open questions

# open q: underlying representations

- neuroscience: a continuum of representations
  - full *semantic* or *grammar* (daw et al 2005; glascher et al 2010)
  - flexible action sequences (doya et al 2002; bornstein & daw 2012)
  - flexible S-S sequences (“successor representation”; dayan 1993; bornstein & daw 2012, 2013)
  - episodes (lengyel & dayan 2008; bornstein & daw 2013; bornstein et al 2017, 2018; vikbladh et al 2018a,b; ritter et al 2018)
- computational RL: many varieties of “model”
  - *sample* models v *distribution* models

## open question: distributional or sample representation?

- is a given behavior guided by a distributional or a sample-based representation?
- this can be difficult to distinguish experimentally, at the level of aggregate behavior, because the predictions are, quantitatively, very similar
- this can even be difficult to distinguish in neural representation. (ma et al 2006; fiszer et al 2010)
- further frontiers: not just states or plans (e.g. categories –  
[http://www.j-paine.org/dobbs/why\\_be\\_interested\\_in\\_categories.html](http://www.j-paine.org/dobbs/why_be_interested_in_categories.html))
- but: general principles apply across representations – learning incrementally, by experience, direct or simulated

# further reading

- all cited papers are at: <http://aaron.bornstein.org/ccnss/>
  - plus some others i think are worth reading
- next edition of sutton & barto book:  
<http://incompleteideas.net/book/the-book-2nd.html>
- forthcoming book: “goal-directed decision making: computations and neural circuits” — ask me for pdfs in a couple months
  - table of contents: <http://aaron.bornstein.org/gdcnc/>
- happy to talk about research any time → [aaron@bornstein.org](mailto:aaron@bornstein.org)

# open q: trajectory sampling?

- no one has yet decoded *multi*-step decisions, either offline or online
- thus it's an open question whether planning is trajectory sampling, or single-step value-function updates

# open q: how do we arbitrate?

- uncertainty-based arbitration thought to require an arbiter
- little evidence for one
- “bottom-up” arbitration? (eisenreich et al 2017; bornstein & cohen in prep)