Biological Learning

Peter Dayan

Gatsby Computational Neuroscience Unit

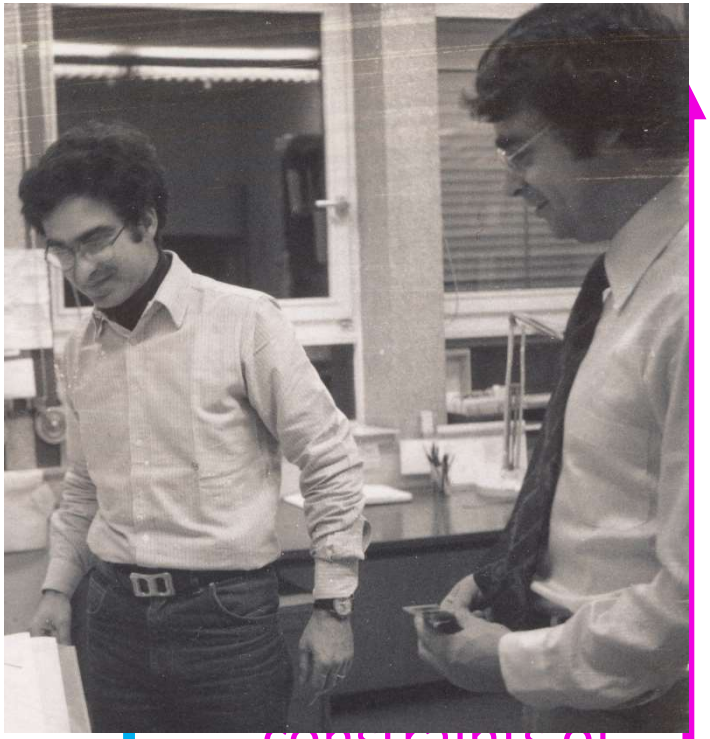
Nathaniel Daw Sam Gershman Sham Kakade Yael Niv

Plan

- biological learning & Marr
- conditioning
  - classical/Pavlovian & prediction
  - instrumental/operant & action selection
  - temporal difference learning & dopamine
  - Pavlovian misbehaviour
- Bayesian conditioning
  - Kalman filtering
  - Chinese restaurant extinction

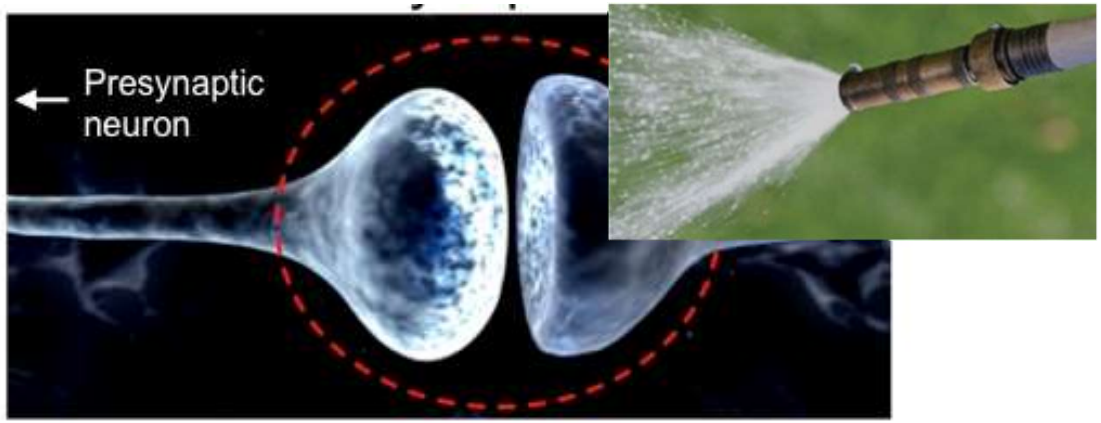
Marrian Cognition

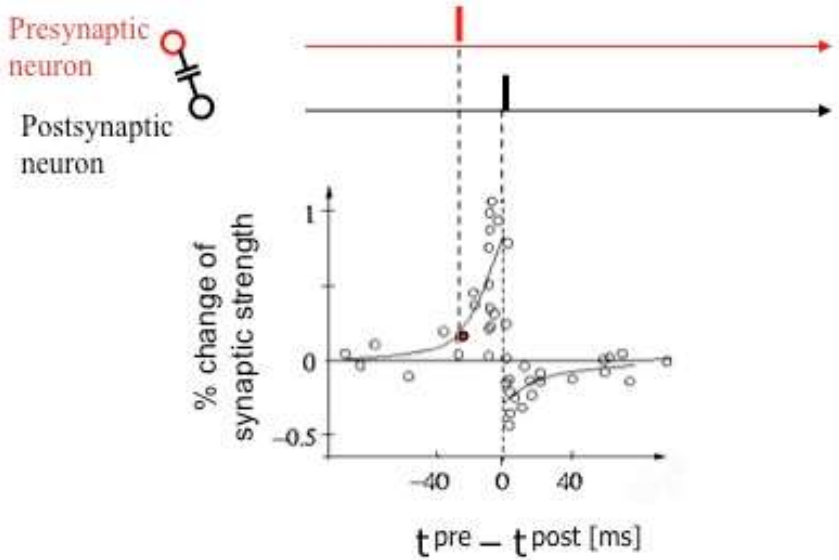
- computation/ethology
  - goal
  - logic of the strategy
- algorithm/psychology
  - effective procedure
  - representations
- implementation/neuroscience
  - neural realization



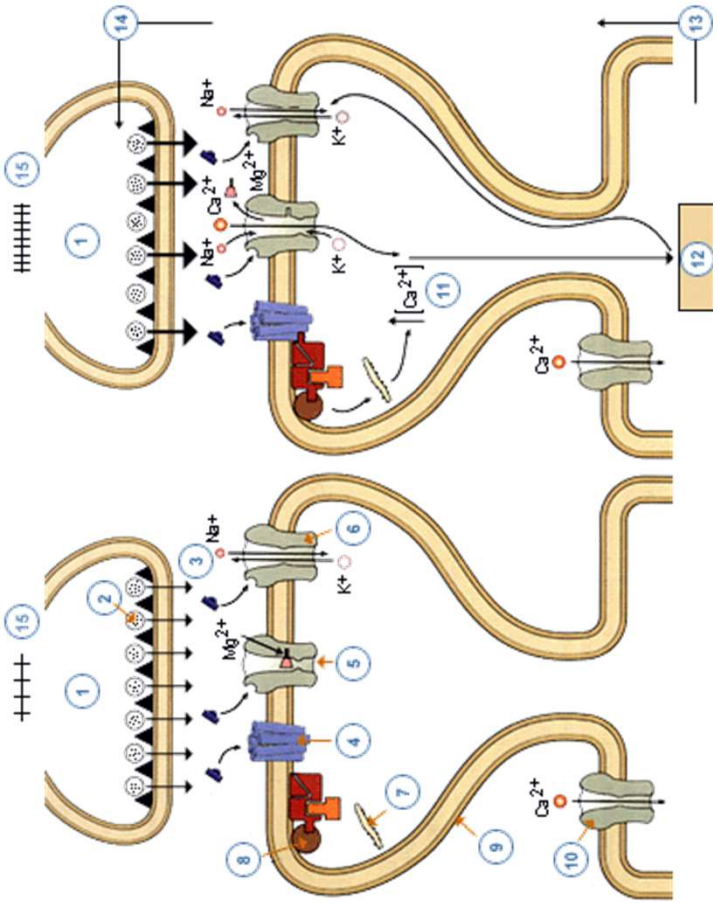
constraints of
the substrate;
heuristics;
approximations

Neuroscience of Learning

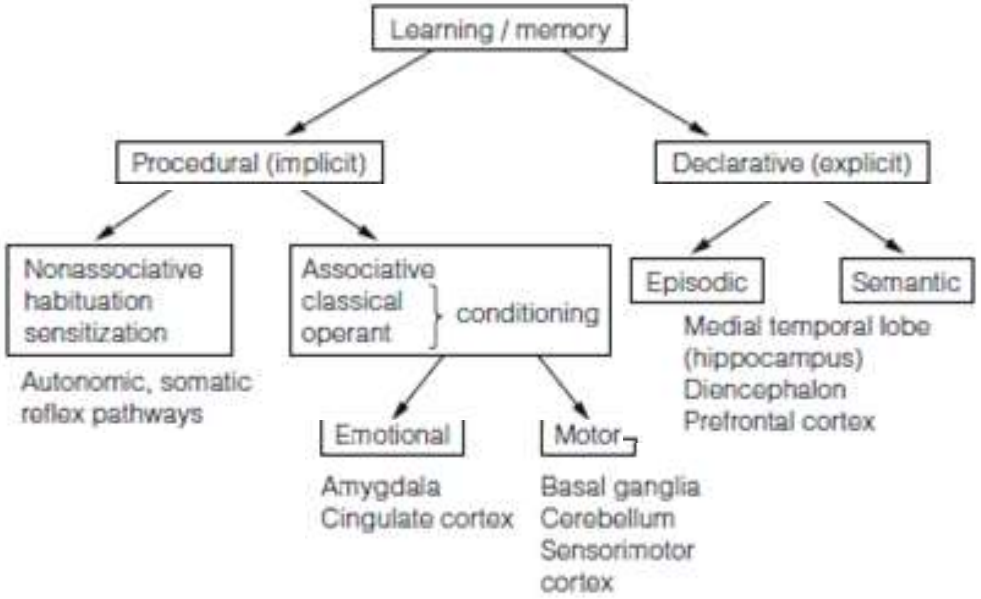


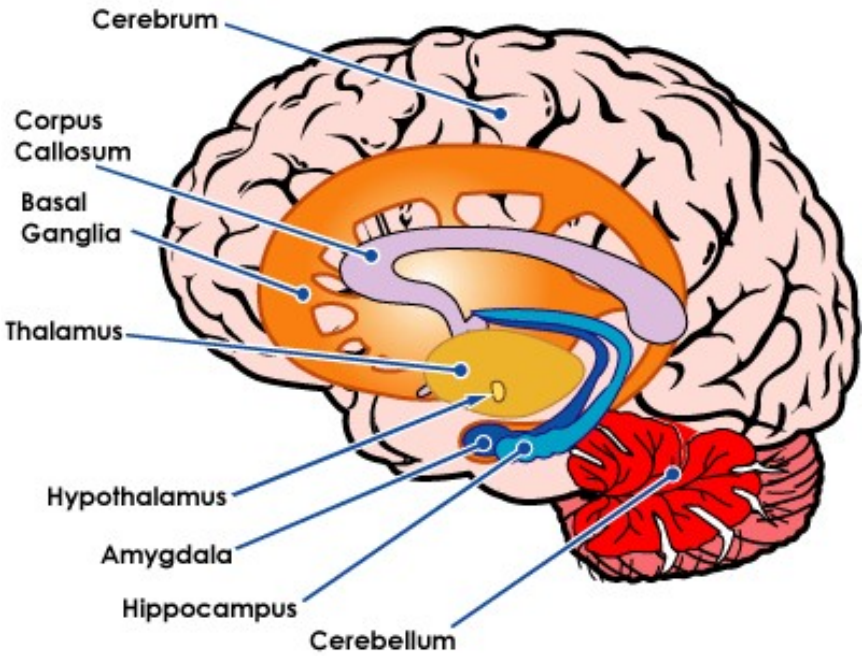


dopamine;
acetylcholine



Conventional Psychobiology of Learning



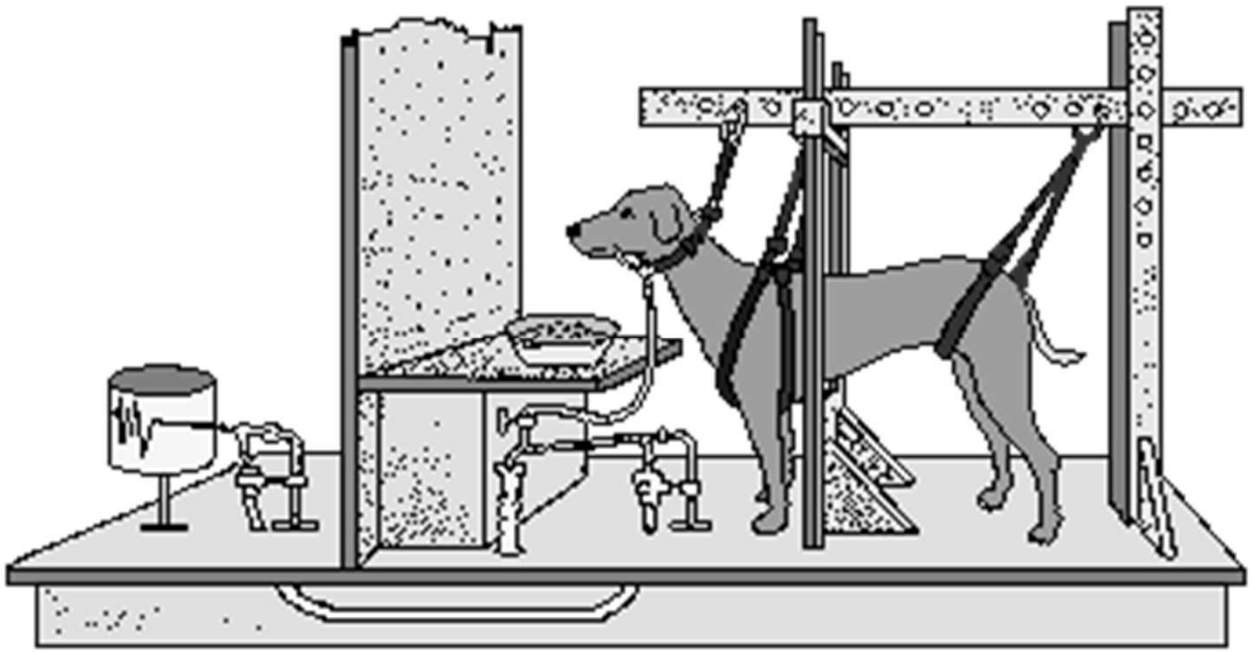


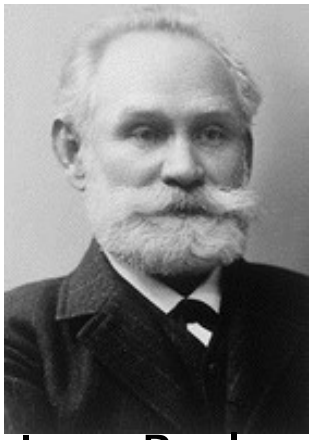
- ubiquitous learning of predictions: misses model-based/declarative control
- forward/inverse models
  - motor control
  - MDPs + policies
  - graphics/vision
- representational learning (Hebb; PCA; InfoMAX)

Bakery LeCun

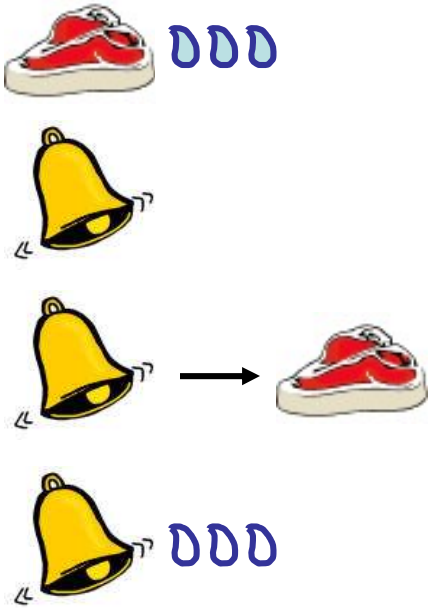
- self-supervised learning
  - e.g., learning representations
- supervised learning
  - e.g., episodic memories
- reinforcement learning
  - e.g., policies

Animals learn predictions





Ivan Pavlov



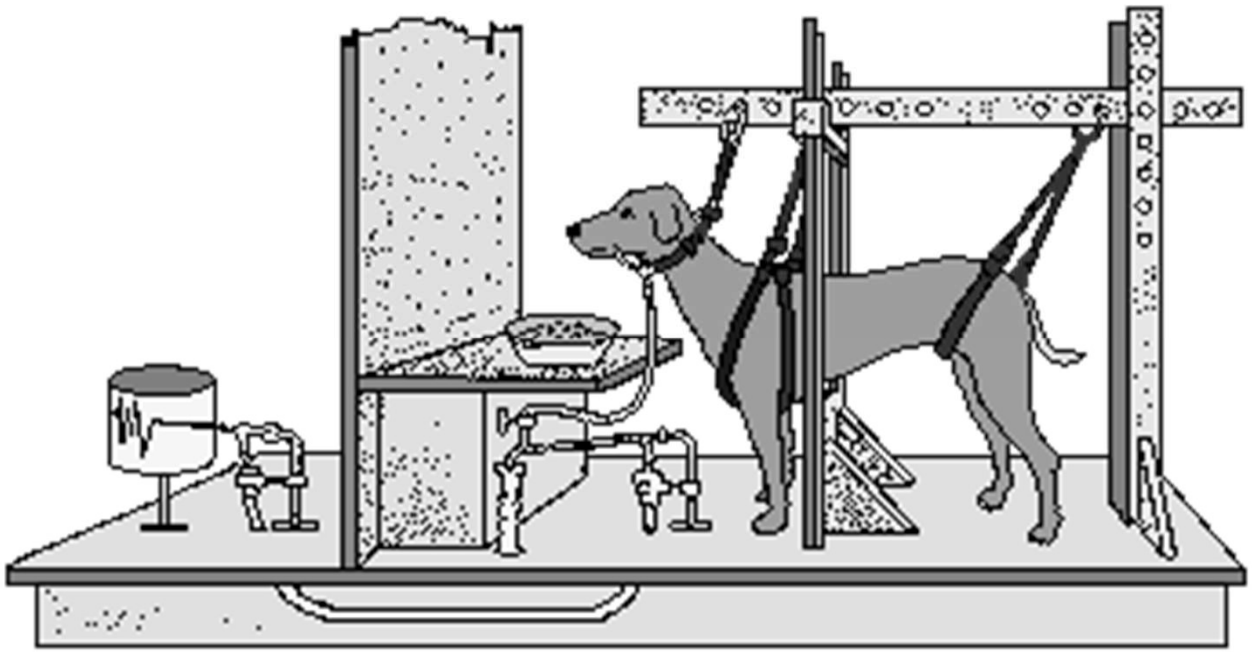
 = Unconditioned Stimulus r

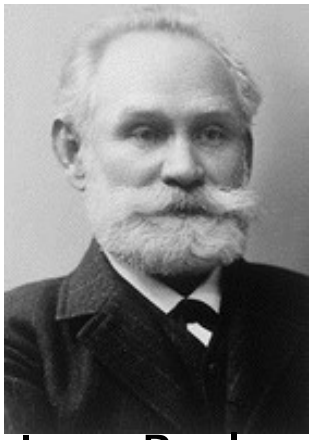
 = Conditioned Stimulus $s_j \in \{0,1\}$

 = Unconditioned Response (reflex);
Conditioned Response (reflex)

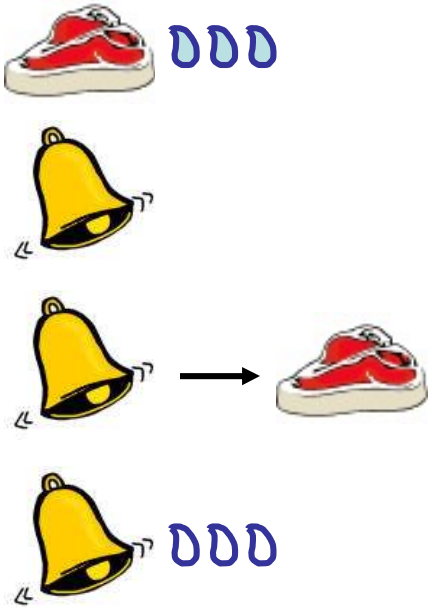
(thanks to Yael Niv)

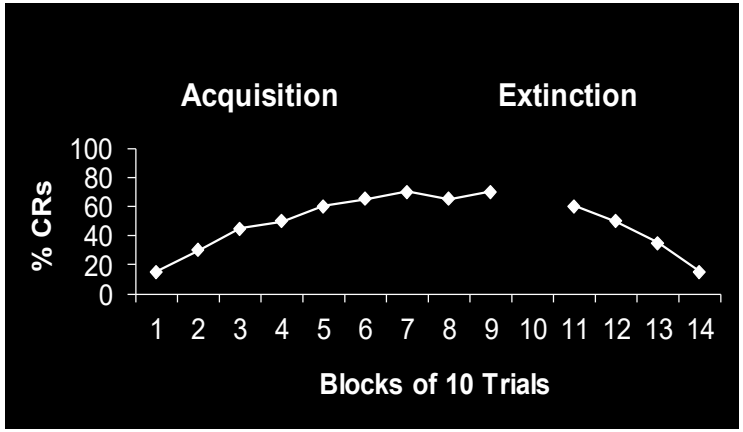
Animals learn predictions





Ivan Pavlov





Spatiotemporal Contiguity?

- stage 1:
  - ❖A (light) → r
- stage 2:
  - ❖A + B (tone) → r
- test:
  - ❖A → r
  - ❖B → ∅
- contingency isn't enough: need surprise

Formalize

- prediction is sum based on binary stimuli:

$$V(\mathbf{s}) = \sum_j w_j s_j \qquad s_j = \{0,1\}$$

- average prediction error:

$$E(\mathbf{w}) = \langle (r - V(\mathbf{s}))^2 \rangle$$

- learning rule as gradient descent:

$$\Delta w_j = \alpha(r - V(\mathbf{s}))s_j$$

$$= \alpha \quad \delta \quad s_j$$

stage 1:

- A (light) → r

stage 2:

- A + B (tone) → r

test:

- A → r
- B → ∅

- highly explanatory:
  - explains: gradual acquisition & extinction, blocking, overshadowing, conditioned inhibition, and more..

Rescorla-Wagner learning

ignore CS: $V_{n+1} = V_n + \alpha(r_n - V_n)$

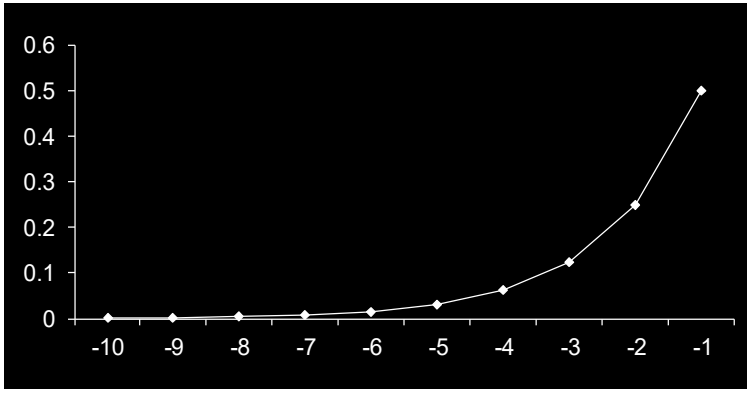
$$V_{n+1} = \alpha r_n + (1-\alpha)V_n$$

$$= \alpha r_n + (1-\alpha)(\alpha r_{n-1} + (1-\alpha)V_{n-1})$$

$$= \alpha r_n + (1-\alpha)\alpha r_{n-1} + (1-\alpha)^2 (\alpha r_{n-2} + (1-\alpha)V_{n-2})$$

the prediction on trial $n+1$ is influenced by rewards on trials (n), $(n-1)$

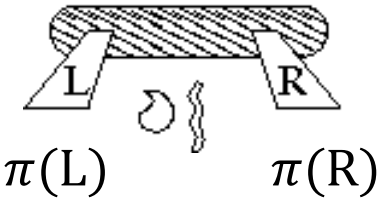
$$V_{n+1} = \alpha \sum_{v=1}^{n} (1-\alpha)^{n-v} r_v + (1-\alpha)^n V_1$$



recent rewards weigh more heavily

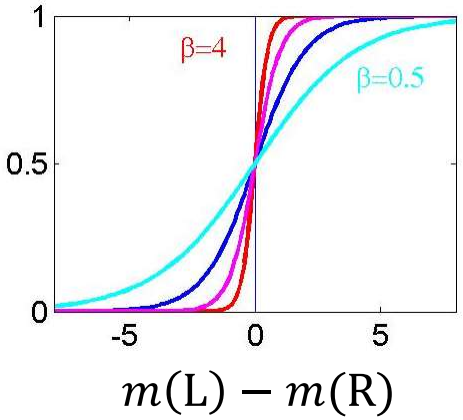
learning rate = forgetting rate!

What about Choice?



- based on action propensities: $m(\mathrm{L}); m(\mathrm{R})$
- stochastic policy:

$$\pi(\mathrm{L}; \mathbf{m}) = \frac{\exp(\beta m(\mathrm{L}))}{\exp\big(\beta m(\mathrm{L})\big) + \exp(\beta m(\mathrm{R}))} = \sigma(\beta(m(\mathrm{L}) - m(\mathrm{R}))$$



use RW rule: Indirect actor: $m(a) = Q(a)$

$$Q_{n+1}(\mathrm{L}) = Q_n(\mathrm{L}) + \alpha\delta_n \quad \text{with } \delta_n = r_n(\mathrm{L}) - Q_n(\mathrm{L})$$

Or Direct Actor

- expected reward:

$$E(\mathbf{m}) = \pi(\mathrm{L};\mathbf{m})\langle r(\mathrm{L})\rangle + \pi(\mathrm{R};\mathbf{m})\ \langle r(\mathrm{R})\rangle$$

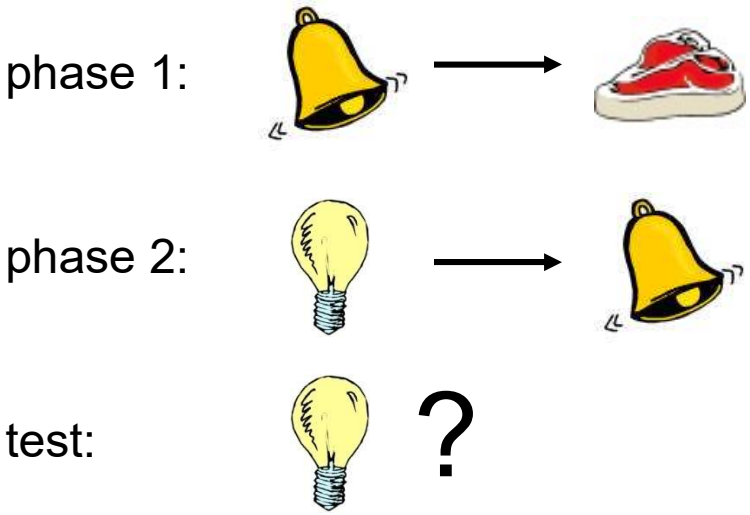
- change parameters according to:
  - relative quality of reward: $(r(a) - V)$
  - sensitivity of choice to parameter: $\frac{\partial \ln \pi(a;\mathbf{m})}{\partial m}$

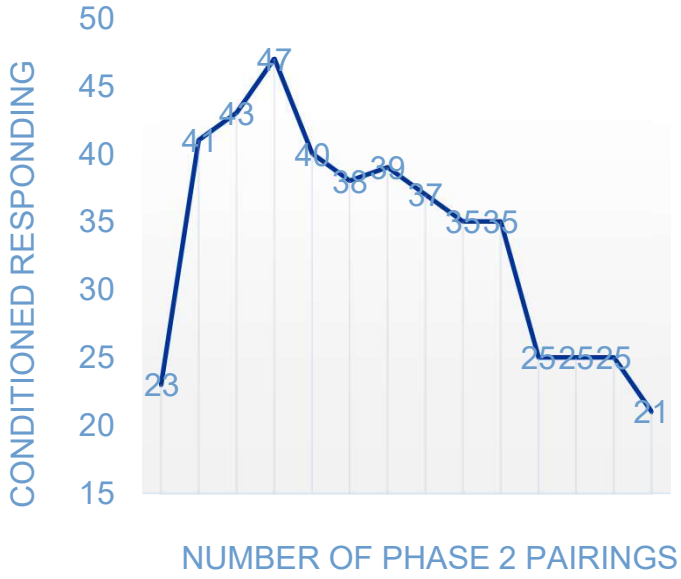
$$\Delta m \propto (r(a) - V)\frac{\partial \ln \pi(a;\mathbf{m})}{\partial m}$$

- does stochastic gradient ascent in $E(\mathbf{m})$

$$\Delta m \propto \frac{\partial}{\partial m} E(\mathbf{m})$$

second order conditioning





cf. Rescorla-Wagner learning

animals learn that a predictor of a predictor, is also a predictor
so not just interested in predicting immediate reward…

start over from the top

- the problem: optimal prediction of future reward

$$V_t = E\left[\sum_{\tau=t}^{T} r_\tau\right]$$

want to predict expected sum of future reward in a trial/episode

N.B. here t, τ index time within a trial

- the equivalent RW prediction error: $\delta^{\mathrm{RW}} = r - \mathrm{V}$

$$\delta_t = \sum_{\tau=t}^{T} r_\tau - V_t$$

- has a key problem of apparently requiring waiting

start over from the top

- The problem: optimal prediction of future reward

$$V_t = E\left[\sum_{\tau=t}^{T} r_\tau\right]$$

want to predict expected sum of future reward in a trial/episode

$$\begin{aligned} V_t &= E[r_t + r_{t+1} + \cdots + r_T] \\ &= E[r_t] + E[r_{t+1} + \cdots + r_T] \\ &= E[r_t] + V_{t+1} \end{aligned}$$

Bellman eqn for policy evaluation

start over from the top

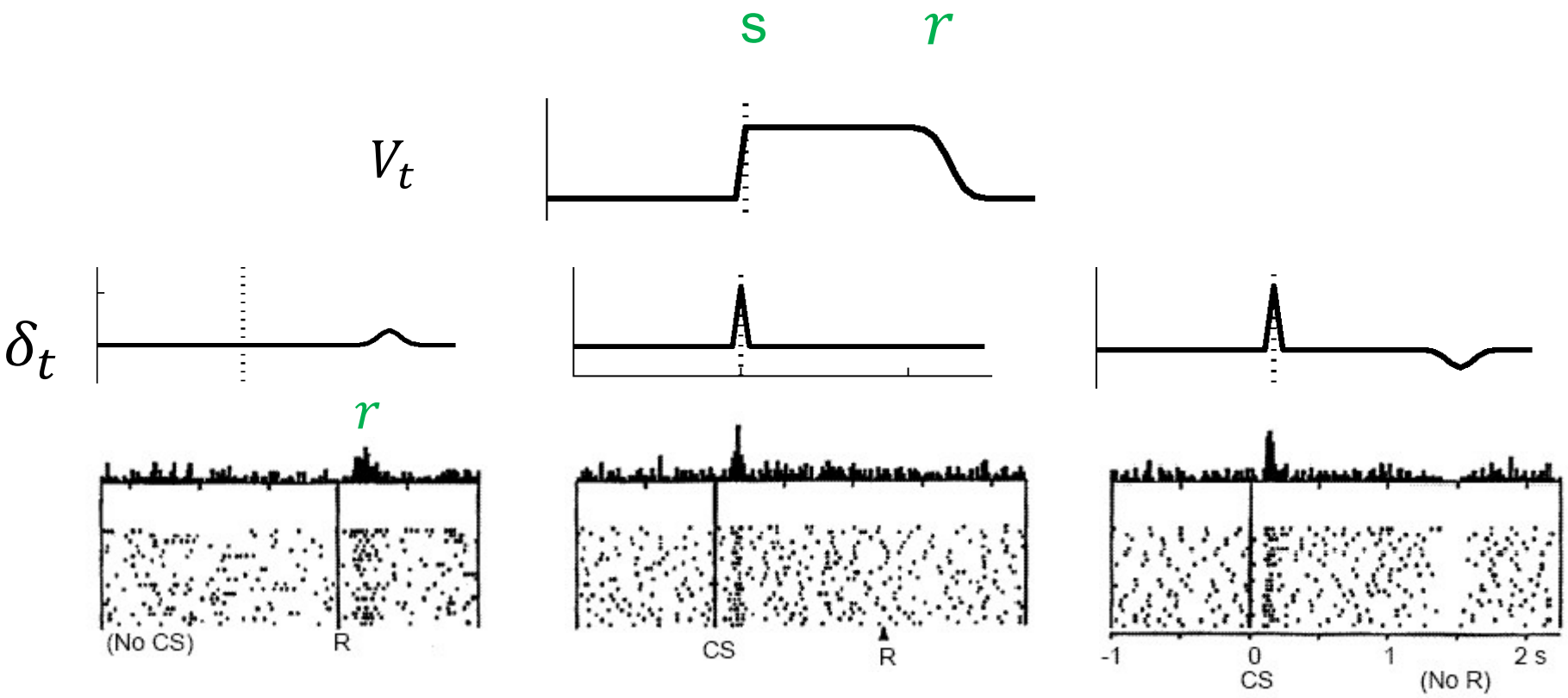
- the problem: optimal prediction of future reward
- the algorithm: temporal difference learning

temporal difference prediction error δ_t

compare to: $V_T \leftarrow V_T + \varepsilon(r_T - V_T)$

dopamine and prediction error

TD error $\delta_t = r_t + V_{t+1} - V_t$

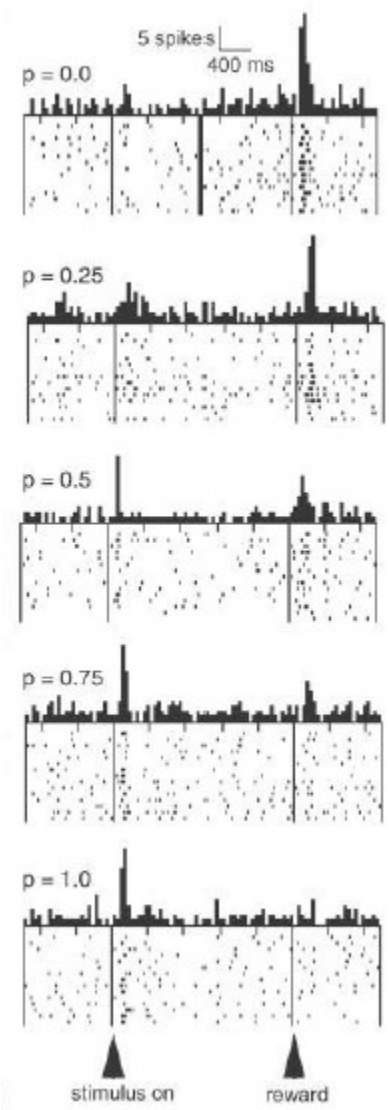


no prediction

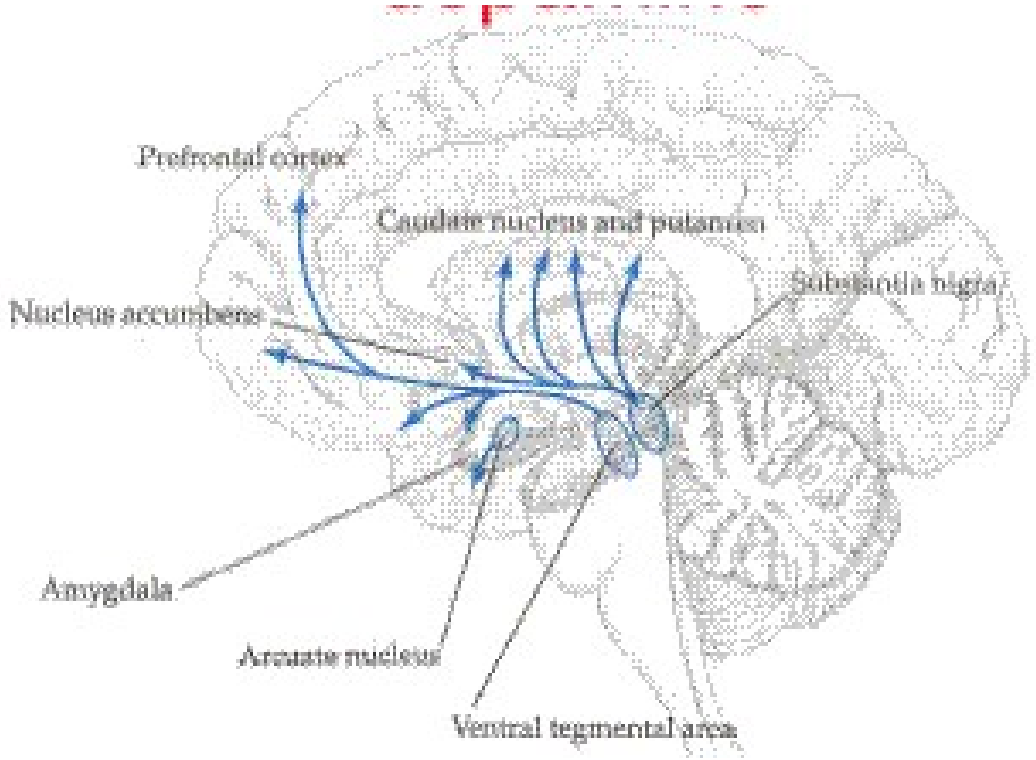
prediction, reward

prediction, no reward

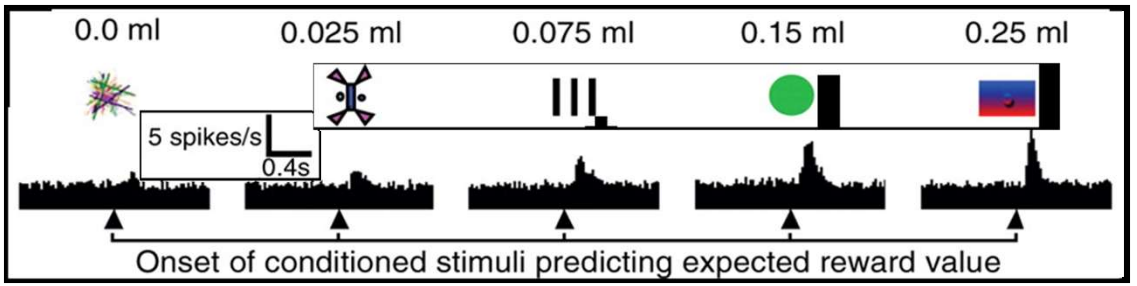
prediction error hypothesis of dopamine



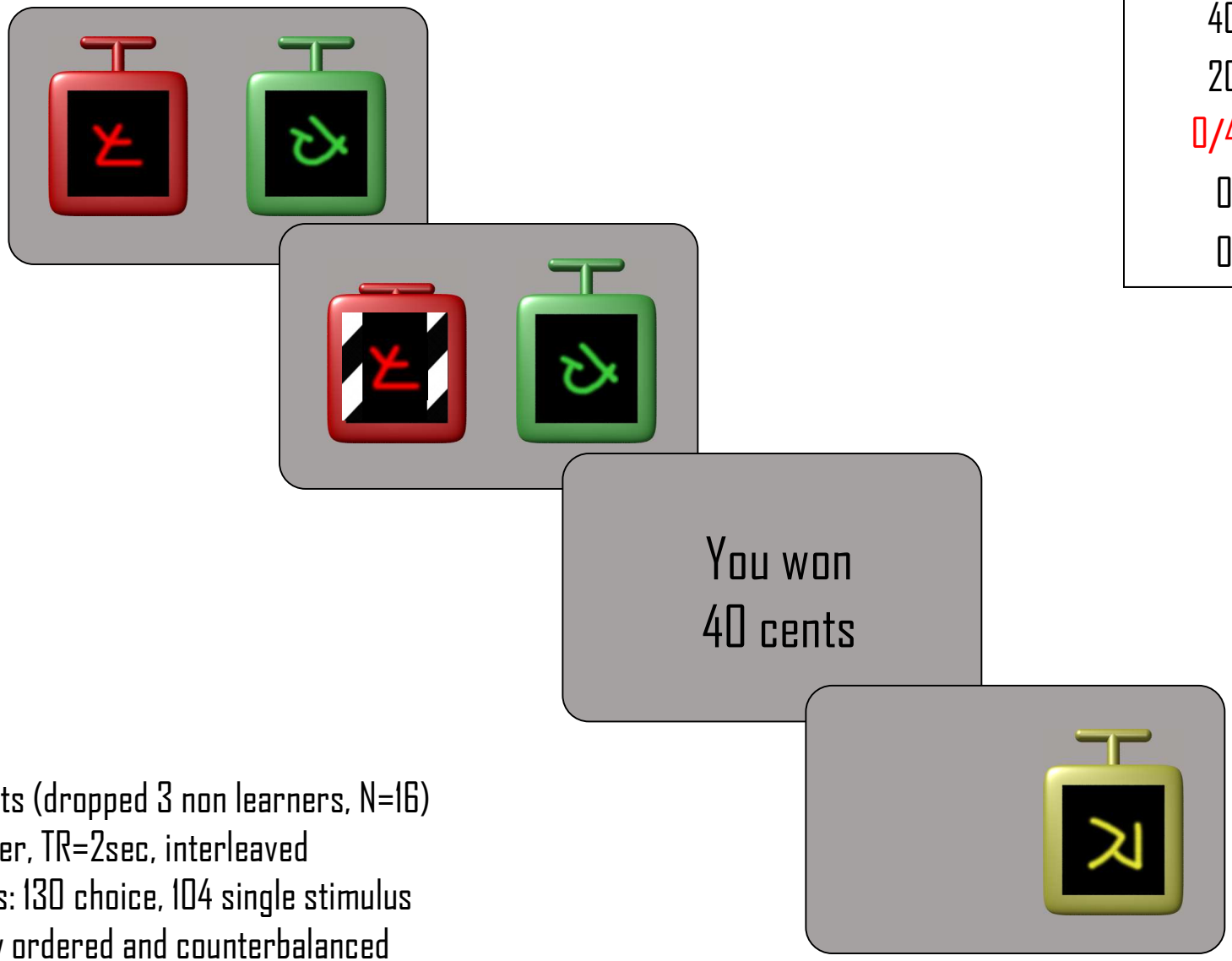
Fiorillo et al, 2003



Tobler et al, 2005



Risk Experiment



5 stimuli:

40¢

20¢

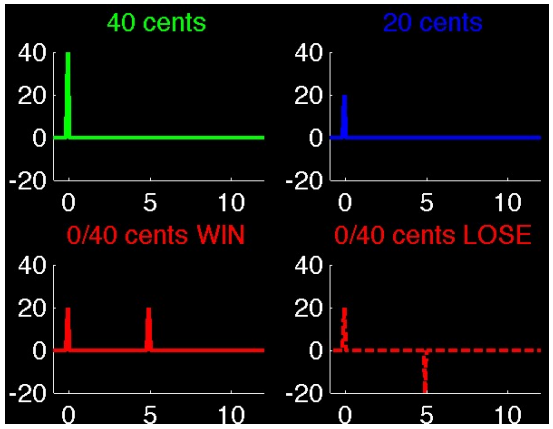
0/40¢

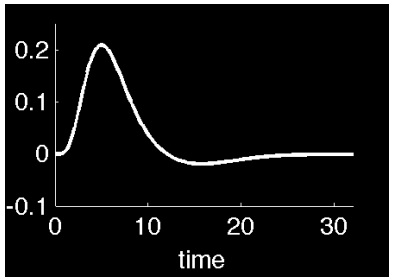
0¢

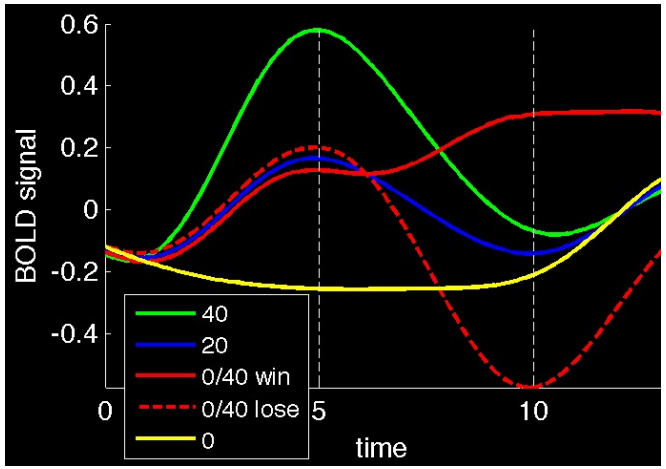
0¢

19 subjects (dropped 3 non learners, N=16)
3T scanner, TR=2sec, interleaved
234 trials: 130 choice, 104 single stimulus
randomly ordered and counterbalanced

Neural results: Prediction Errors

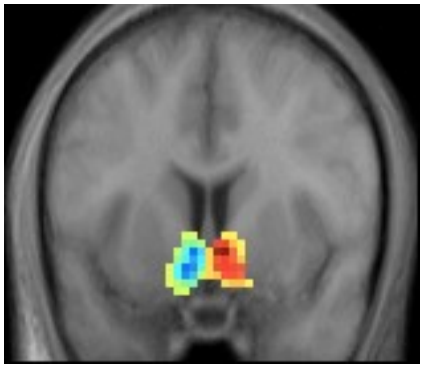




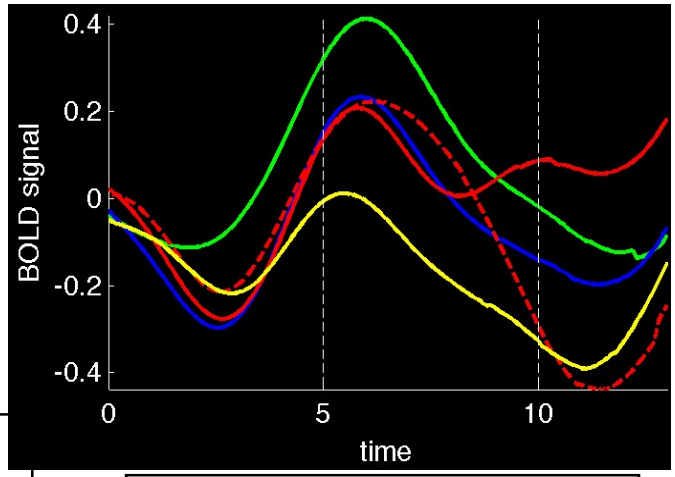


what would a prediction error look like (in BOLD)?

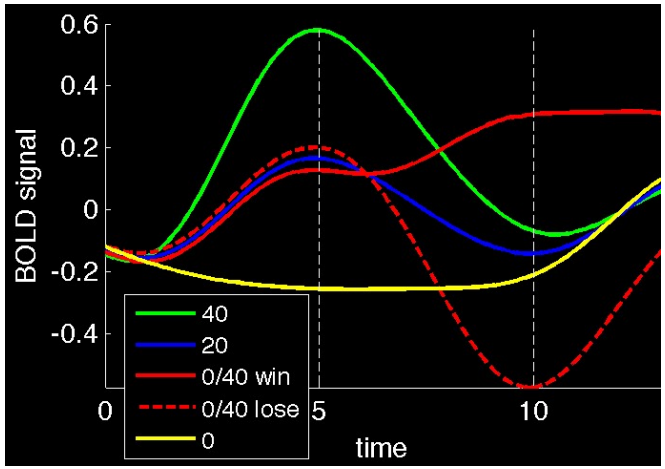
Neural results: Prediction errors in NAC



unbiased anatomical ROI
in nucleus accumbens
(marked per subject*)



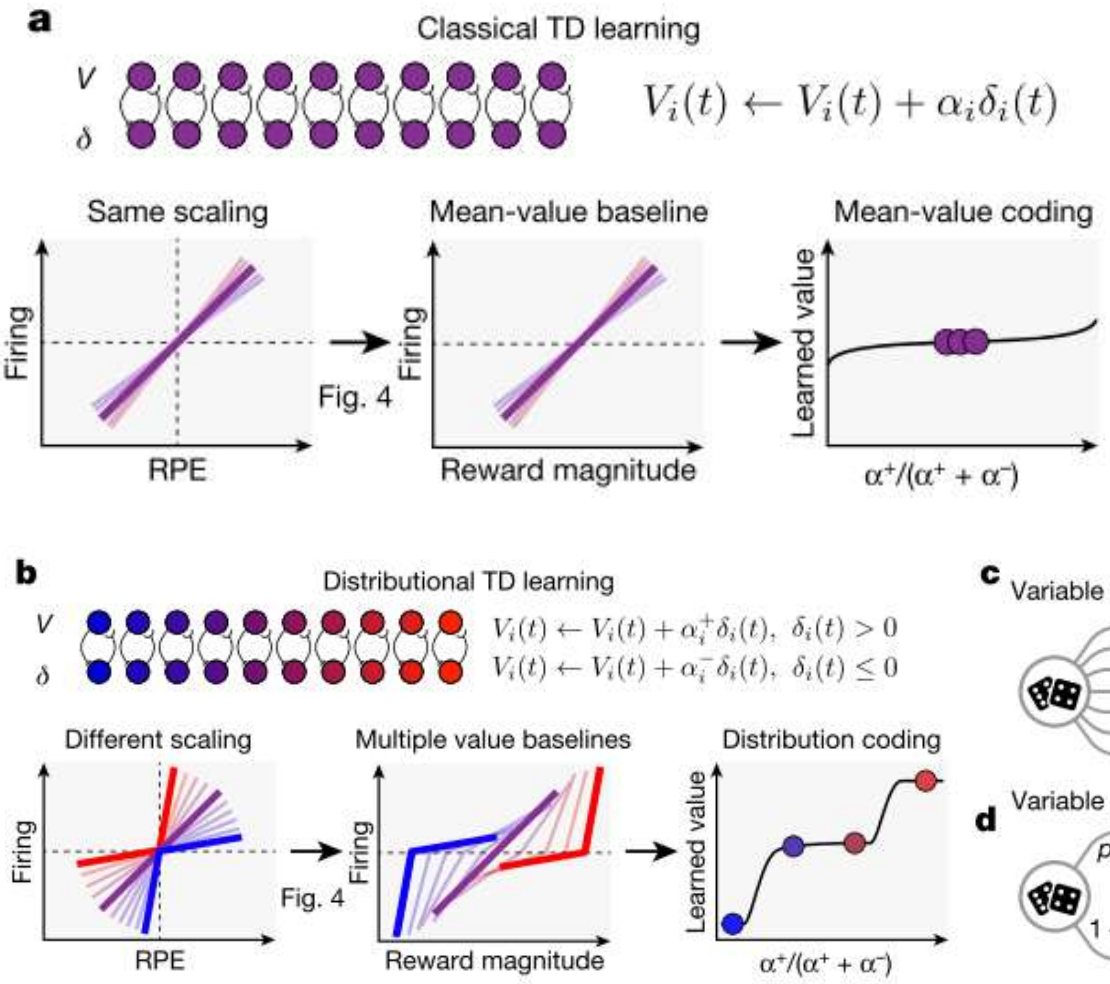
raw BOLD
(avg over all subjects)

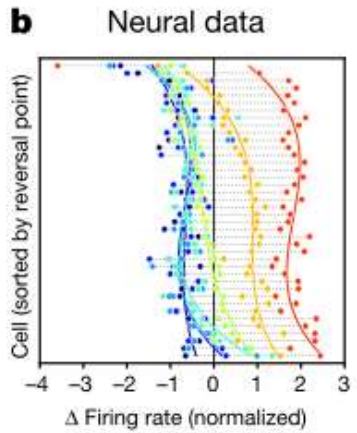


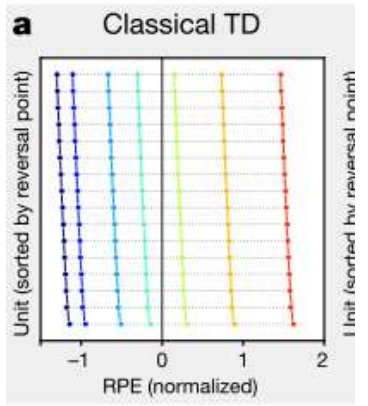
can actually decide between different neuroeconomic models of risk

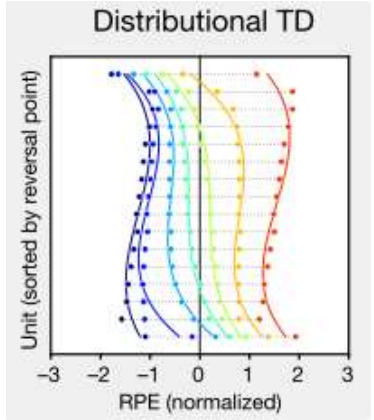
* thanks to Laura deSouza

Distributional TD





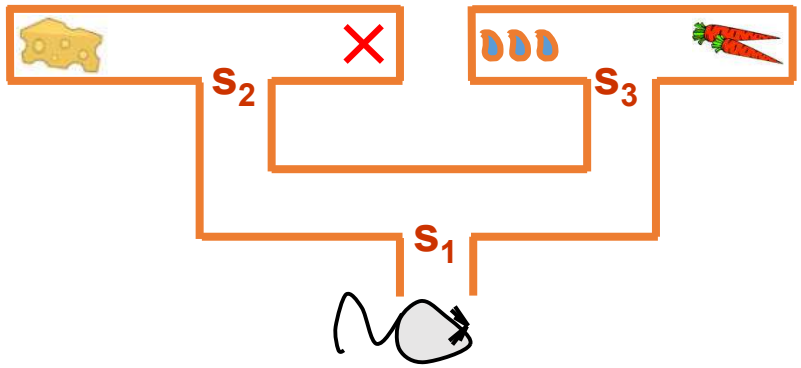


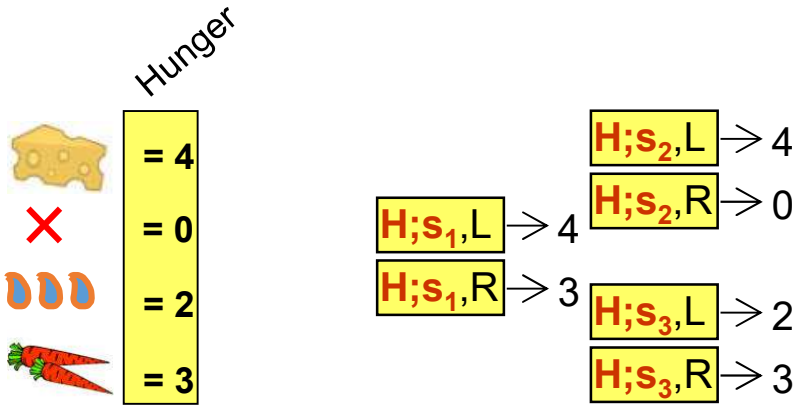


Dabney, Kurth-Nelson et al., 2019

Using prediction for control

Daw & Niv





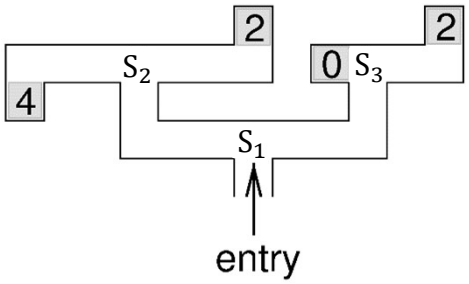
use prediction as surrogate reward: $\Delta m \propto (r(a) - V)\frac{\partial \ln \pi(a;\mathbf{m})}{\partial m}$

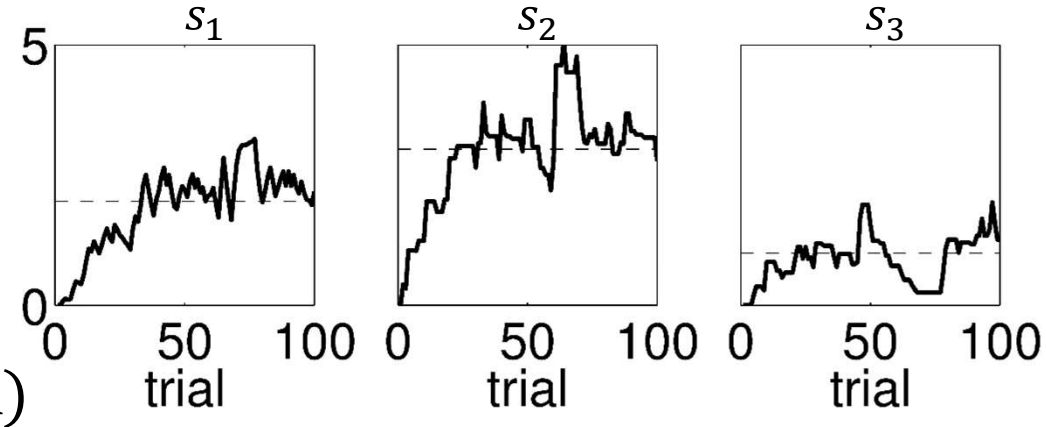
not: $r(a) - V$ instead: $r_t(a_t) + V_{t+1} - V_t = \delta_t$

Direct: Actor-Critic

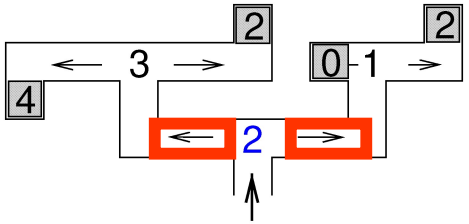
start with policy: $\pi(s, \mathrm{L}; \mathbf{m}) = \sigma(\beta(m(s, \mathrm{L}) - m(s, \mathrm{R}))$

evaluate it: $V(s_1), V(s_2), V(s_3)$





improve it: $\Delta m(s, \mathrm{L}), \Delta m(s, \mathrm{R})$



$$\delta_t = r_t + V_{t+1} - V_t$$

$$\delta_t = 0 + 3 - 2 = +1$$

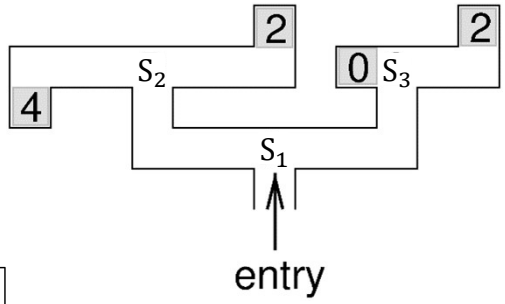
$$\delta_t = 0 + 1 - 2 = -1$$

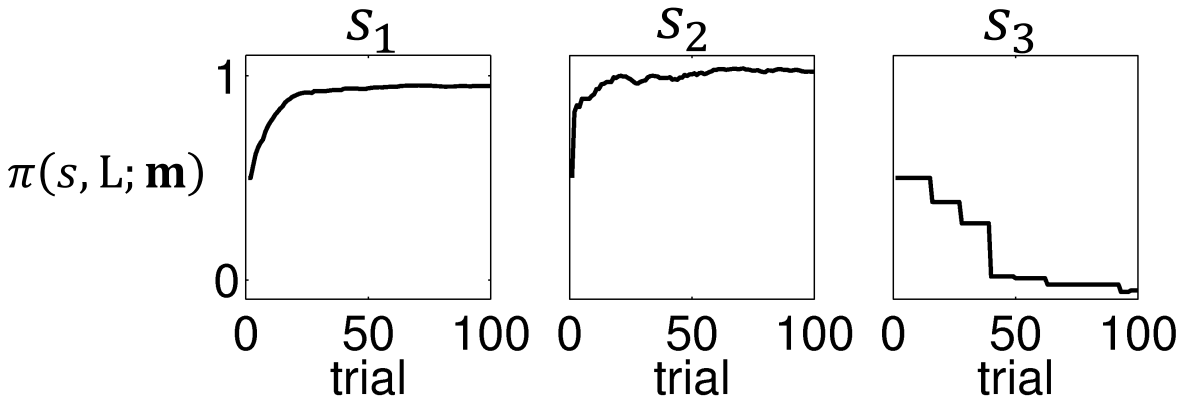
thus choose L more frequently than R $\qquad \Delta m(s, a) \; \alpha \; \delta$

Direct: Actor-Critic

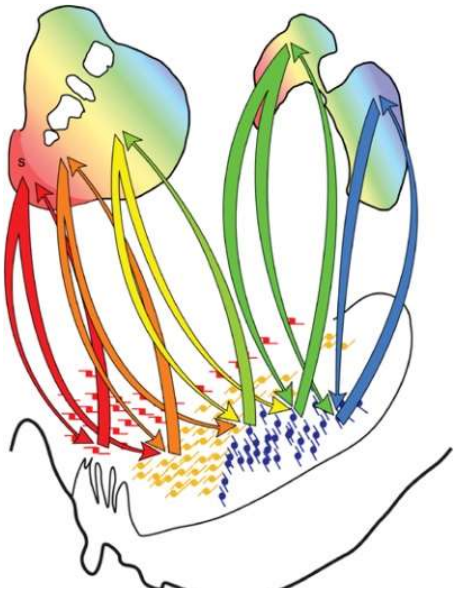
$\delta > 0$ if

- value is too pessimistic $\Rightarrow \Delta V$
- action is better than average $\Rightarrow \Delta m \Rightarrow \Delta \pi$





- spiraling links between striatum and dopamine system: ventral to dorsal

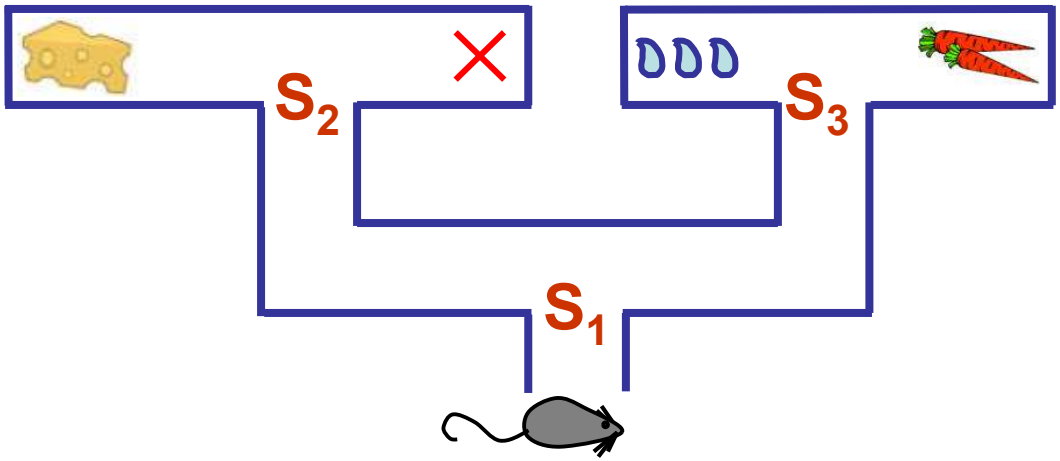


vmPFC
OFC/dACC
dPFC
SMA
Mx

Indirect Algorithms

- Bellman equation for Q-values:
  - $Q^*(s_t, a_t) = E[r(s_t, a_t) + \max_b Q^*(s_{t+1}, b)]$
- Q-learning: off-policy
  - $\Delta Q(s_t, a_t) \propto r_t + \max_b Q(s_{t+1}, b) - Q(s_t, a_t)$
  - rats (Roesch et al)
- SARSA: on-policy
  - $\Delta Q(s_t, a_t) \propto r_t + Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$
  - monkeys (Morris et al)

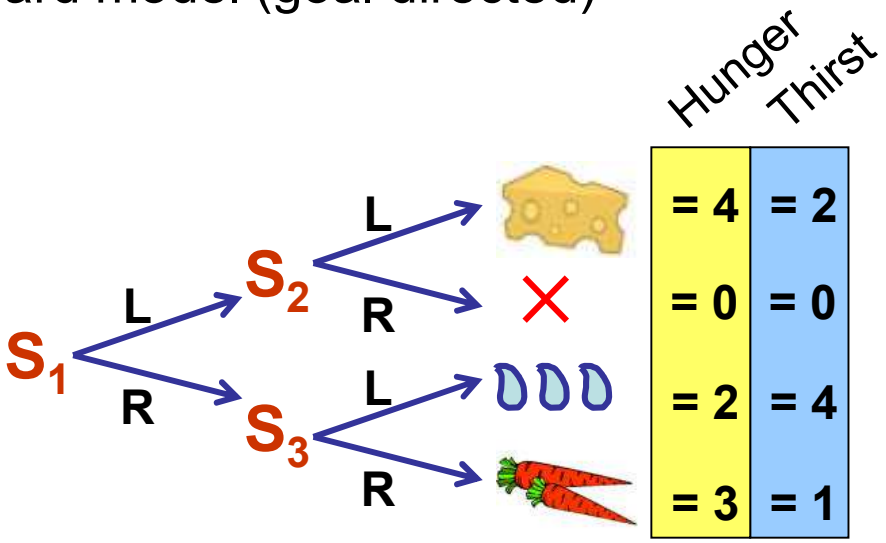
Reinforcement Learning

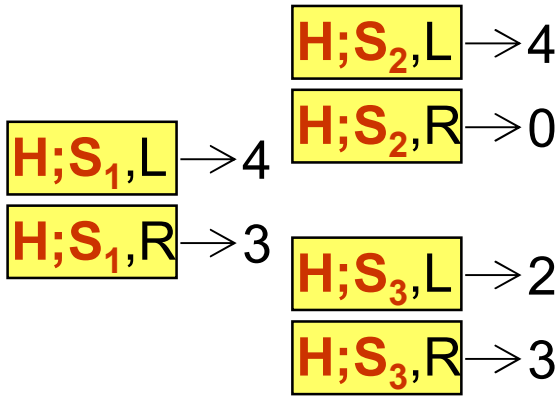


forward model (goal directed)

caching (habitual)

(NB: trained hungry)

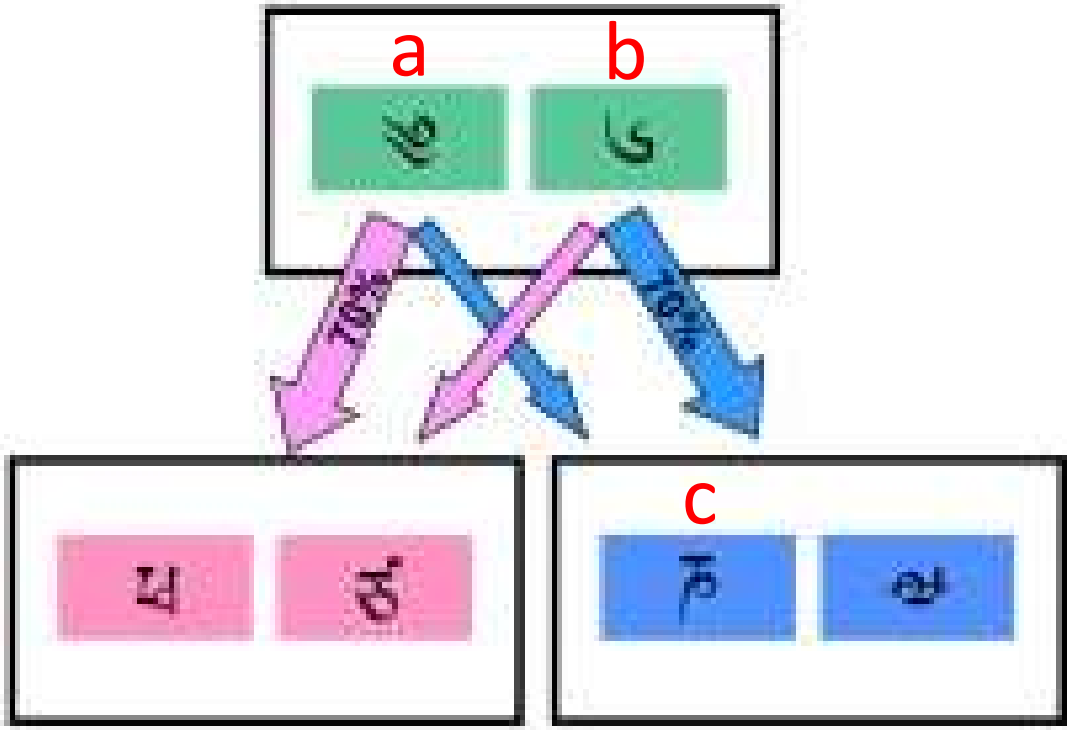




acquire with simple learning rules

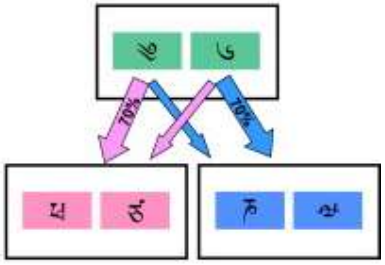
acquire recursively

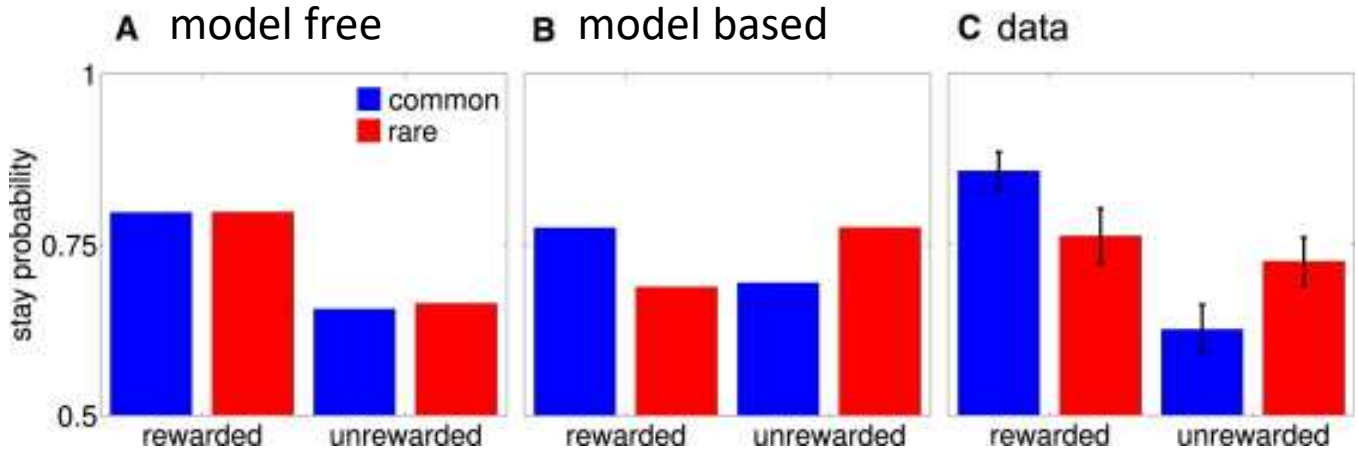
Human Canary...



- if a → c and c → **£££** , then do more of a or b?
  - MB: b
  - MF: a (or even no effect)

Behaviour





- assume a mix

$$Q_{tot}(s,a) = (1-\beta)Q_{MF}(s,a) + \beta Q_{MB}(s,a)$$

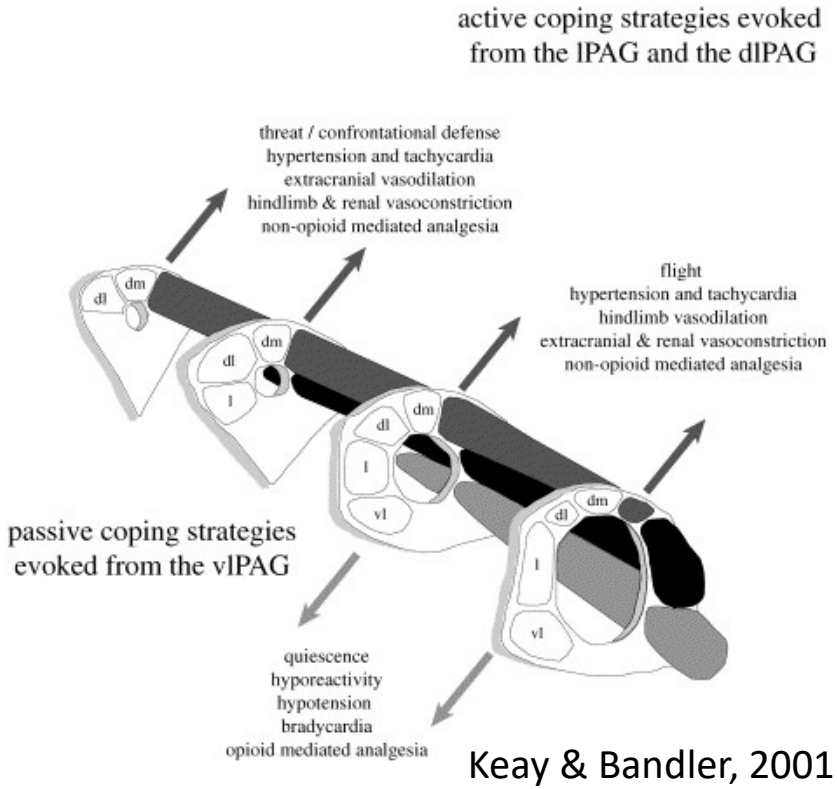
- expect that β will vary by subject (but be fixed)

Plan

- biological learning & Marr
- conditioning
  - classical/Pavlovian & prediction
  - instrumental/operant & action
  - temporal difference learning & dopamine
  - Pavlovian misbehaviour
- Bayesian conditioning
  - Kalman filtering
  - Chinese restaurant extinction

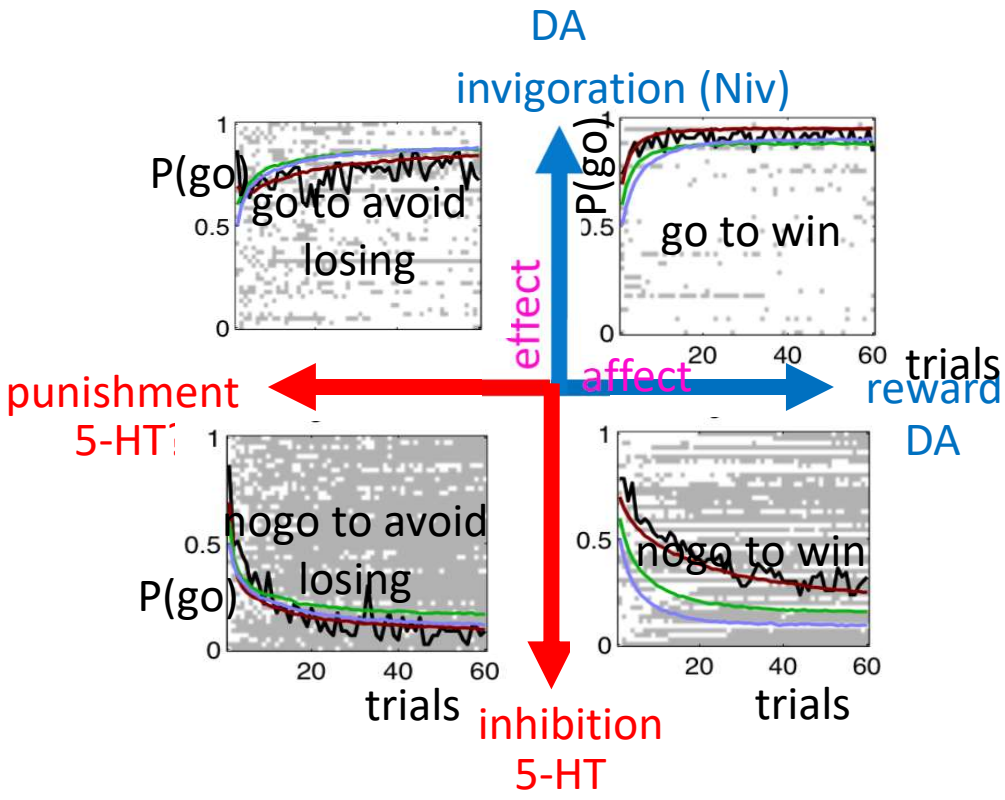
Pavlovian Control

Pavlovian Mechanisms



Keay & Bandler, 2001

Pavlovian Misbehaviour



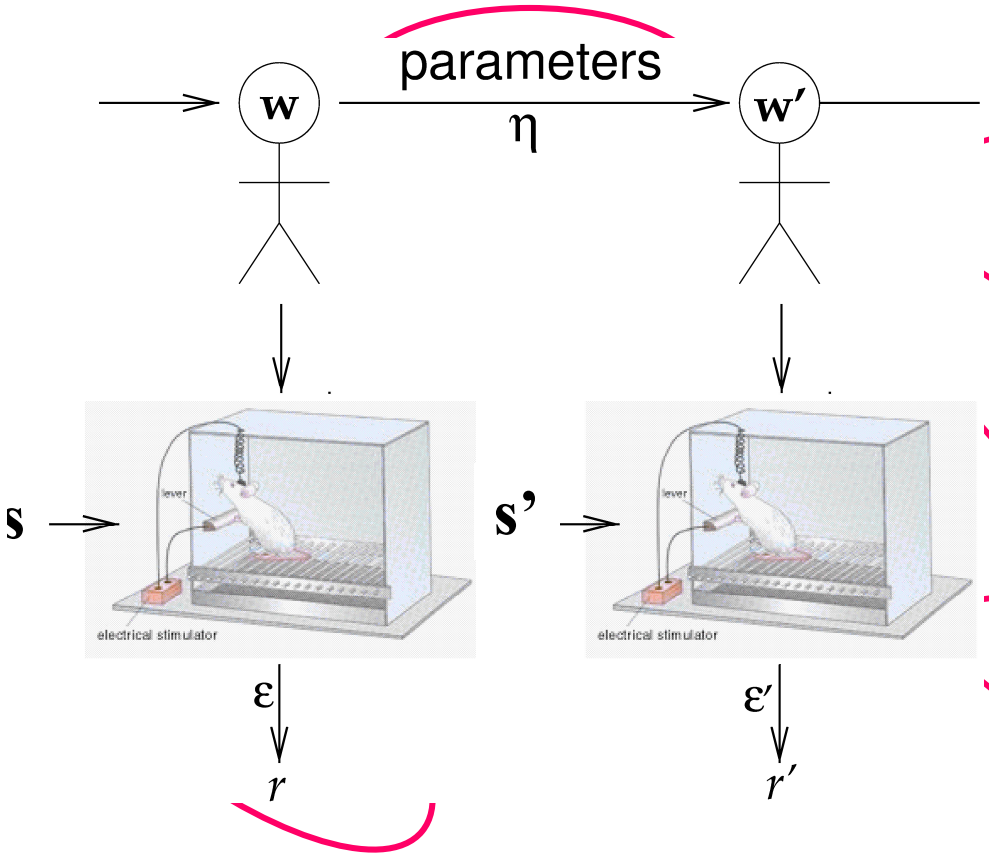
Crockett et al, 2009; Guitart-Masip et al, 2011; 2012

- **data**
- (delta rule)
- bias
  - go+b
- tremble
- Pavlovian
  - go+ωV_t(s)

Plan

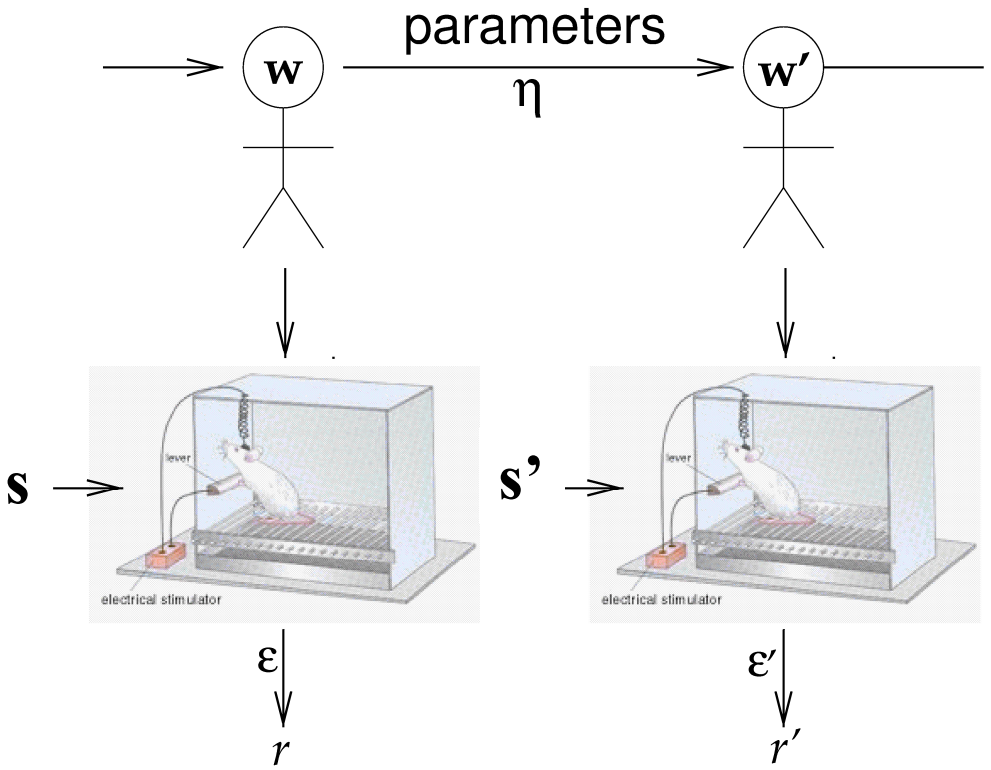
- biological learning & Marr
- conditioning
  - classical/Pavlovian & prediction
  - instrumental/operant & action
  - temporal difference learning & dopamine
  - Pavlovian misbehaviour
- Bayesian conditioning
  - Kalman filtering
  - Chinese restaurant extinction

Computational Conditioning



- output model
  - stimulus combination
- update model
  - expected/unexpected uncertainty
  - uncertainty-dependent learning rates
  - correlations
  - contexts
- decision model
  - action system competition
  - reinforcement learning

Kalman Filter



expt $\mathbf{w}' = \mathbf{w} + \boldsymbol{\eta}$

reward given $\boldsymbol{r} = \mathbf{w} \cdot \mathbf{s} + \varepsilon$

allowable drift $\boldsymbol{\eta} \sim \boldsymbol{N}[\mathbf{0}, \boldsymbol{\sigma}^{\mathbf{2}}\mathbb{I}]$

output noise $\varepsilon \sim \boldsymbol{N}[\mathbf{0}, \boldsymbol{\rho}^{\mathbf{2}}]$

- Markov random walk (or OU process)
- no punctate changes
- additive model of combination
- forward inference

Kalman Posterior

The Kalman filter maintains uncertainty:

$$P(V) = \mathcal{N}[\hat{\mathbf{w}} \cdot \mathbf{s}\,,\ \mathbf{s} \cdot \Sigma \cdot \mathbf{s}\,|$$

where

Assumed Density KF

Diagonal approx to $\Sigma = \text{diag}(\sigma_i^2)$

If $\mathbf{w} \sim \mathcal{N}\left[\hat{\mathbf{w}}, \text{diag}(\sigma_i^2)\right]$, then

$$\Delta \hat{w}_i = \frac{\sigma_i^2}{\sum_j \sigma_j^2 + \rho^2}(r - \mathbf{s} \cdot \hat{\mathbf{w}})u_i$$

- Rescorla-Wagner error correction
- competitive allocation of learning
  - Pearce & Hall

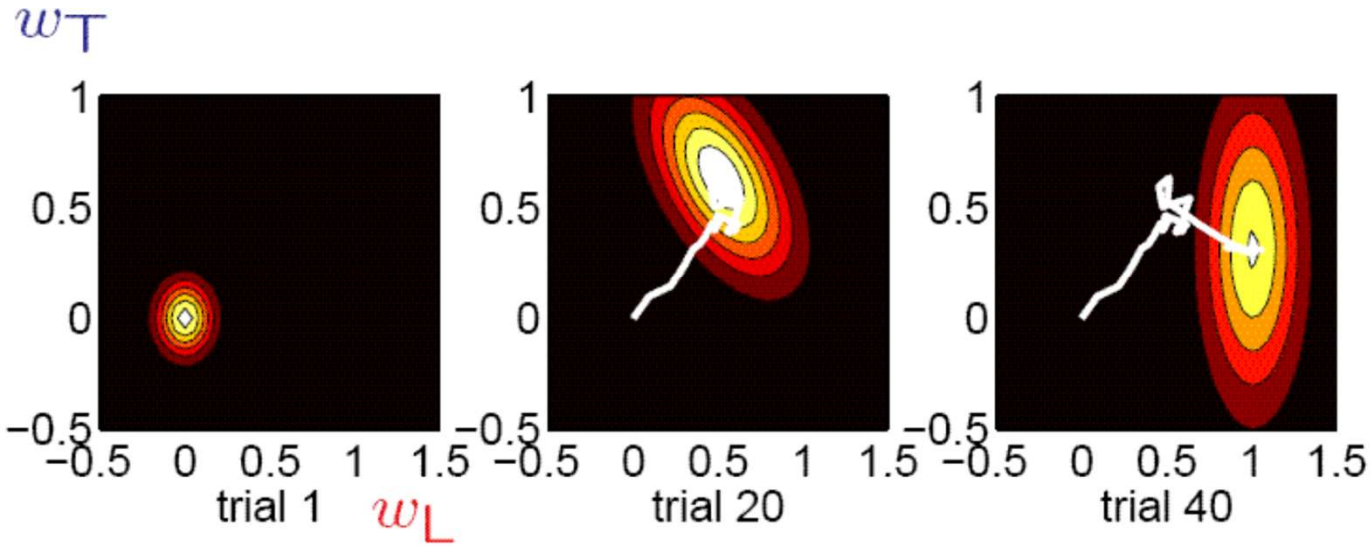
Blocking

forward	L→r	L+T→r	T→ ·
backward	L+T→r	L→r	$T \rightarrow \cdot$

- forward blocking: error correction

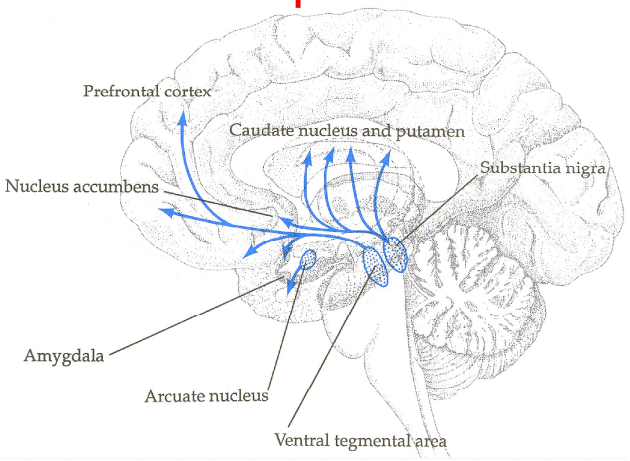
$$(r - \mathbf{s} \cdot \hat{\mathbf{w}})$$

- backward blocking: -ve **off-diag** $\Sigma_{LT} < 0$

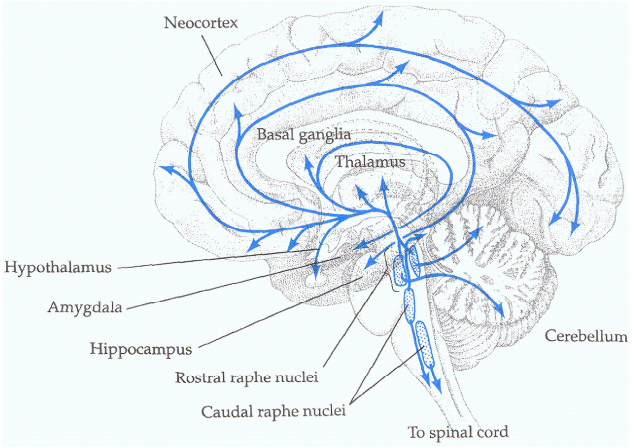


Meta-learning

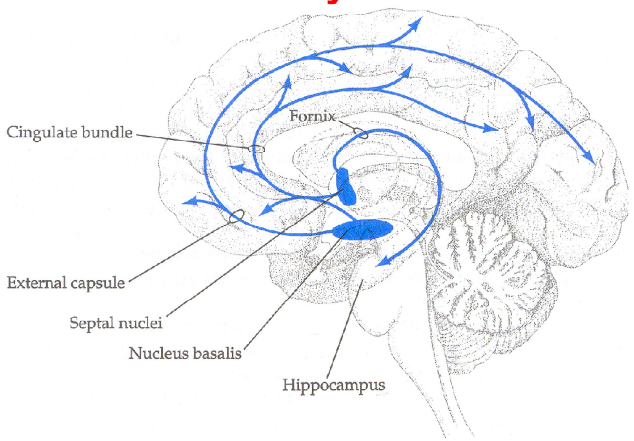
dopamine



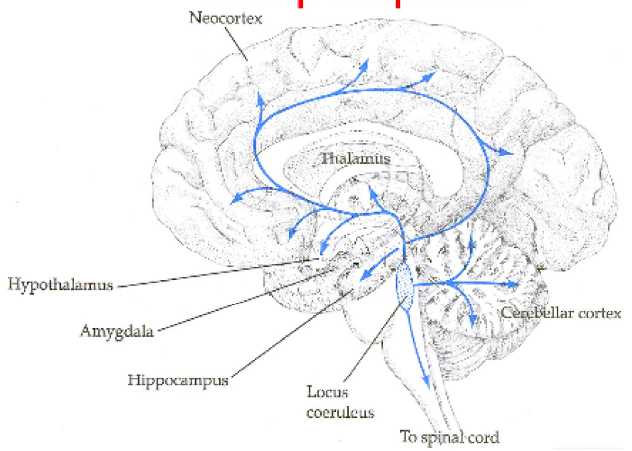
5HT



acetylcholine



norepinephrine



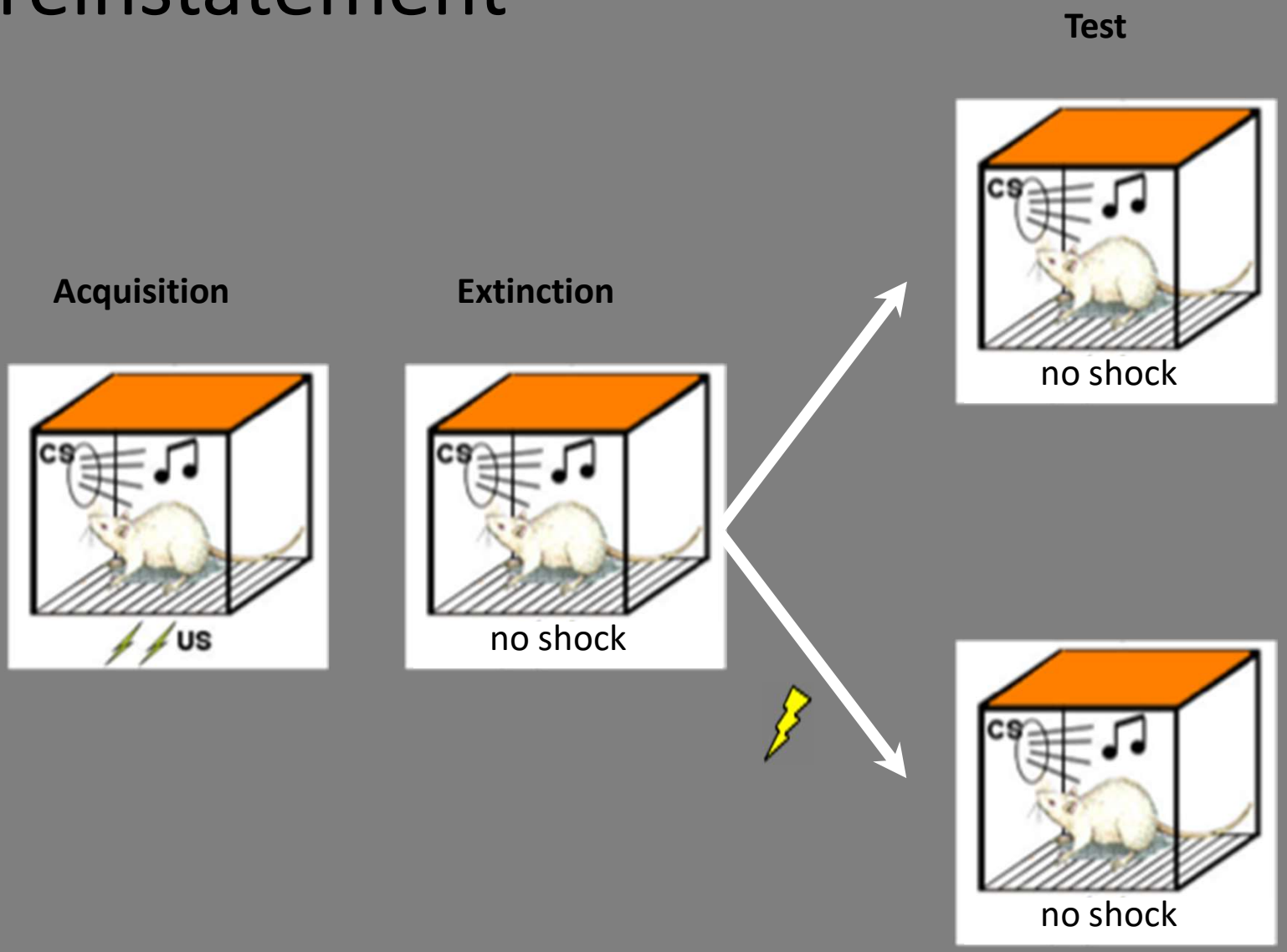
general excitability, signal/noise ratios

specific prediction errors, uncertainty signals

Plan

- biological learning & Marr
- conditioning
  - classical/Pavlovian & prediction
  - instrumental/operant & action
  - temporal difference learning & dopamine
  - Pavlovian misbehaviour
- Bayesian conditioning
  - Kalman filtering
  - Chinese restaurant extinction

reinstatement



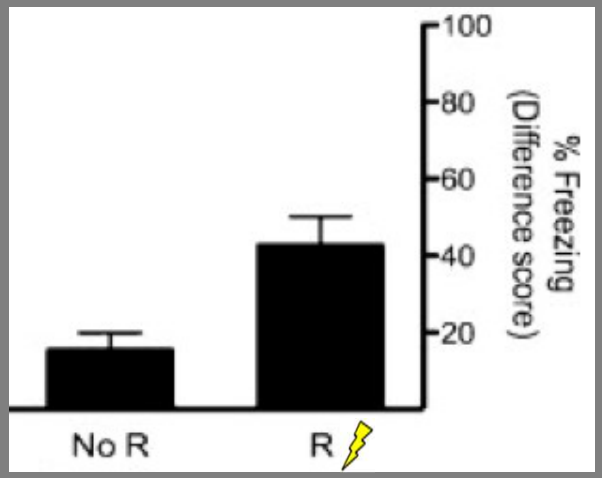
slides from Yael Niv

extinction ≠ unlearning

Acquisition

Extinction

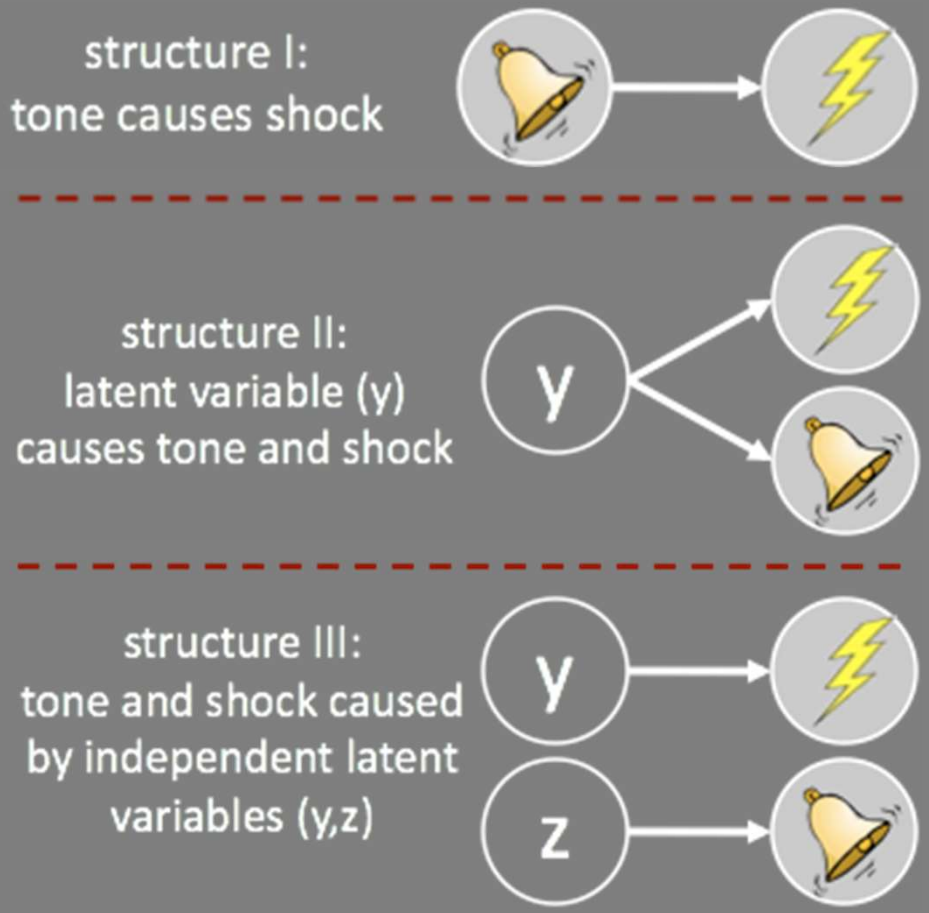
Test



Storsve, McNally & Richardson, 2012

slides from Yael Niv

learning causal structure: Gershman & Niv

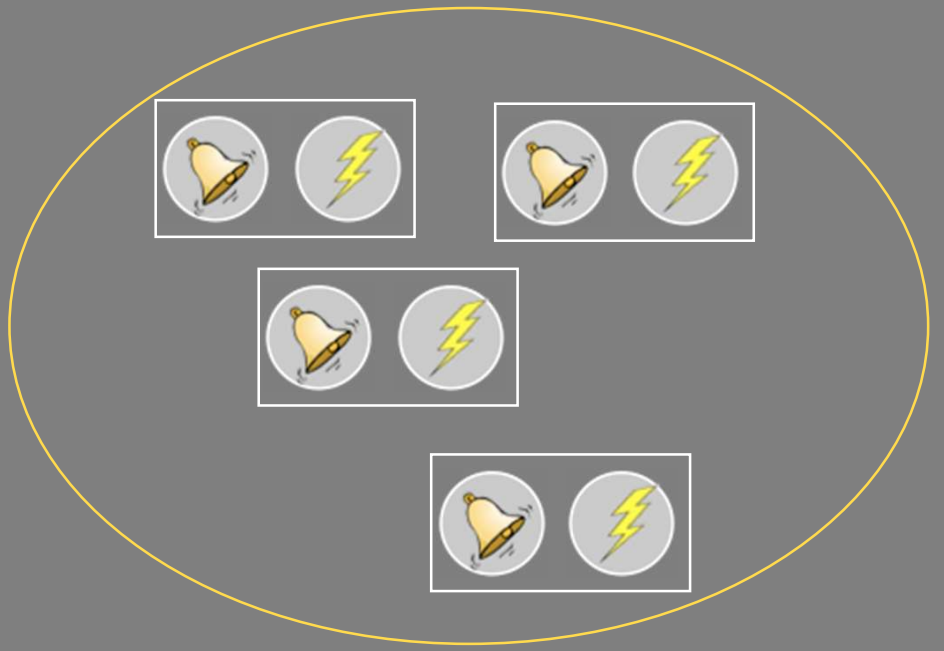


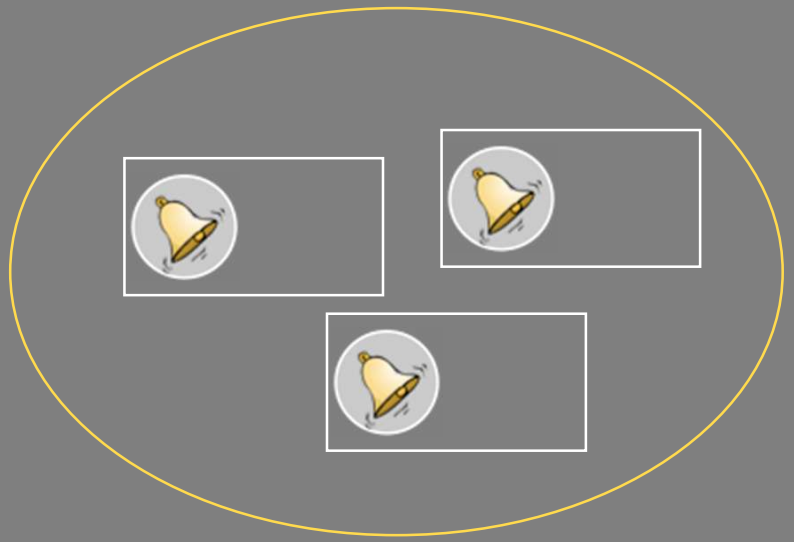
Sam Gershman

conditioning as clustering: DPM

Gershman & Niv;

Daw & Courville; Redish

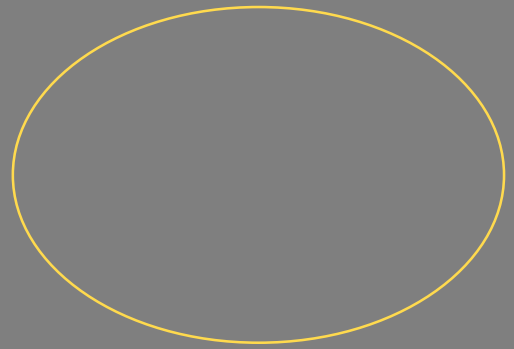


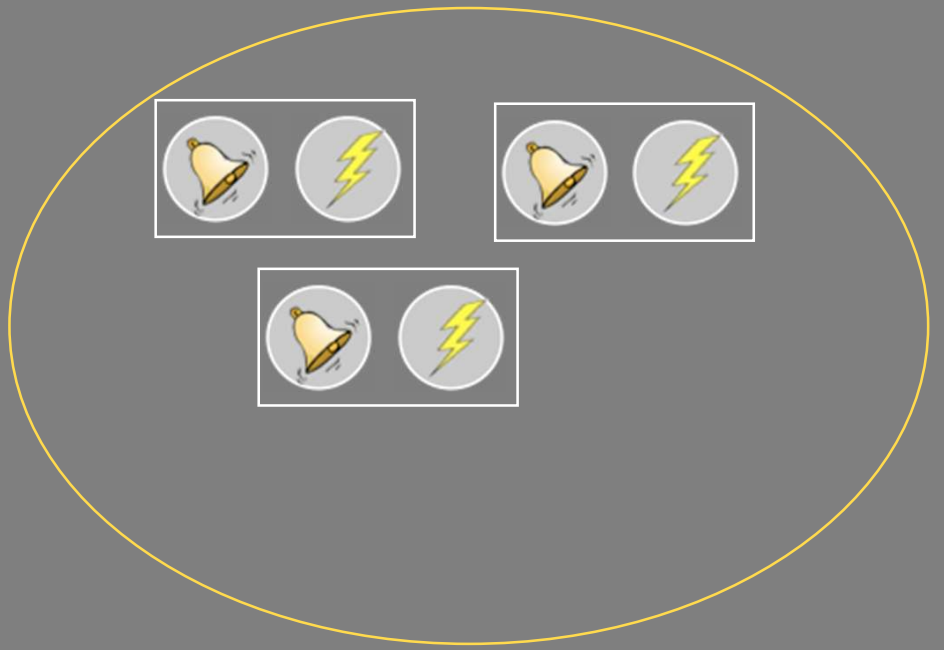


Within each cluster:
"learning as usual"
(Rescorla-Wagner, RL etc.)

associative learning versus state learning

Gershman & Niv

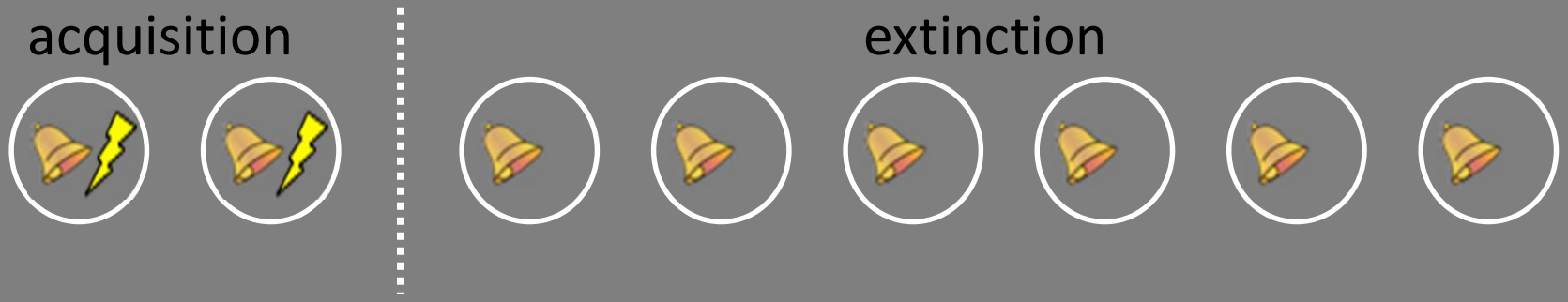




structural learning
(create new state)

how to erase a fear memory

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

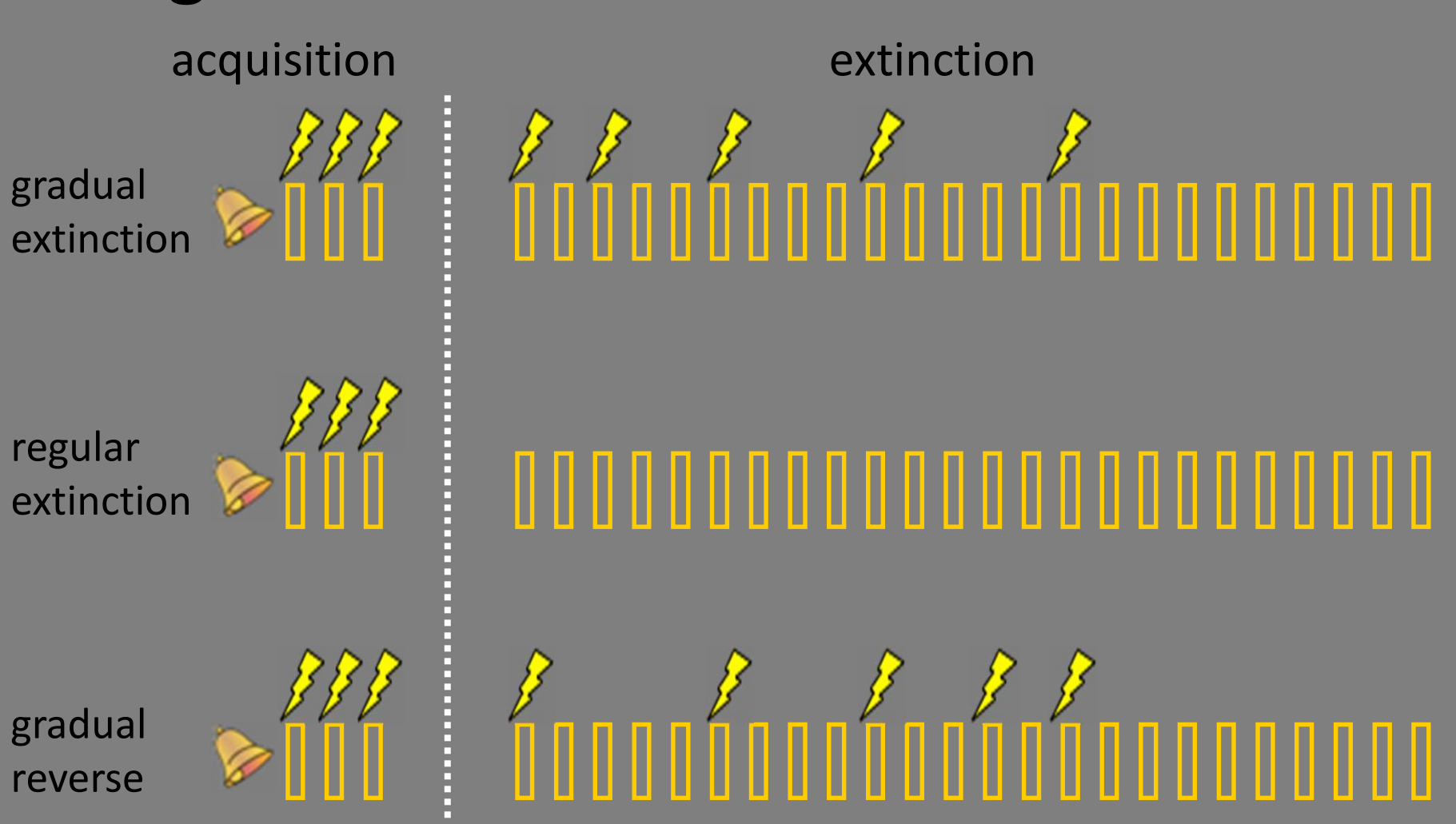


what if we make extinction a bit more similar to acquisition?

slides from Yael Niv

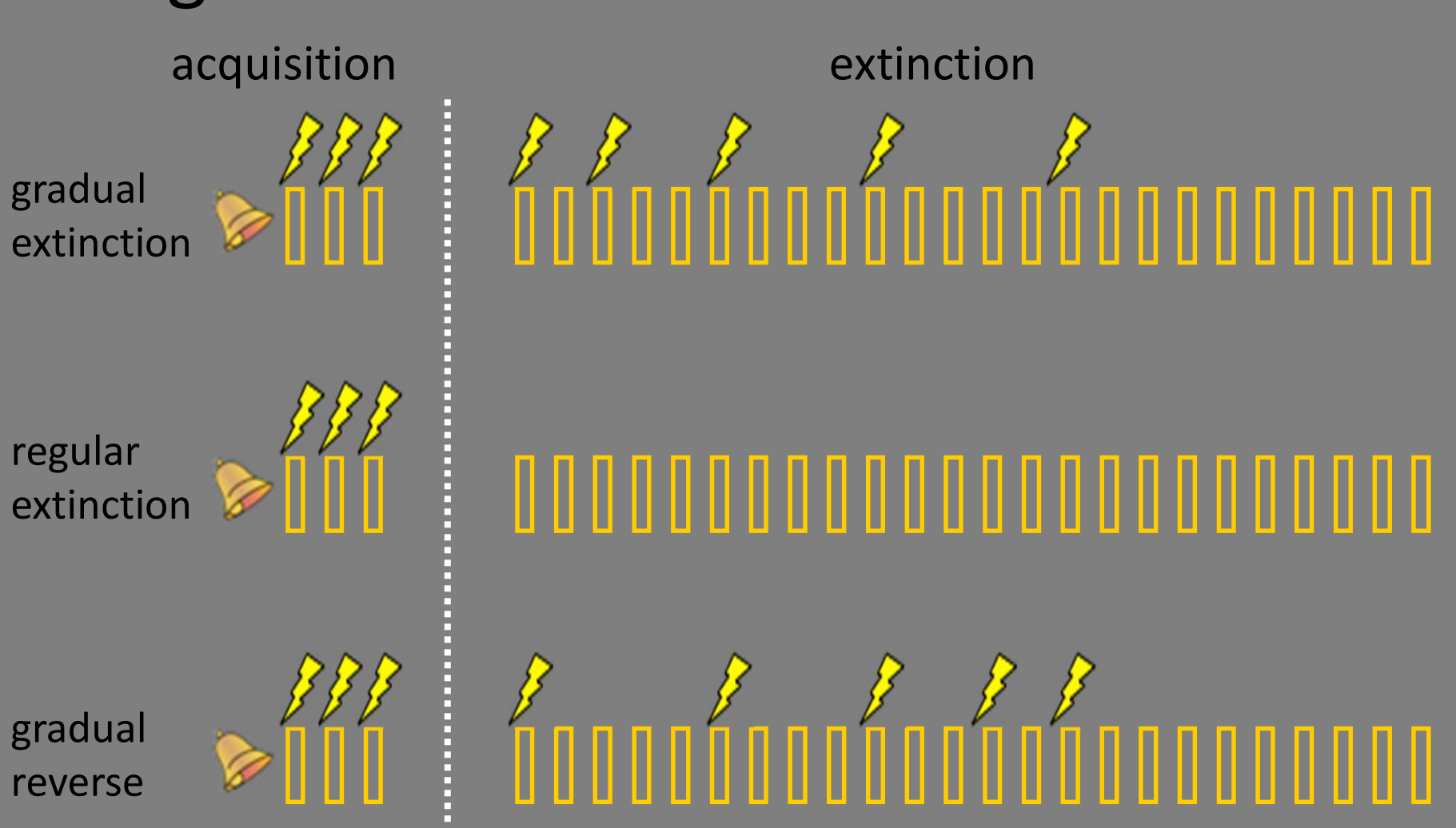
gradual extinction

Gershman, Jones, Norman, Monfils & Niv



gradual extinction

Gershman, Jones, Norman, Monfils & Niv

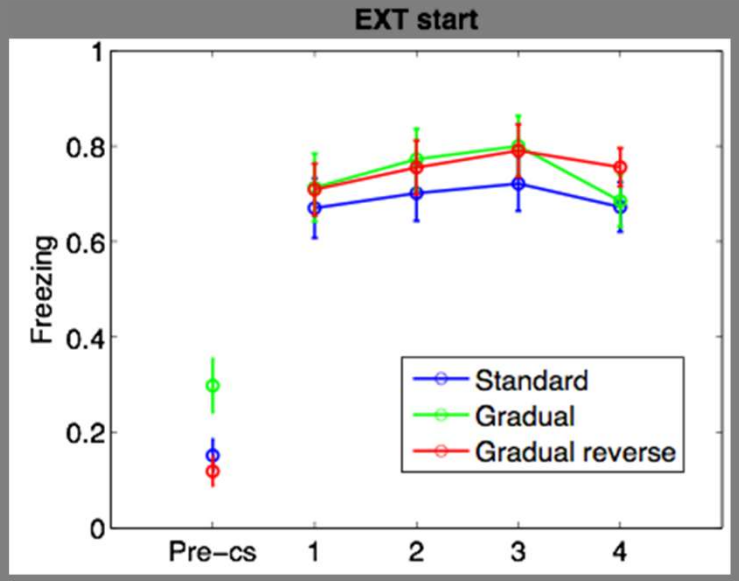


test one day (reinstatement) or 30 days later (spontaneous recovery)

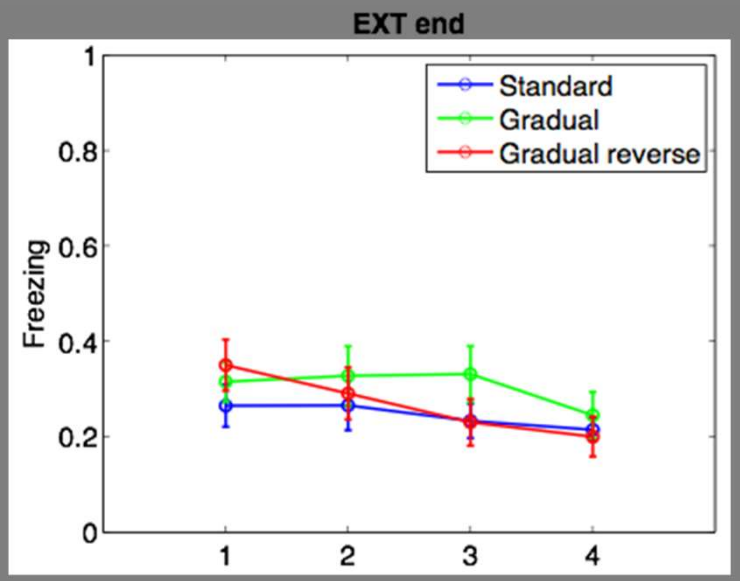
gradual extinction

Gershman, Jones, Norman, Monfils & Niv - under review

first trials of extinction

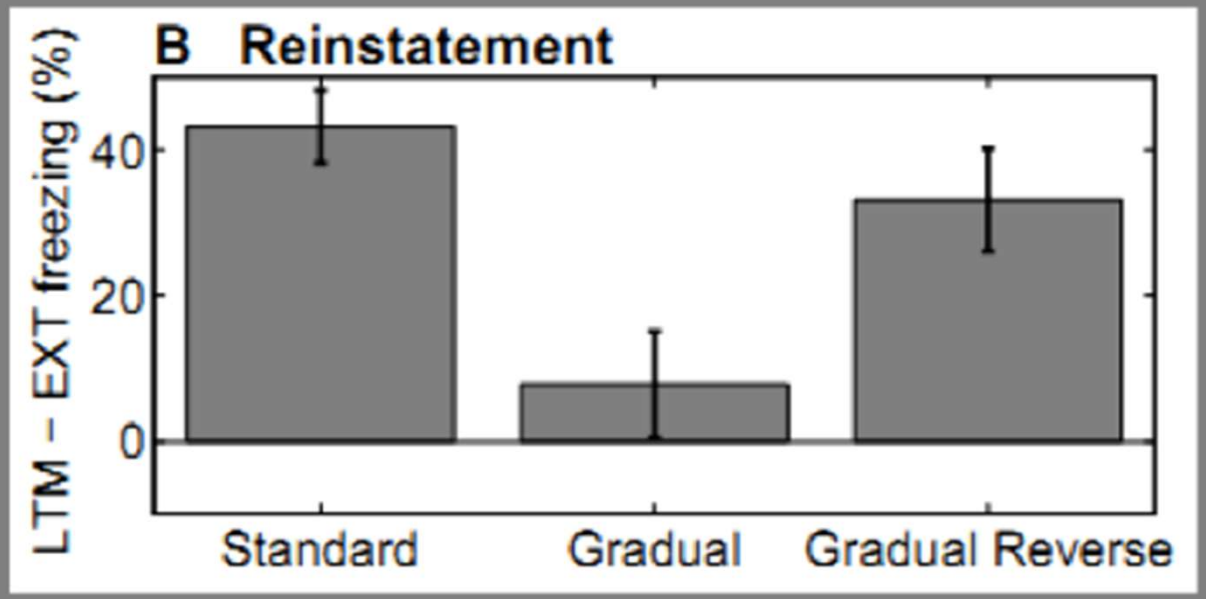


last trials of extinction



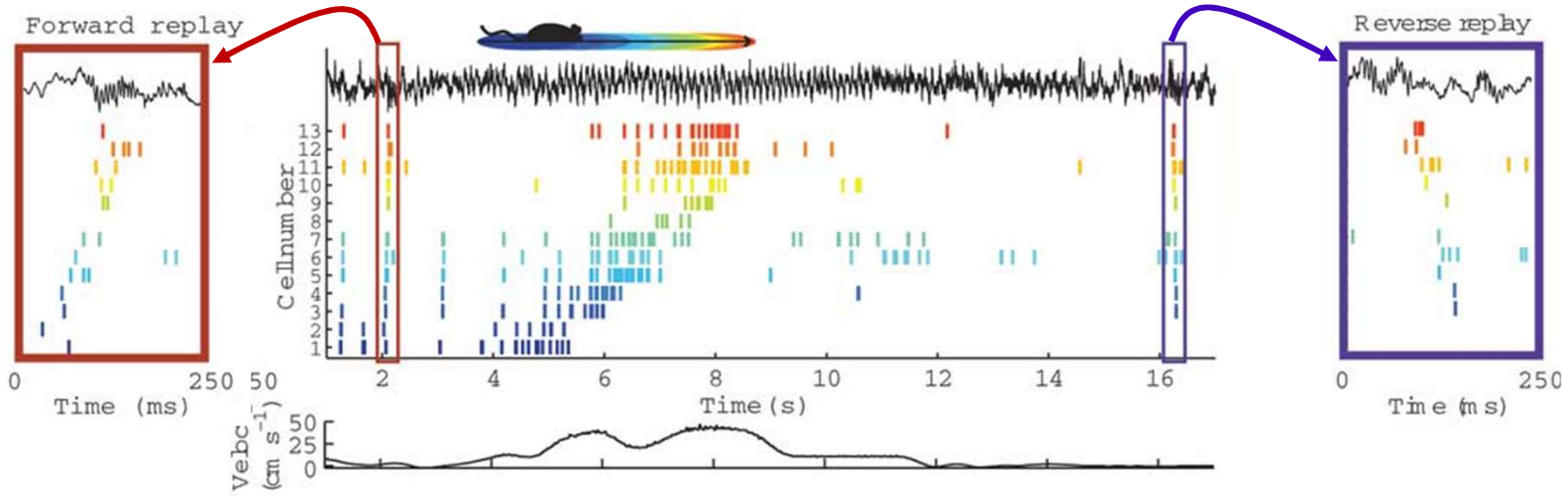
gradual extinction

Gershman, Jones, Norman, Monfils & Niv



only gradual extinction group shows no reinstatement

Wake/'Sleep'



Diba & Buzsaki 2007; Carr et al (2011) Nat Neurosci 14:147

- replay: offline model
  - maintenance
  - inversion

Discussion

- different ‘sorts’ of learning
  - supervised
  - self-supervised
  - reinforcement
- different algorithms
  - back-propagation
  - Hebb/anti-Hebb/delta + MCMC
  - three-term rules
- different neural substrates:
  - hippocampus striatum
  - cortex cerebellum