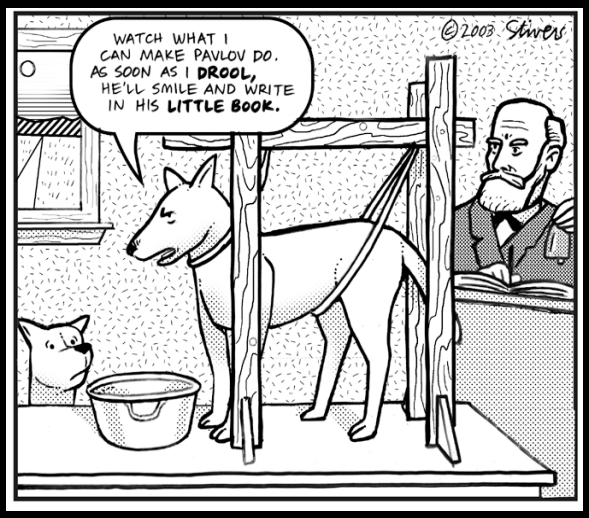
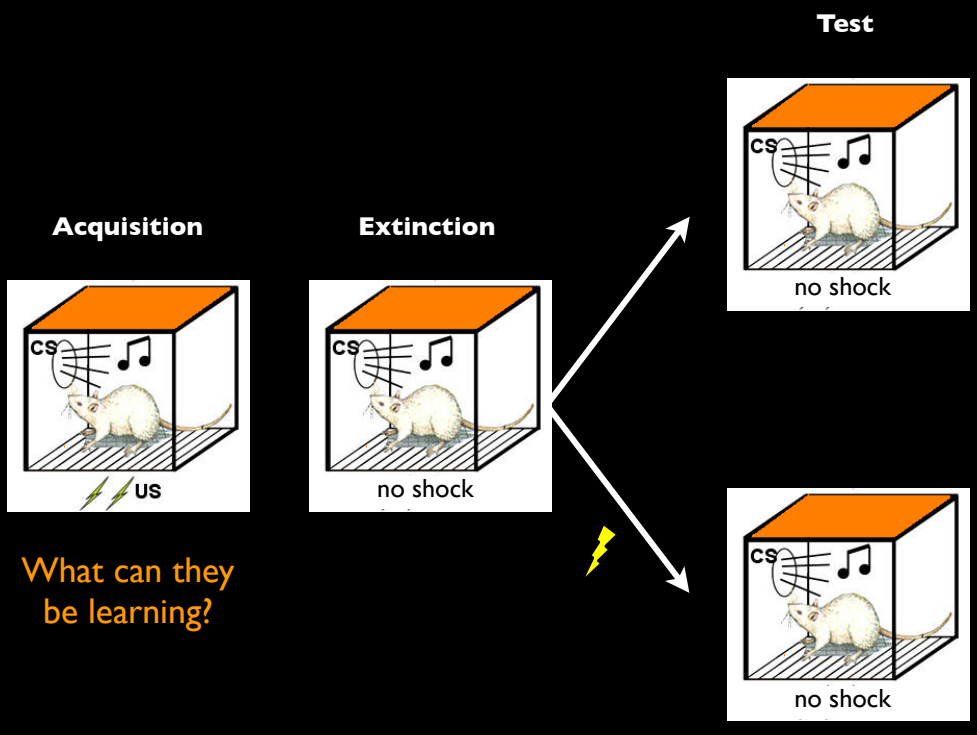
What are animals *really* learning? The case of extinction



PSY/NEU338: Animal learning and decision making: Psychological, computational and neural perspectives

Case in point: Reinstatement



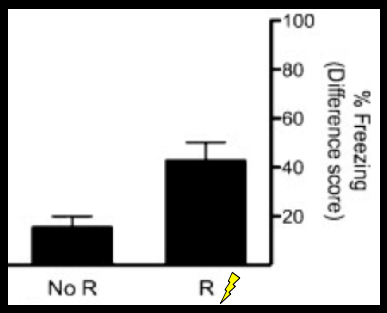
What can they be learning?

Extinction ≠ Unlearning

Acquisition

Extinction

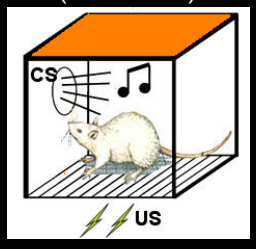
Test



Storsve, McNally & Richardson, 2012

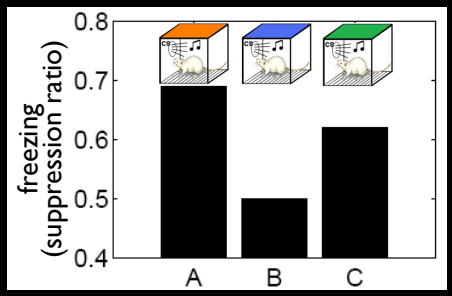
Extinction ≠ Unlearning

Acquisition
(Context A)



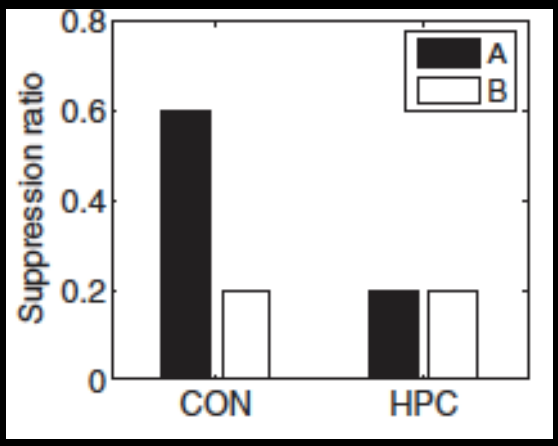
Extinction
(Context B)

Test (renewal)



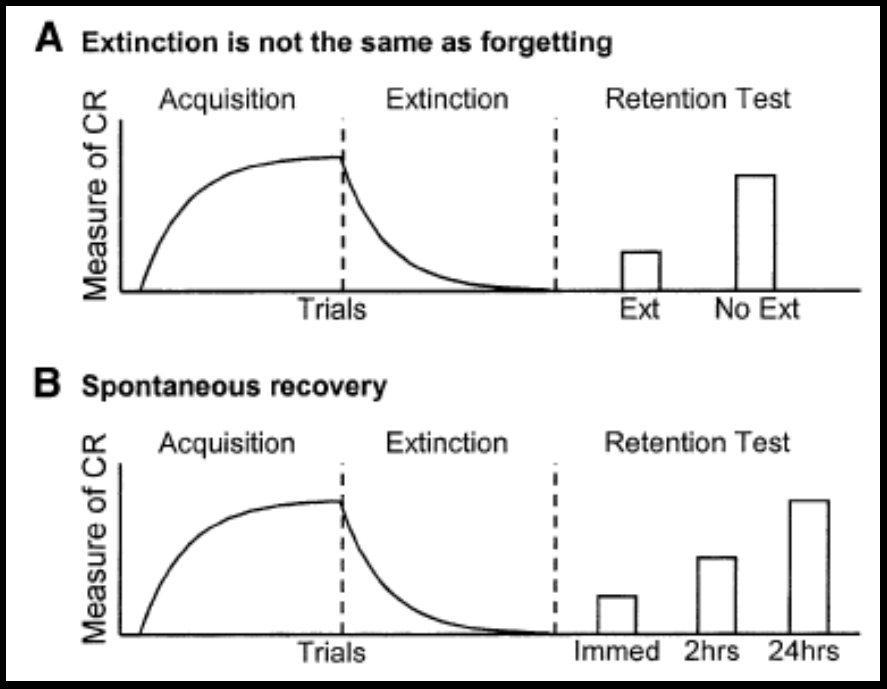
Bouton & Bolles 1979

And… this depends on integrity of hippocampus

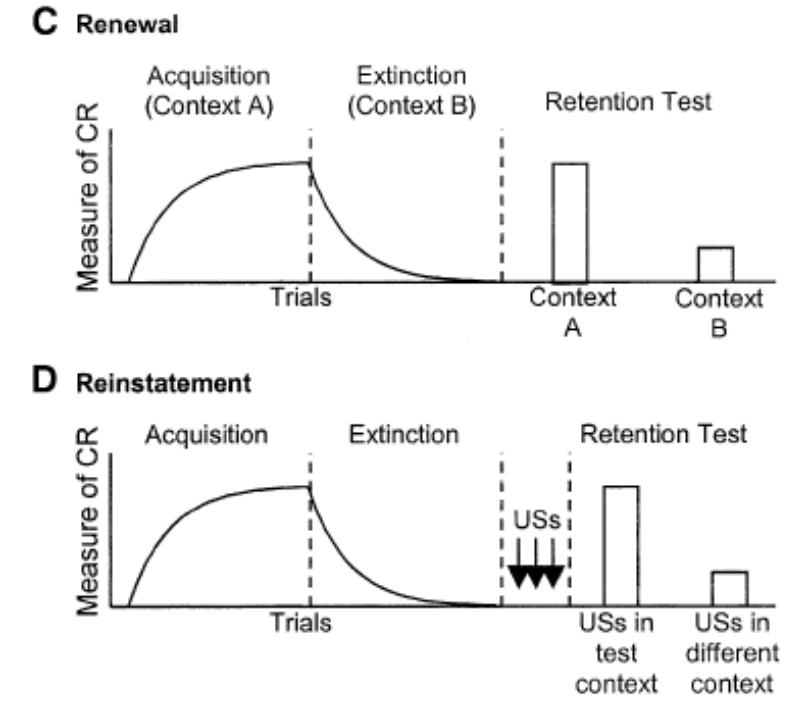


Ji & Maren 2005

Extinction does something, but not what we want it to do



Extinction does something, but not what we want it to do



Redish: Animals also learn the *states* of the task

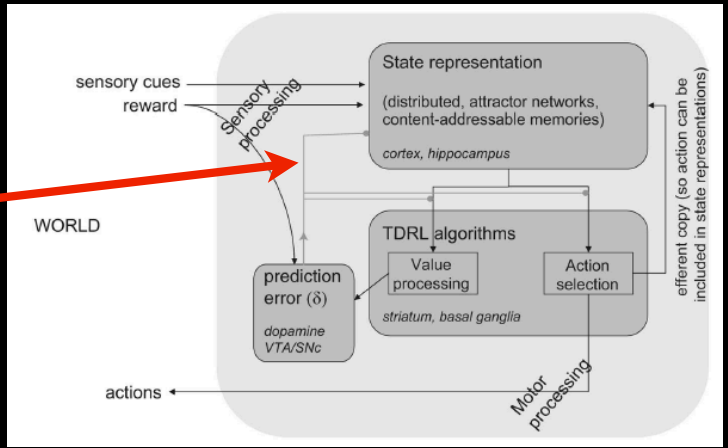
Psychological Review
2007, Vol. 114, No. 3, 784–805

0033-295X/07/$12.00 DOI: 10.1037/0033-295X.114.3.784

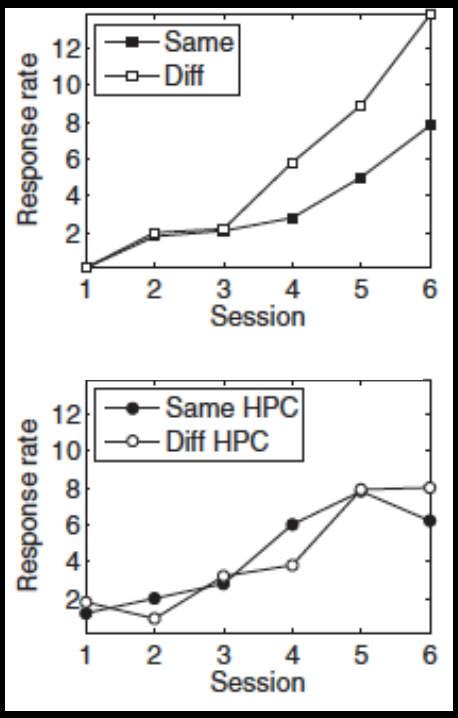
Reconciling Reinforcement Learning Models With Behavioral Extinction and Renewal: Implications for Addiction, Relapse, and Problem Gambling

A. David Redish, Steve Jensen, Adam Johnson, and Zeb Kurth-Nelson
University of Minnesota

idea: persistently negative prediction errors (low reward rate) cause splitting of new state



But: this does not explain similar phenomena in LI

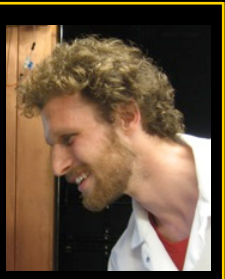


no LI if conditioning is in different context than preexposure

and this too depends on hippocampus… What is the hippocampus doing in these tasks?

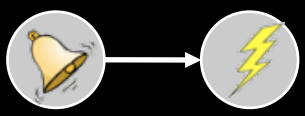
Honey & Good 1993

going one step further: learning causal structure

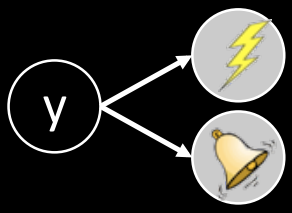


Sam Gershman

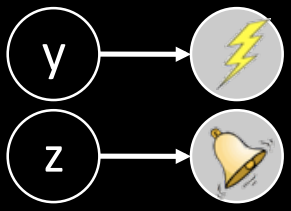
structure I:
tone causes shock



structure II:
latent variable (y)
causes tone and shock



structure III:
tone and shock caused
by independent latent
variables (y,z)



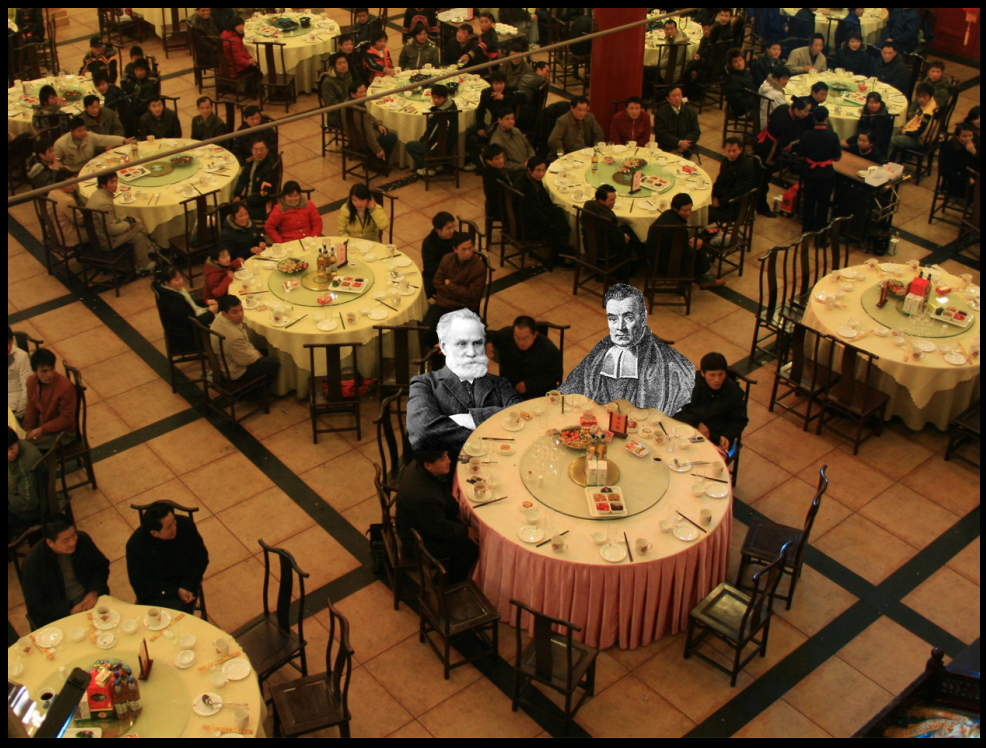
Bayesian inference with an infinite capacity prior

Gershman, Blei & Niv 2010

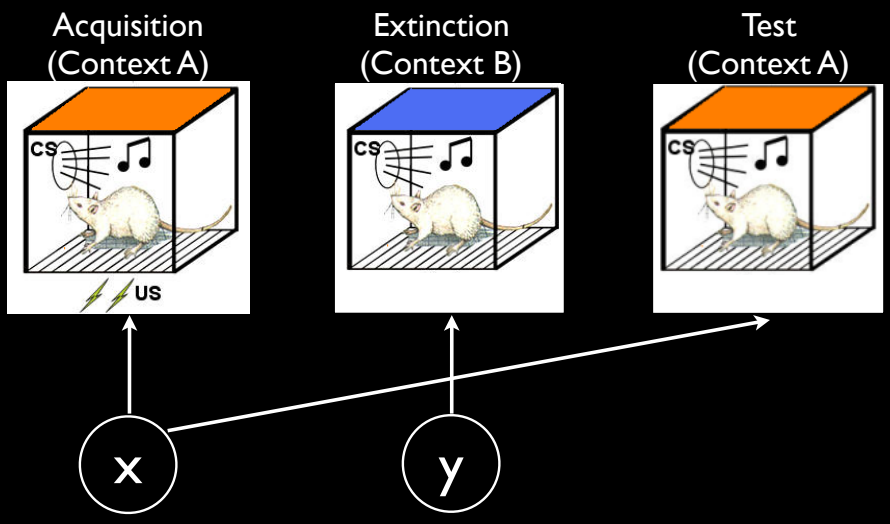
Hypothesis: Animals assume a *generative model* that is flexible enough to capture the complex structure of the environment, but constrained enough to allow learning

1. Each trial is caused by a single latent cause
2. Each latent cause tends to produce similar trials (ie, has a characteristic probability of "emitting" each of the stimuli)
3. All else equal, a (recently) prolific latent cause (ie, has caused many trials) is more likely to cause another trial
4. The number of possible latent causes is unbounded. That is, there is some (small) probability that the current trial is generated by a completely new latent cause

Chinese Restaurant Prior (Infinite capacity mixture model)

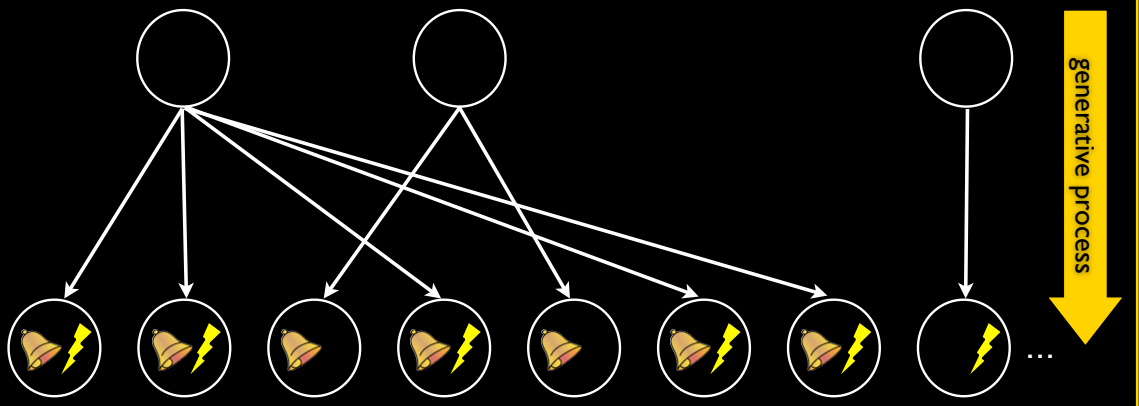


A latent cause theory

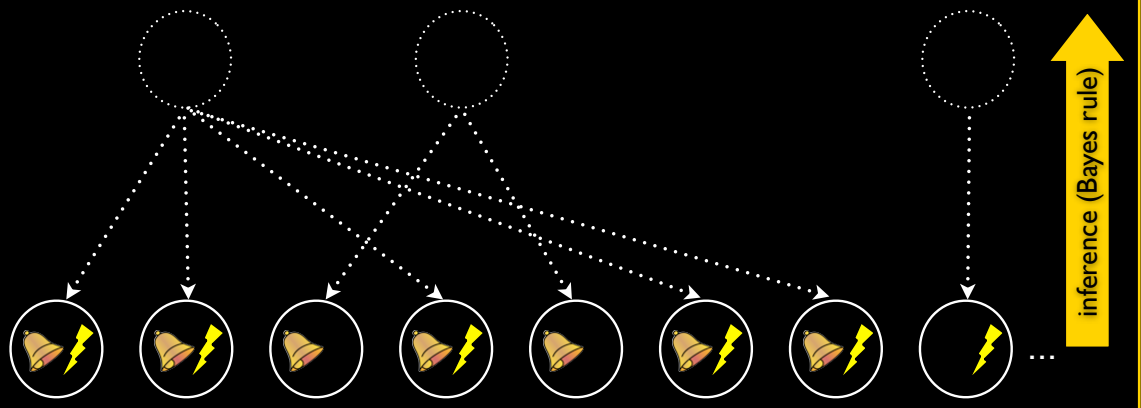


1. latent causes have characteristic emission probabilities
2. similarity between trials allows inference about the relevant latent cause

Inference: "inverting" a generative model



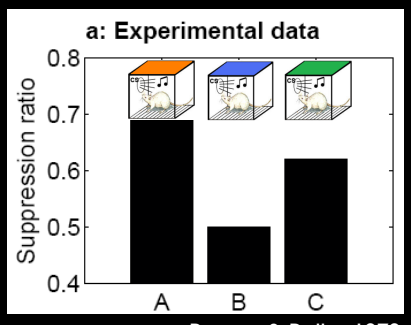
Inference: "inverting" a generative model



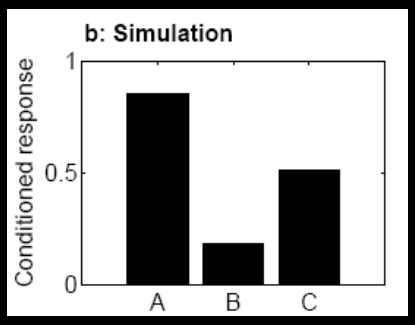
$$p(\text{cause}|\text{data}) \propto p(\text{data}|\text{cause})p(\text{cause})$$

Model explains renewal of fear

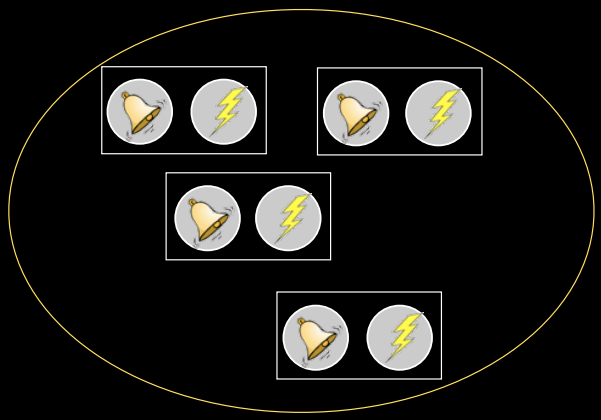
Gershman, Blei & Niv 2010

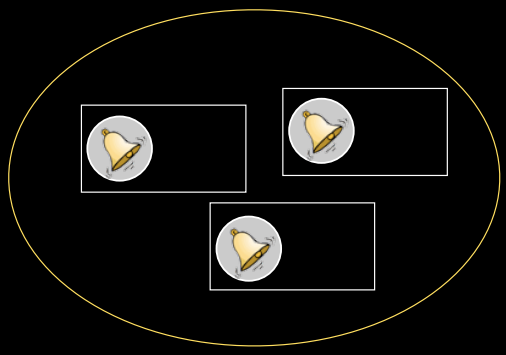


Bouton & Bolles 1979



Conditioning as clustering

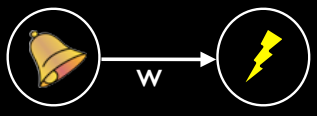




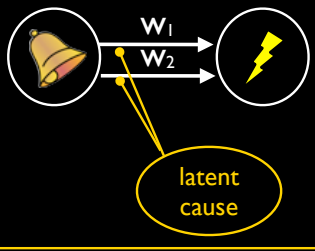
Within each cluster: "learning as usual"
(Rescorla-Wagner, RL etc.)

Equivalent to multiple associations

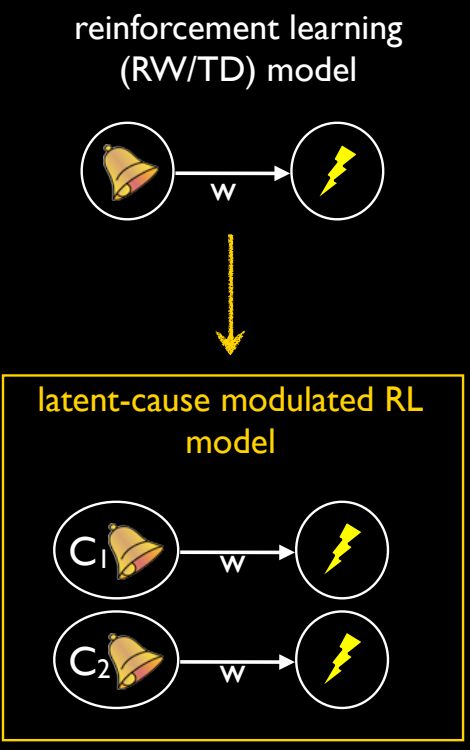
reinforcement learning (RW/TD) model



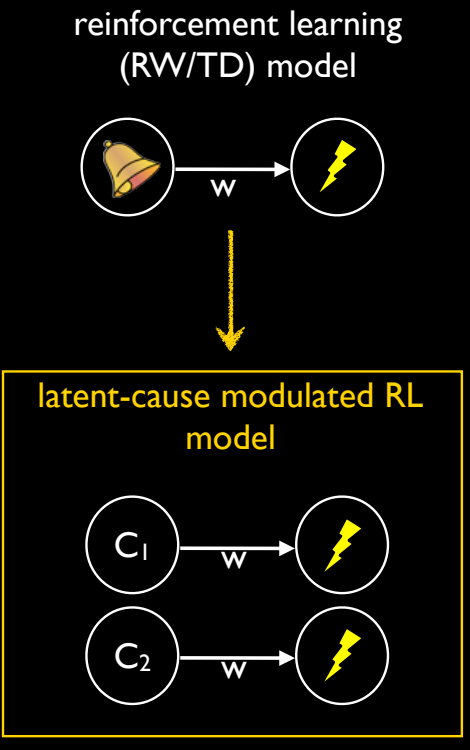
latent-cause modulated RL model



Equivalent to "compound cues"



Latent causes as states

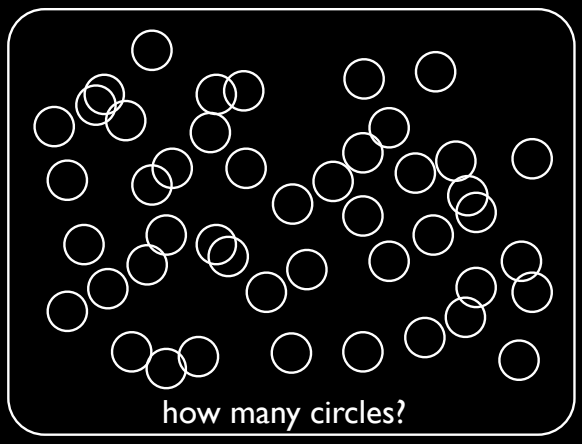


Testing the model I : "Circles Task"

Gershman & Niv 2013

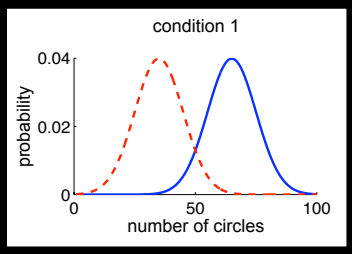
similarity is crucial for clustering observations

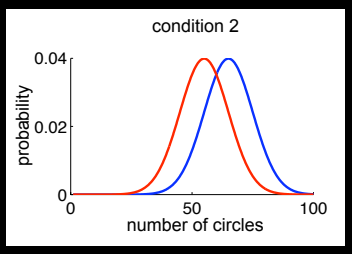
inference about latent causes determines "internal state"

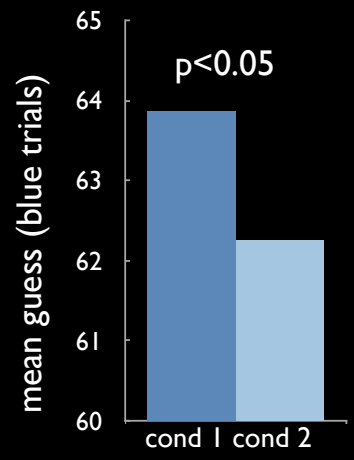


Testing the model I : "Circles Task"

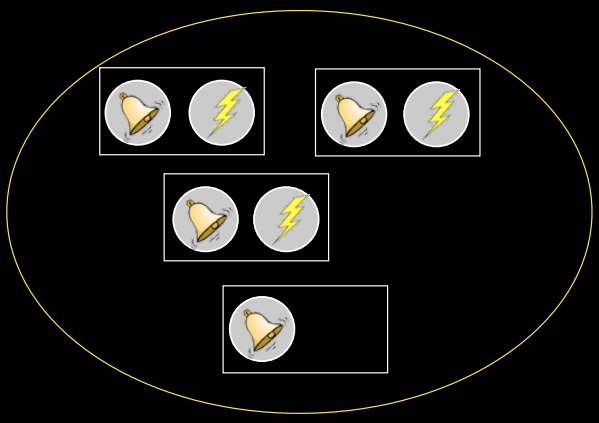
Gershman & Niv 2013





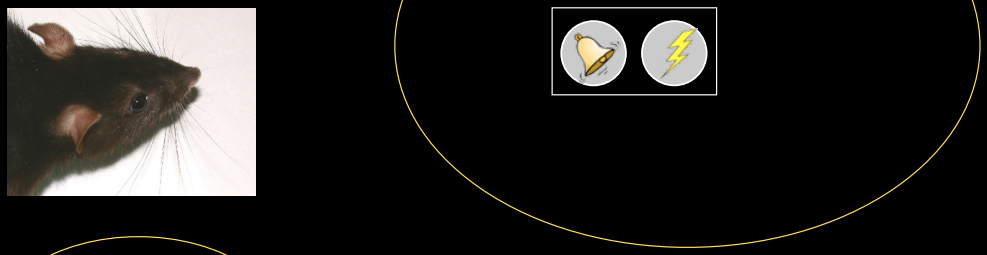


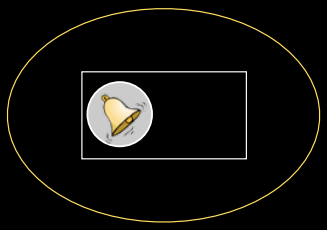
A "battle" between learning and memory



associative learning (modify existing cluster)

A "battle" between learning and memory

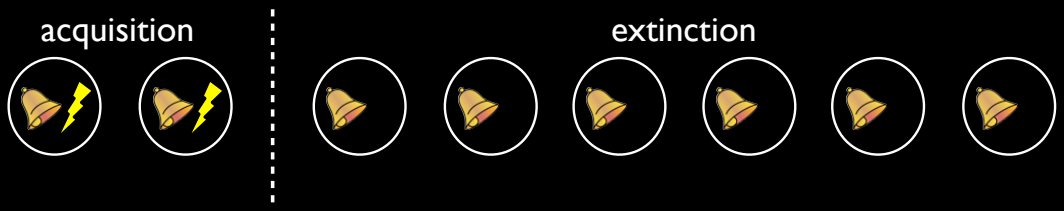




structural learning
(create new state)

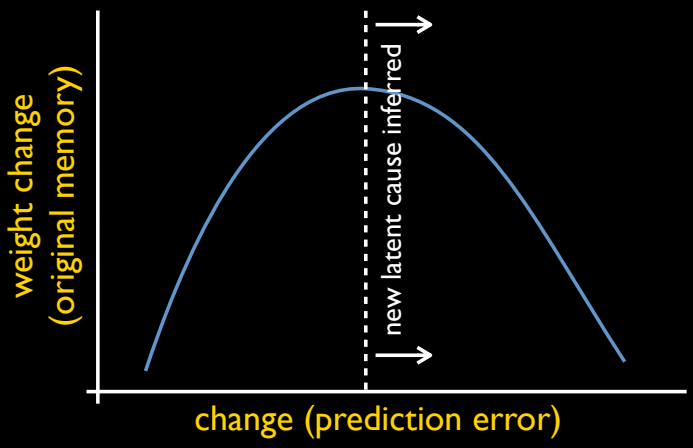
Testing the model II: how to erase a fear memory

hypothesis: prediction errors (dissimilar data) lead to new states



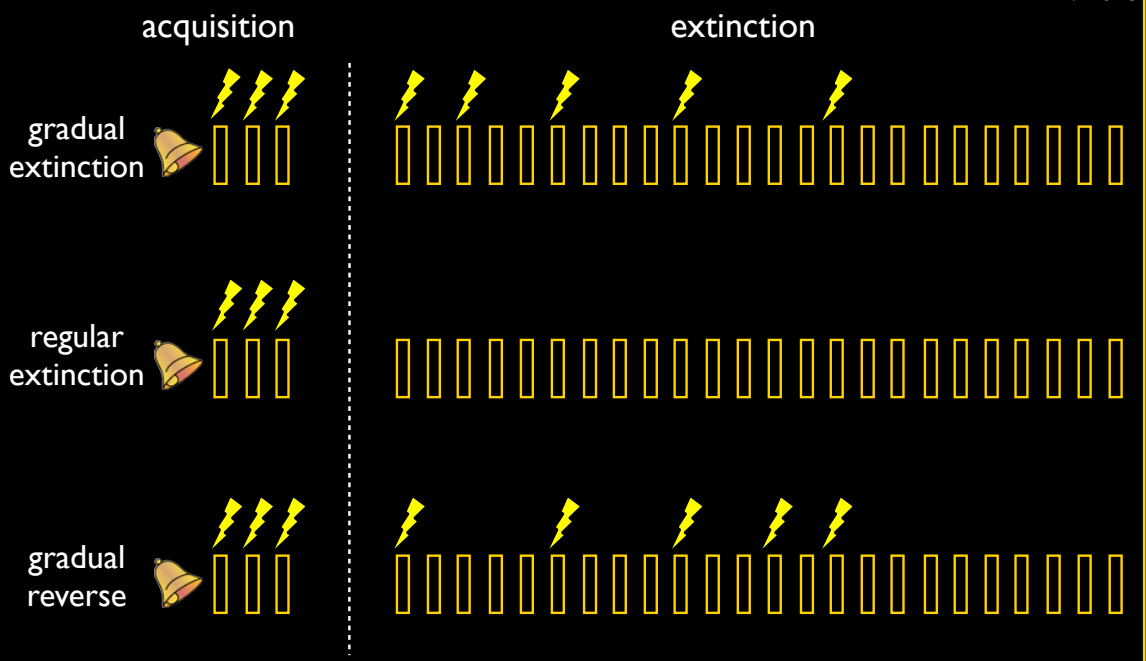
what if we make extinction a bit more similar to acquisition?

Testing the model II: how to erase a fear memory



Testing the model II: gradual extinction

Gershman, Jones, Norman, Monfils & Niv 2013

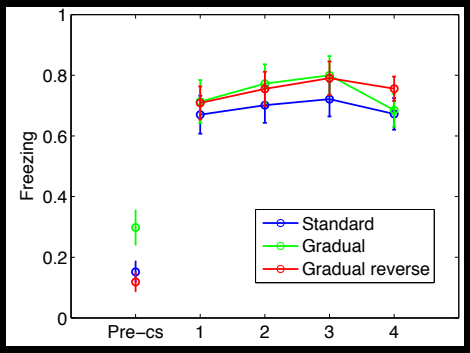


test one day (reinstatement) or 30 days later (spontaneous recovery)

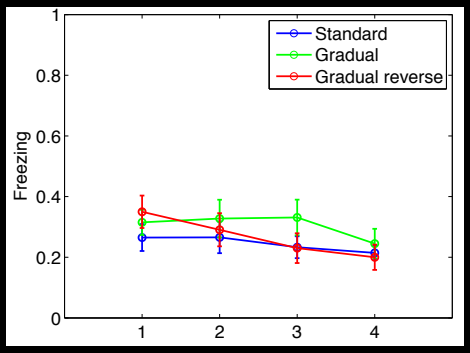
Testing the model II: gradual extinction

Gershman, Jones, Norman, Monfils & Niv 2013

first trials of extinction

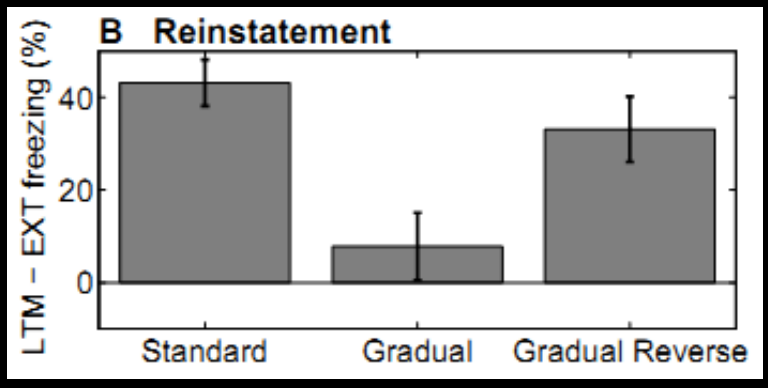


last trials of extinction



Testing the model II: gradual extinction

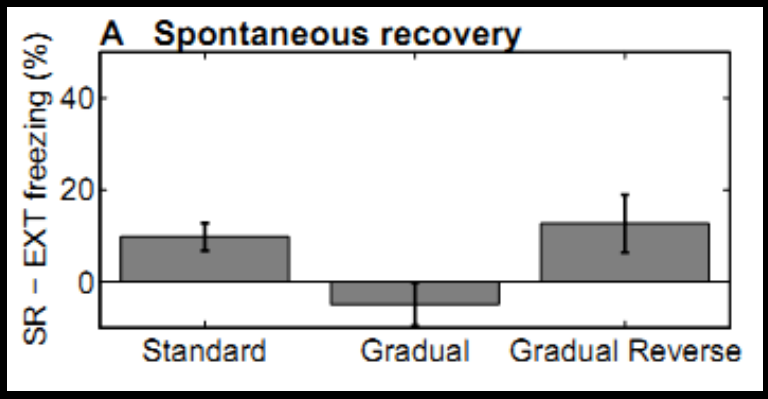
Gershman, Jones, Norman, Monfils & Niv 2013



only gradual extinction group shows no reinstatement

Testing the model II: gradual extinction

Gershman, Jones, Norman, Monfils & Niv 2013

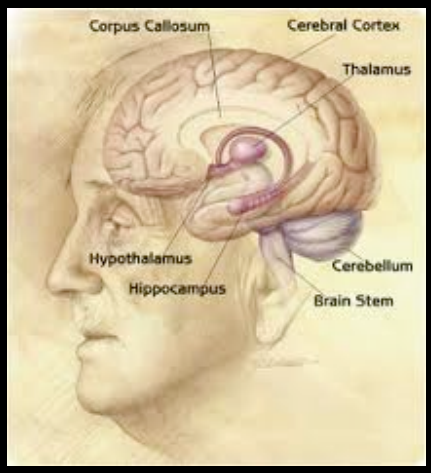


only gradual extinction group freeze LESS after 30 days than at the end of extinction (= no spontaneous recovery)

Where does this "clustering" occur?

key decision: is current observation (trial) similar or different from previous observations?

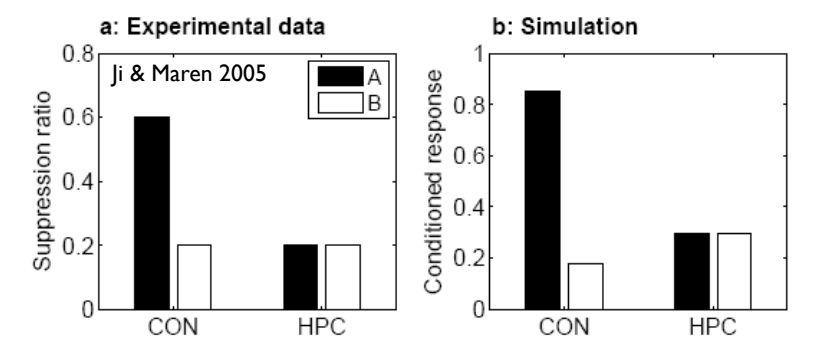
an expert in such decisions: HIPPOCAMPUS



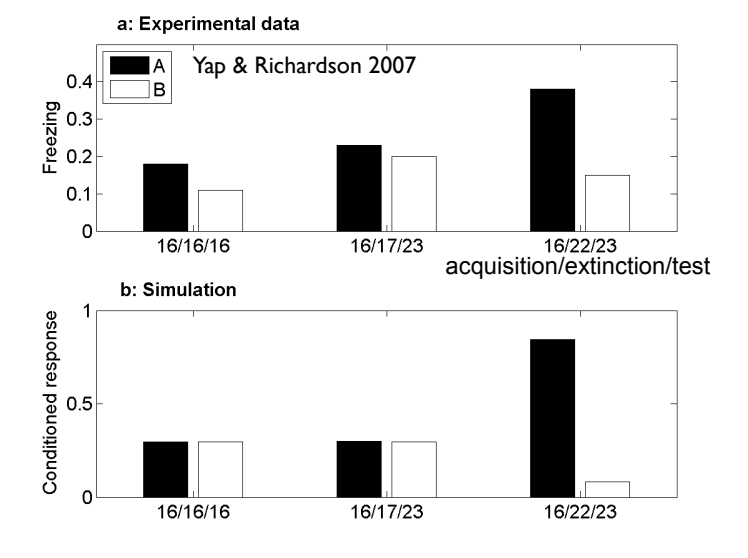
Where does this "clustering" occur?

key decision: is current observation (trial) similar or different from previous observations?

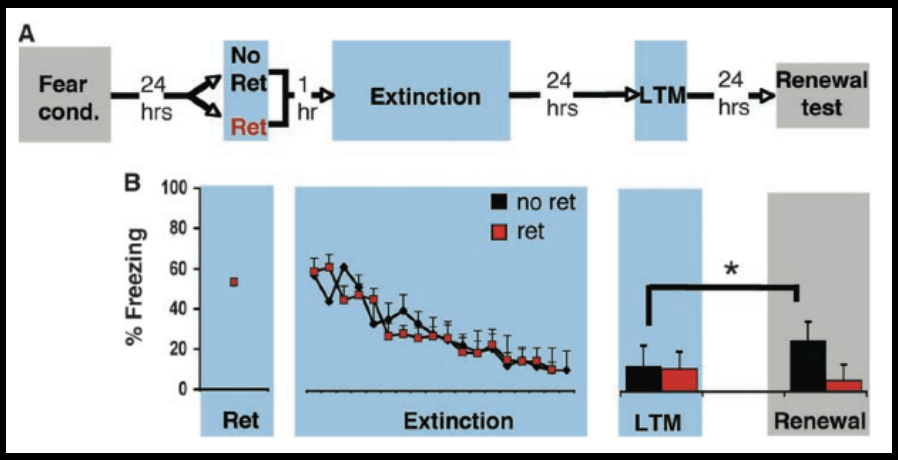
an expert in such decisions: HIPPOCAMPUS



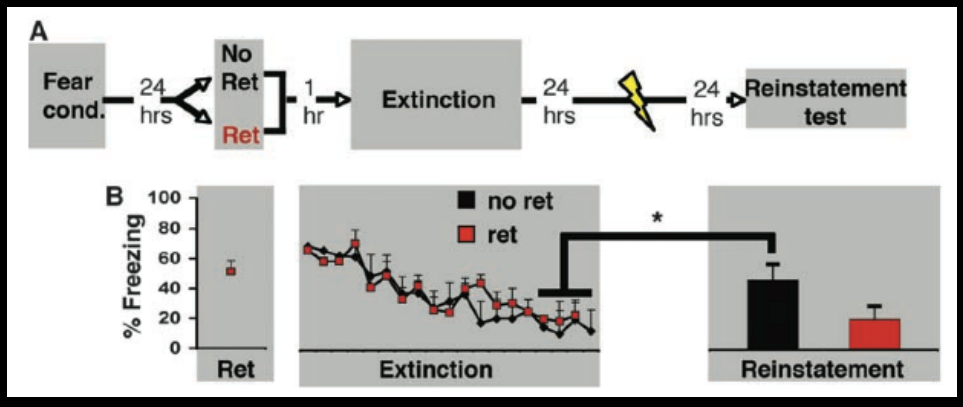
Where does this "clustering" occur?



But now the *really* crazy stuff: Monfils-Schiller paradigm

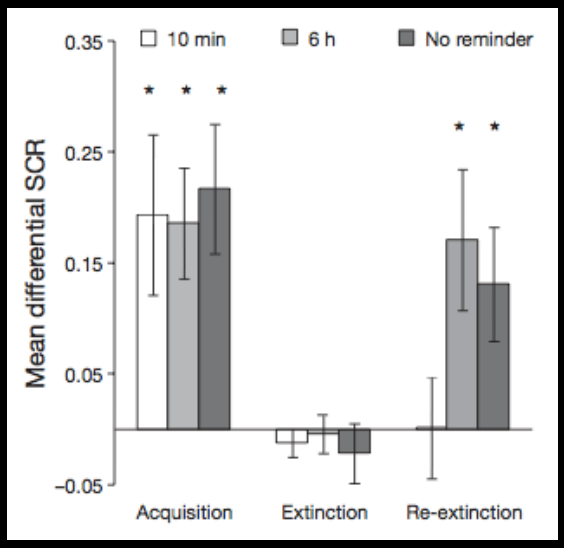


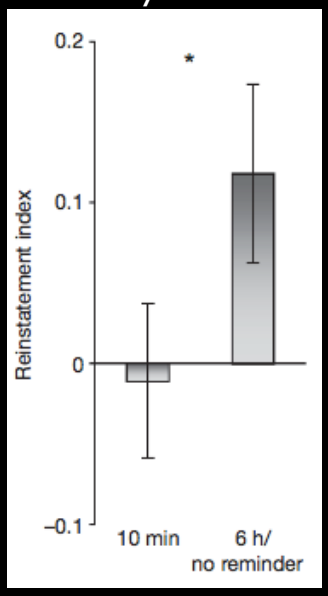
But now the *really* crazy stuff: Monfils-Schiller paradigm



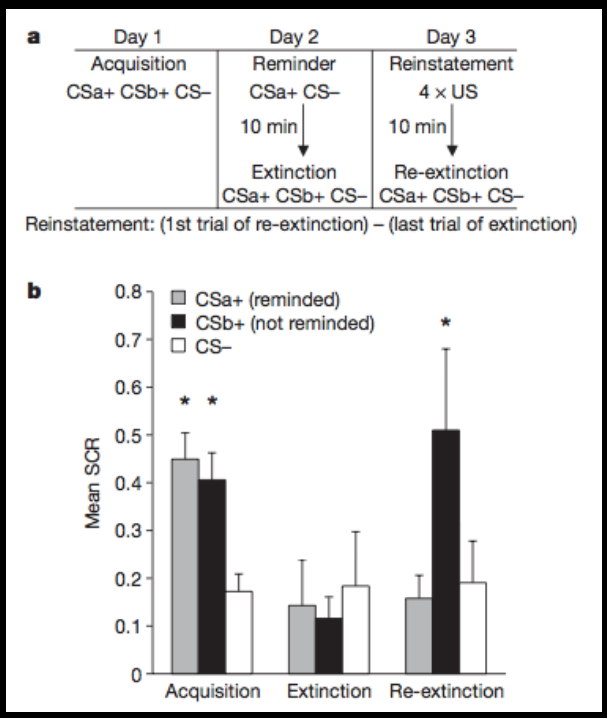
In case you are skeptic: this also works in humans...

Even a year later!

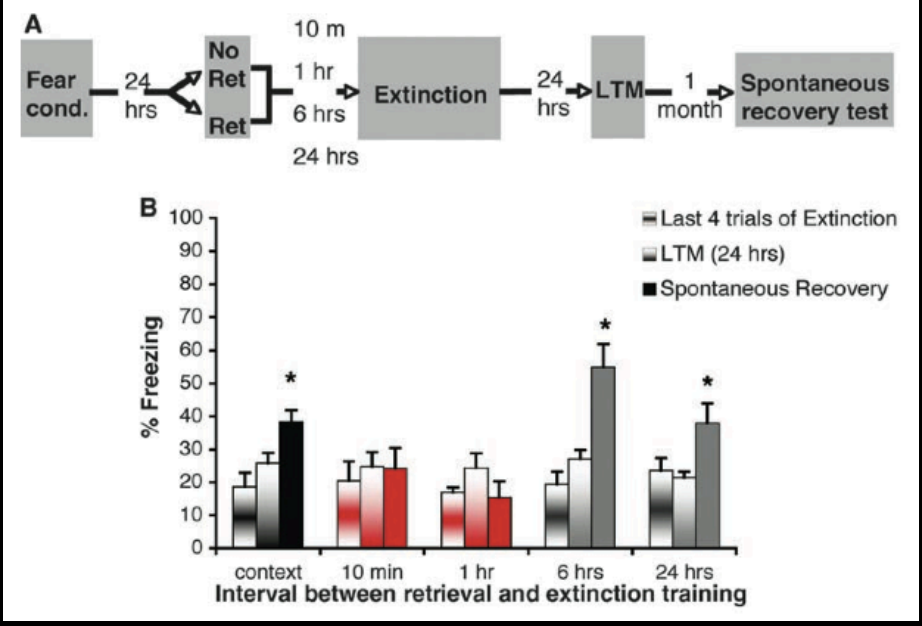




In case you are skeptic: this also works in humans...

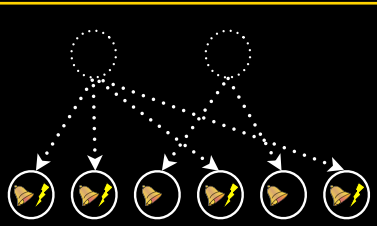


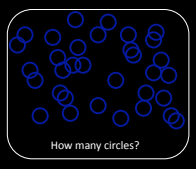
An explanation: reconsolidation

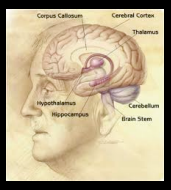


Summary

- Animals might be smarter than we think: make inferences about the latent causes of observations as part of prediction learning
- similarity determines when different observations will interfere with each other: create a new memory or alter an old one? Inferred structure determines the "battle" between learning and memory in the brain
- we can control memory modification by titrating prediction errors
- memory is not just a passive record: depends on the animal's beliefs about the underlying causal structure (subjective!)
- potential interaction between dopamine and hippocampus in reinforcement learning







Bayesian inference: what is all the hype about?

- random variables (discrete, continuous)
- probability distribution
- probability distributions as beliefs
- Bayes rule

Bayesian inference: intuition

$$P(A|B) = P(B|A)P(A)/P(B)$$
$$P(model|data) = P(data|model)p(model)/p(data)$$

- sometimes P(A|B) is really hard but P(B|A) is easy
  → this is why inversion using Bayes rule helps
- eg. you saw 3 heads and 7 tails, what is the probability that the coin is fair?
- If the coin is fair, what is the probability of seeing 3 heads and 7 tails?
- eg. Brian says: "I'll pass" - what is Brian talking about?
- If Brian is talking about an exam, how probable is this sentence?

The importance of priors in Bayesian inference

- interpretation of Bayes rule: we care about both prior and likelihood in inference
- eg. test for disease came out positive
  accuracy of test is 99%
  disease a-priori in 1/10000 people
  what are the odds that the patient has the disease?