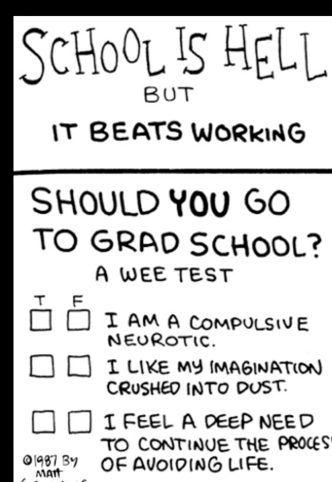
SOBP Satellite: Computational Psychiatry

Reinforcement Learning: Theory

Yael Niv
Psychology Department & Princeton Neuroscience Institute
yael@princeton.edu

why is decision making hard?



- Reward/punishment may be delayed
- Outcomes may depend on a series of actions

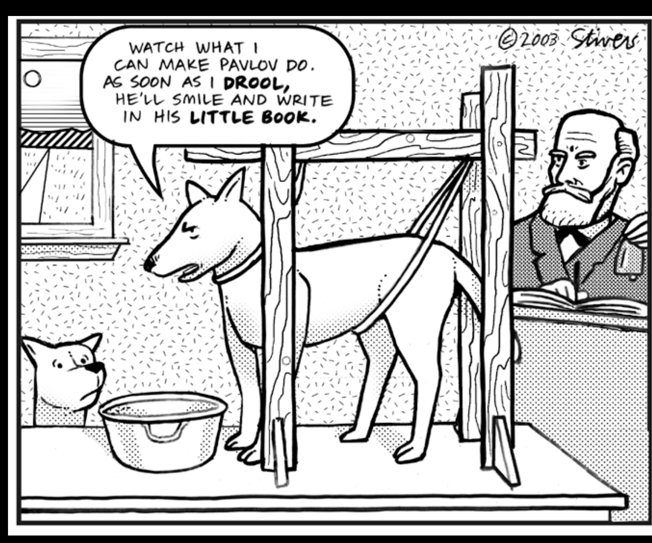
⇒ "credit assignment problem" (Sutton, 1978)

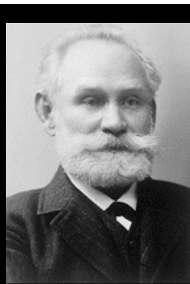
How does the brain solve this problem?

for this: we need to learn two basic things

1. what is going to happen (prediction learning)
2. what to do about it (action learning)

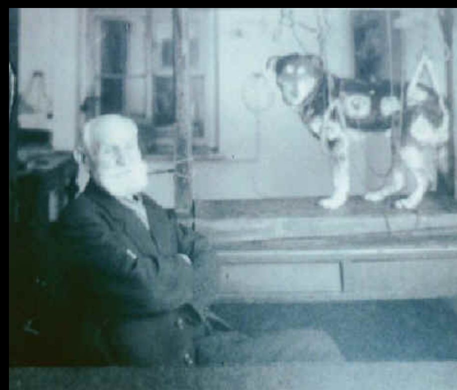
Act I: what are animals really learning?





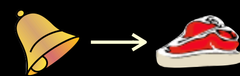
Ivan Pavlov
(Nobel prize portrait)

animals learn predictions

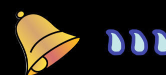


pair stimulus

...with significant event



measure anticipatory behavior



 = Unconditional Stimulus (US)

 = Conditional Stimulus (CS)

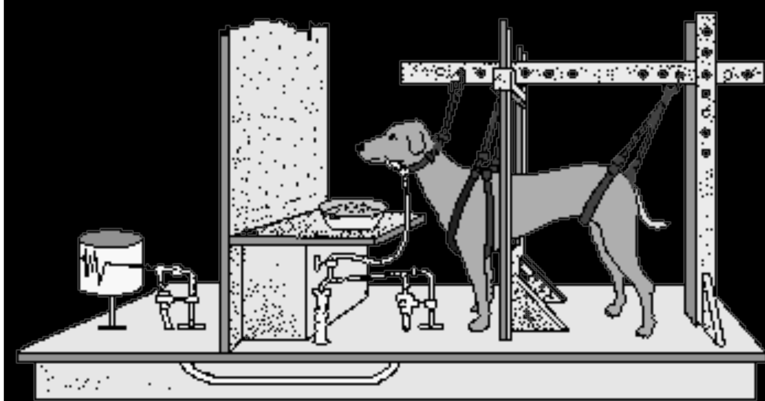
 = Conditional Response (CR)
(here, also Unconditional Response; UR)

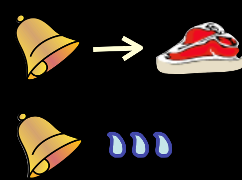
Very general form of learning from experience (snails - humans)

example: pigeon appetitive conditioning

- behavior reveals predictions
- behavior seems compulsive -- hard to avoid
- even at a cost
- and if it prevents the appetitive outcome altogether

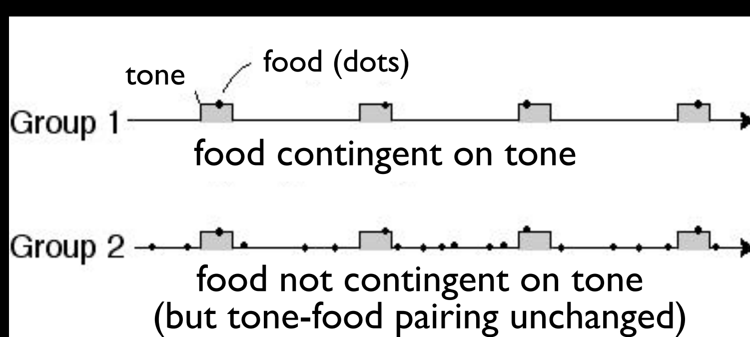
back to basic classical conditioning





Under what conditions does learning occur?

1) Rescorla's control condition



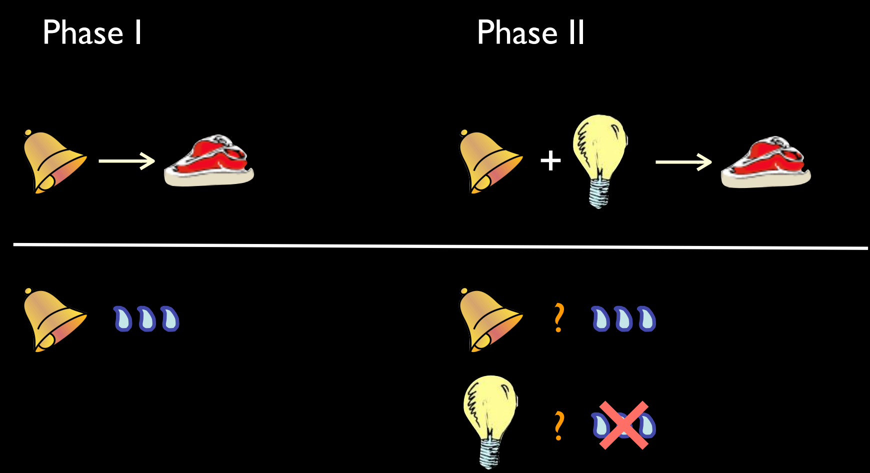
will Group 2 show a conditioned response to the tone?

temporal contiguity is not enough - need contingency

$P(\text{food} \mid \text{tone}) \neq P(\text{food} \mid \text{no tone})$

Credits: Randy Gallistel

2) Kamin's blocking



contingency is also not enough.. need <u>surprise</u>

P(food | noise+light) ≠ P(food | noise alone)

Summary so far...

- Naïvely it had seemed that pairing a neutral stimulus with a motivationally significant one is enough for prediction learning...
- ...but we also need contingency and surprise
- A super simple theory ("where is the theory? I only see one equation"):

Rescorla & Wagner (1972)

The idea: error-driven learning*

Change in value is proportional to the difference between actual and predicted outcome

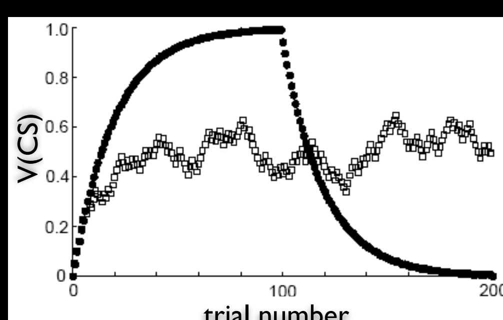
$$\Delta V(CS_i) = \underbrace{\eta}_{\text{learning rate}}[\underbrace{R_{US}}_{\text{actual outcome value}} - \sum_{j \in \text{trial}} \underbrace{V(CS_j)}_{\text{value as prediction}}]$$

* strictly speaking it was Bush & Mosteller's (1951) idea

Rescorla & Wagner (1972)

$$V_{T+1} = V_T + \eta[R_T - V_T]$$

- what would happen with random 50% reinforcement? eg. 1 1 0 1 0 0 1 1 1 0 0
  - what would V be *on average* after learning?
  - what would the error term look like after learning?



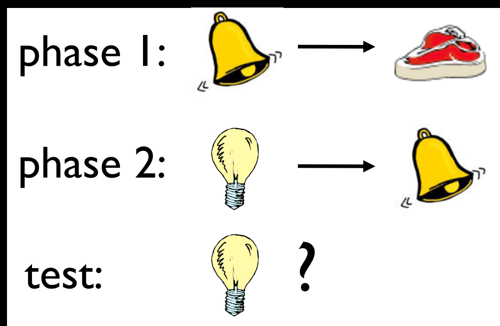
- can you estimate what learning rate (or step size) η was used in this simulation? (try to think how you could do the same from behavioral data)

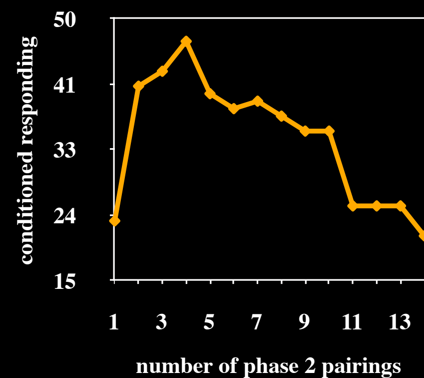
Summary so far...

- Animals (including humans) learn predictions
- Prediction learning can be explained by an error-correcting learning rule:
  predictions are learned from experiencing the world and comparing predictions to reality
  (ie, learning from prediction errors)
- Rescorla-Wagner: A simple but very powerful model

Act 2: Is that so?
(or: there is always a "but..")

But: second order conditioning





what do you think will happen?

animals learn that a predictor of a predictor is also a predictor!
⇒ not interested solely in predicting immediate reinforcement..

helpful heritage from computer science:

David Marr (1945-1980, computational vision) proposed three levels of analysis:

1. the problem (Computational Level)
2. the strategy (Algorithmic Level)
3. how it is actually done by networks of neurons (Implementational Level)

developing a model, now more formally

The problem: optimal prediction of future reinforcement

$$V_t = E\left[\sum_{i=t+1}^{\infty} r_i\right]$$

want to predict expected sum of future reinforcement

$$V_t = E\left[\sum_{i=t+1}^{\infty} \gamma^{i-t-1} r_i\right]$$

want to predict expected sum of *discounted* future reinforcement ($0<\gamma<1$)

$$V_t = E\left[\sum_{i=t+1}^{t_{end}} r_i\right]$$

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

developing a model, now more formally

The problem: optimal prediction of future reinforcement

$$\begin{aligned} V_t &= E[r_{t+1} + r_{t+2} + ... + r_{t_{end}}] \\ &= E[r_{t+1}] + E[r_{t+2} + ... + r_{t_{end}}] \\ &= E\left[r_{t+1}\right] + V_{t+1} \end{aligned}$$

(note: t indexes time <u>within a trial</u>)

$$V_t = E\left[\sum_{i=t+1}^{t_{end}} r_i\right]$$

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

developing a model, now more formally

The problem: optimal prediction of future reinforcement

$$\begin{aligned} V_t &= E[r_{t+1} + r_{t+2} + \ldots + r_{t_{end}}] \\ &= E[r_{t+1}] + E[r_{t+2} + \ldots + r_{t_{end}}] \\ &= E\left[r_{t+1}\right] + V_{t+1} \end{aligned}$$

(note: *t* indexes time within a trial)

Temporal Difference (TD) learning

The problem: optimal prediction of future reinforcement

The algorithm: $V_t = E\left[r_{t+1}\right] + V_{t+1}$

$$V_t^{T+1} = V_t^T + \eta \underbrace{\left(r_{t+1}^T + V_{t+1}^T - V_t^T\right)}_{\text{temporal difference prediction error } \delta(t+1)}$$

(note: *t* indexes time within a trial, T indexes trials)

compare to: $V^{T+1} = V^T + \eta\left(r^T - V^T\right)$

Sutton & Barto 1983, 1990

Act 3 - remedies for a faulty fortune teller (dopamine and prediction errors)

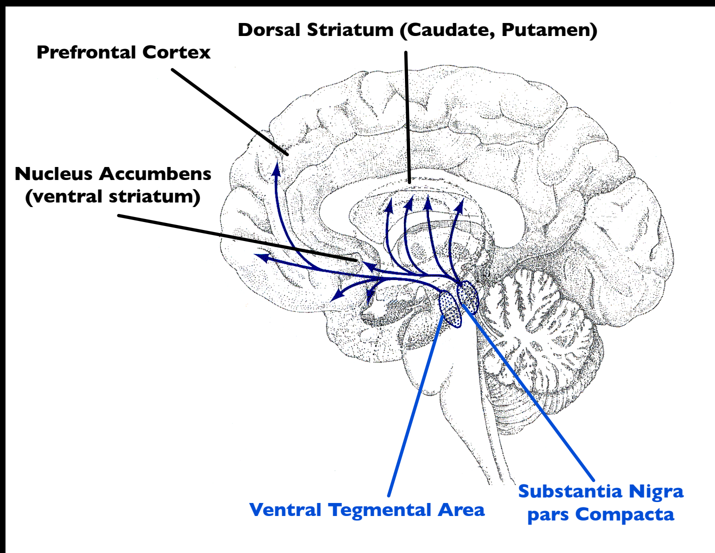
Back to Marr's three levels

The problem: prediction of future reward/punishment

The algorithm: Rescorla-Wagner/temporal difference learning, aka, learning from prediction errors

Neural implementation: does the brain use prediction errors for learning?

dopamine does everything



Parkinson's Disease
→ Motor control

but also: drug addiction, gambling, natural rewards

→ Reward pathway?
→ Learning?

Also involved in:

- Working memory
- Novel situations
- ADHD
- Schizophrenia
- …

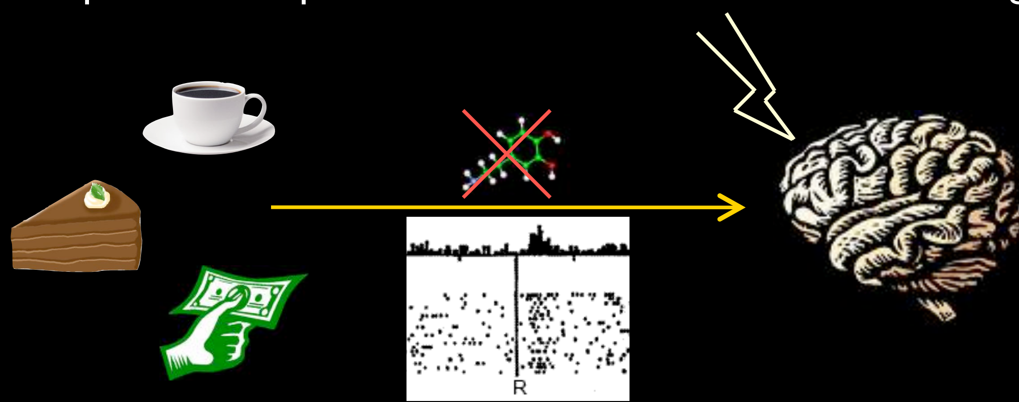
dopamine and conditioning

- Dopamine antagonists:
  disrupt regular Pavlovian conditioning
- Self-stimulation experiments:
  stimulation of dopamine pathways is "rewarding"

Peter Shizgal, Concordia

the anhedonia hypothesis (Wise, ’80s)

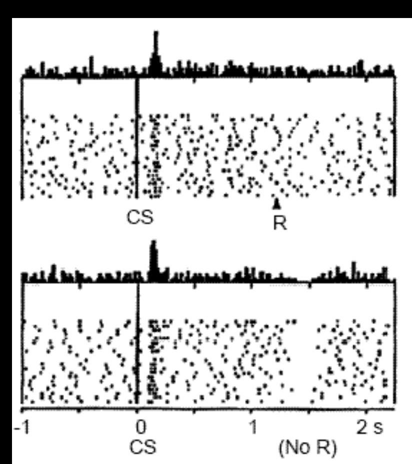
- Anhedonia = inability to experience positive emotional states derived from obtaining a desired or biologically significant stimulus
- Neuroleptics (dopamine antagonists) cause anhedonia
- Dopamine is important for reward-mediated conditioning



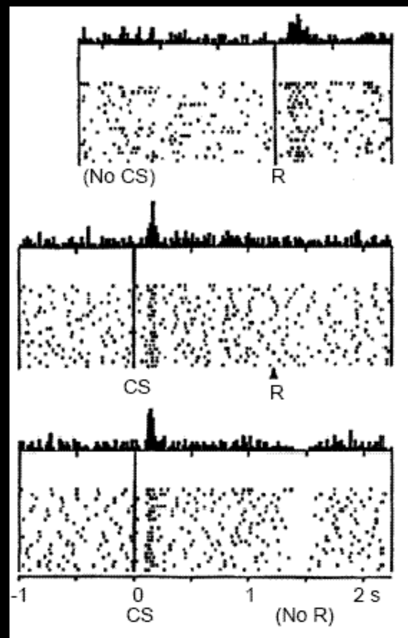
but...

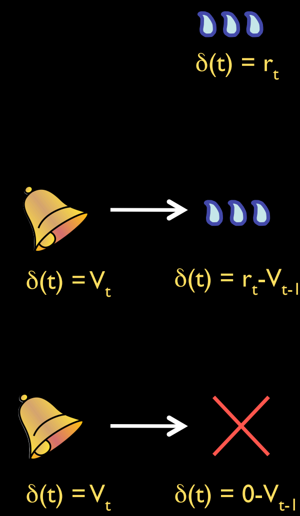
predictable reward

omitted reward



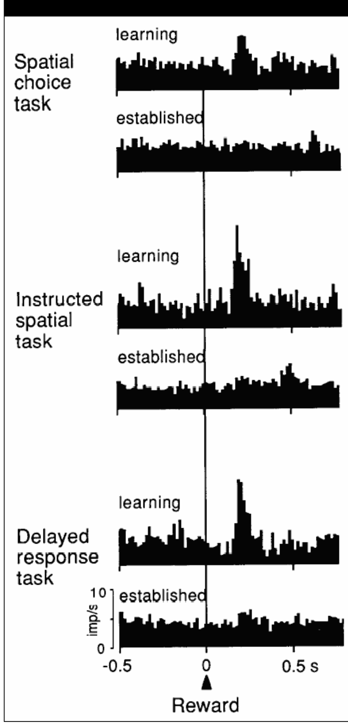
what are we looking at?





Schultz, Dayan & Montague 1997

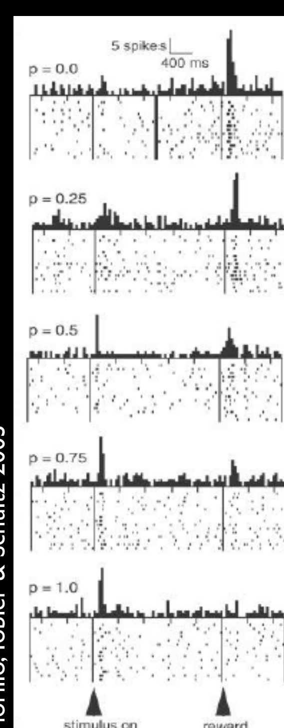
prediction error hypothesis of dopamine



Schultz, Apicella & Ljungberg, 1993

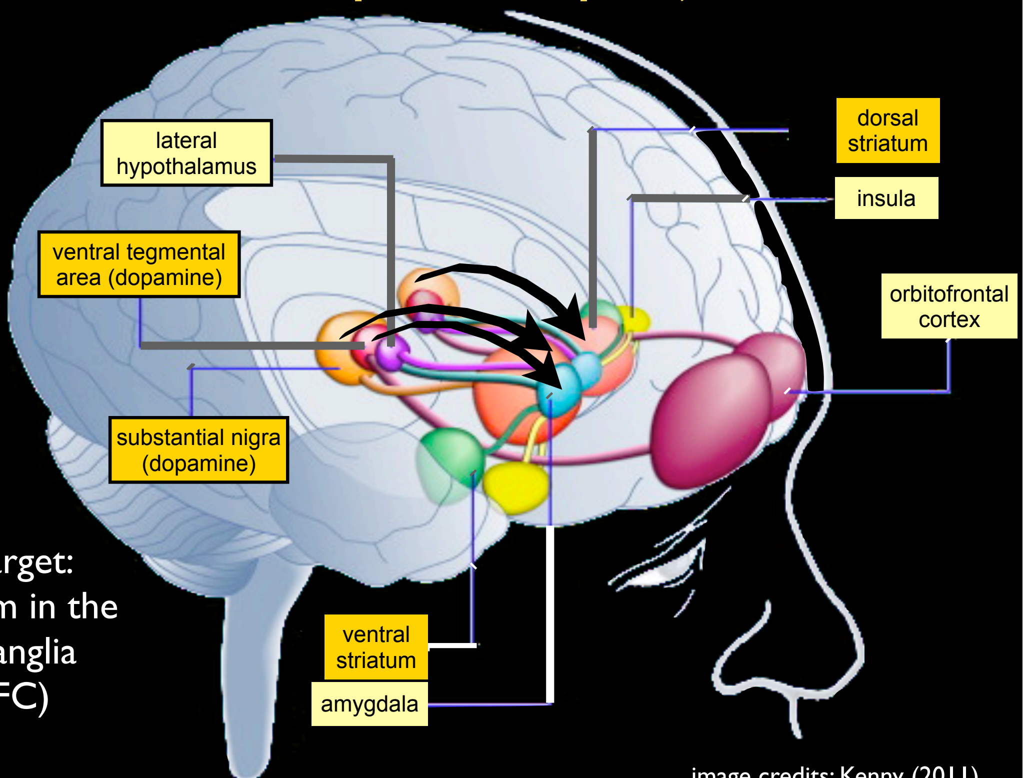
The idea: Dopamine encodes a reward prediction error

(Montague, Dayan, Barto mid 90's)



Fiorillo, Tobler & Schultz 2003

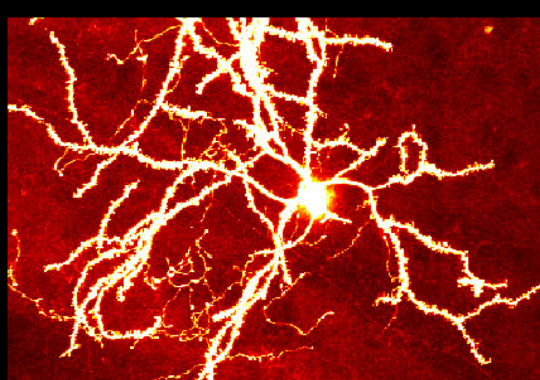
where does dopamine project to?

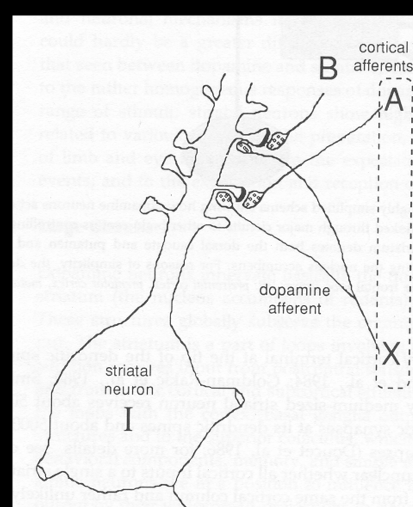


main target:
striatum in the
basal ganglia
(also PFC)

image credits: Kenny (2011)

organization of cortico-striatal synapses

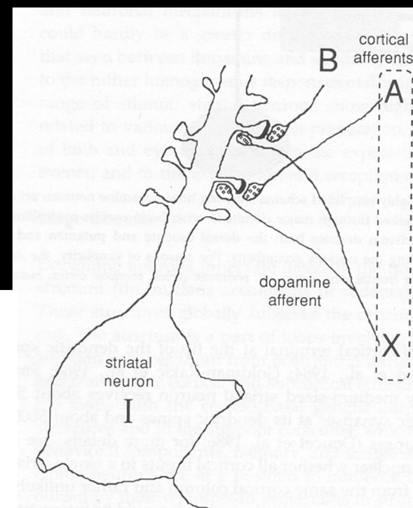


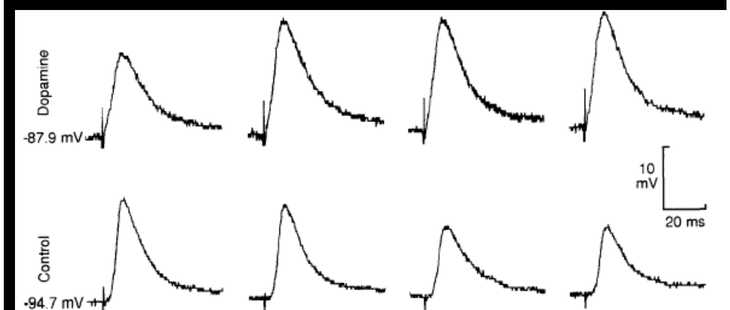


Wickens et al, 1996

dopamine and synaptic plasticity

- prediction errors are for learning…
- cortico-striatal synapses show dopamine-dependent plasticity
- three-factor learning rule: need presynaptic+postsynaptic+dopamine



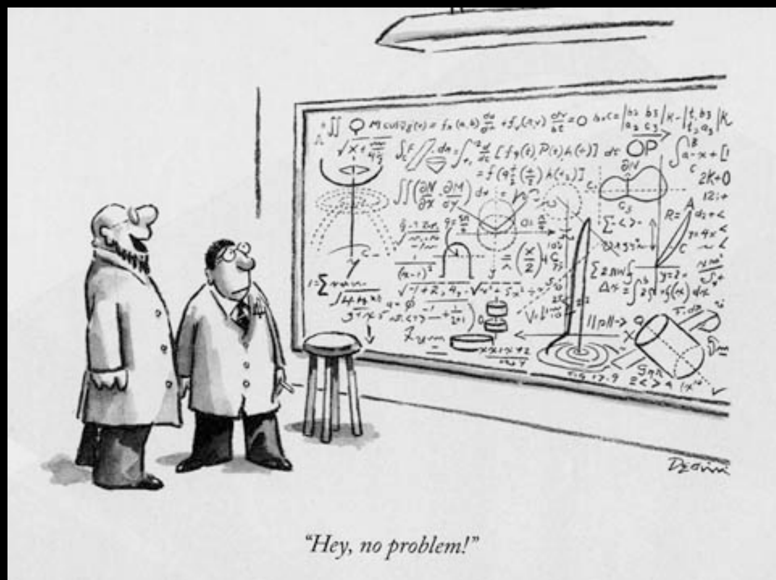


Wickens, Begg & Arbuthnott 1996

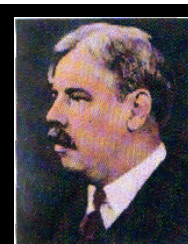
Summary so far...

- Temporal difference learning is a "better" version of Rescorla-Wagner learning
- derived from first principles (from definition of problem)
- explains everything that R-W does, and more (eg. 2nd order conditioning)
- basically a generalization of R-W to real time

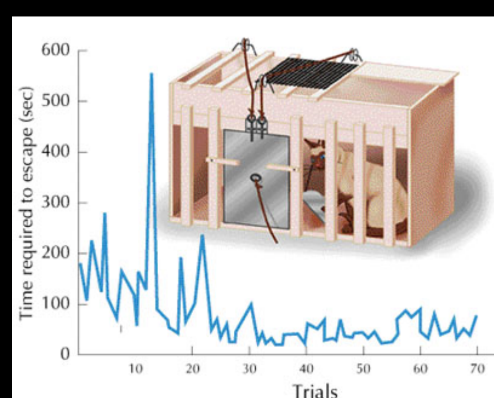
Act 4: Now what do we do?



Edward Thorndike (1874-1949)

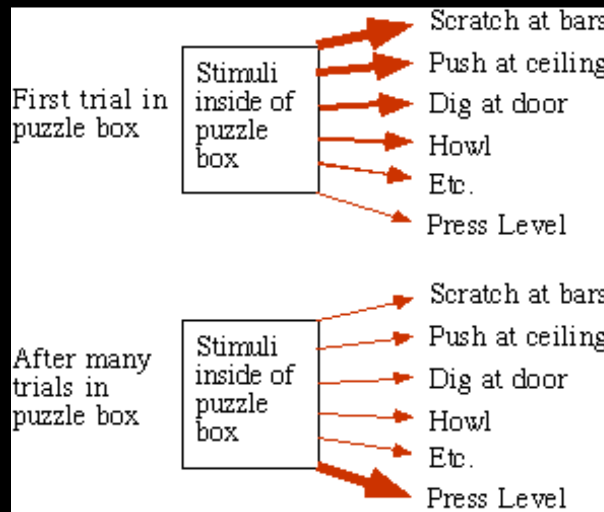


- Background: Darwin, attempts to show that animals are intelligent
- Tested hungry cats in "puzzle boxes"
- Operational definition for learning: time to escape
- Gradual learning curves, trial and error rather than 'insight'

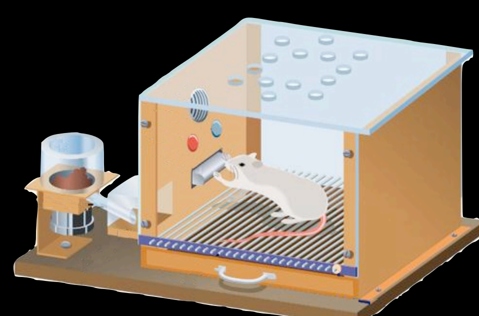


Thorndike: The Law of Effect

Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond.

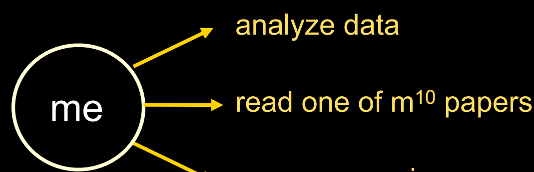


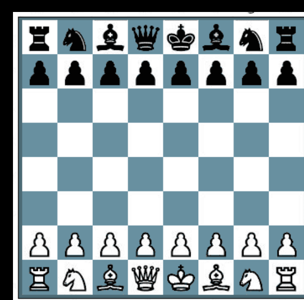
instrumental conditioning as adaptive control



how to model instrumental conditioning?

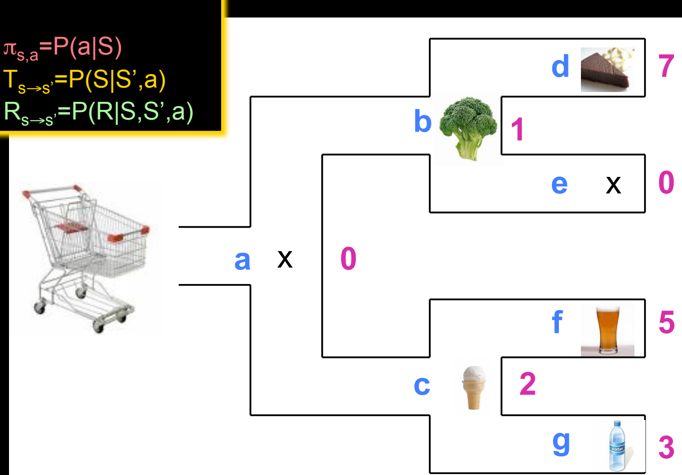
- The problem: find the best behavioral policy
  (i.e., what to do in what situation)
  best in terms of?
- The real problem: the credit assignment problem
- Algorithms: Reinforcement learning





more formally: MDPs

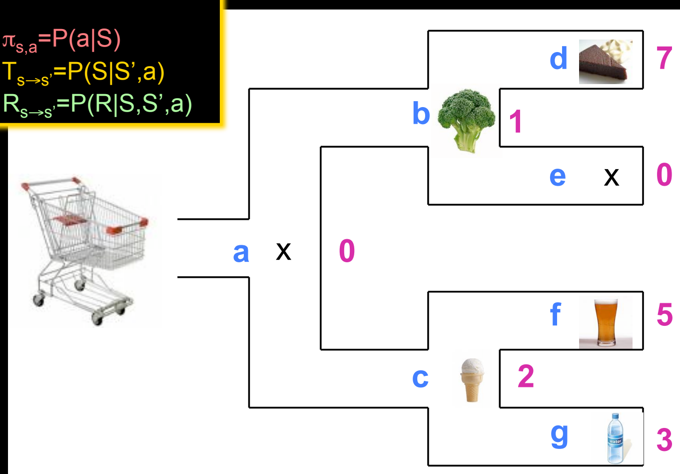
- States S
- Actions $\pi_{s,a}$=P(a|S)
- Transitions $T_{s\to s'}$=P(S|S',a)
- Rewards $R_{s\to s'}$=P(R|S,S',a)



transitions: P(b|a,left) = 90%; P(c|a,left) = 10% etc.
(wonky shopping cart)

The Markov property

- States S
- Actions $\pi_{s,a}$=P(a|S)
- Transitions $T_{s \to s'}$=P(S|S',a)
- Rewards $R_{s \to s'}$=P(R|S,S',a)



- The idea: given the current situation, history does not matter
- $P(S_{t+1}|S_1,S_2,\ldots,S_t,a_1,a_2,\ldots,a_t) = P(S_{t+1}|S_t,a_t)$
- $P(r_t|S_1,S_2,\ldots,S_t,a_1,a_2,\ldots,a_t) = P(r_t|S_t,a_t)$

Stylized task: described fully by S,A,R,T

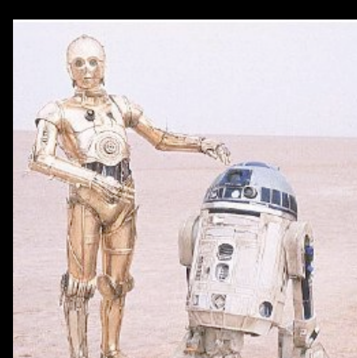
World: "You are in state 34. Your immediate reward is 3. You have 2 actions"

Robot: "I'll take action 1"

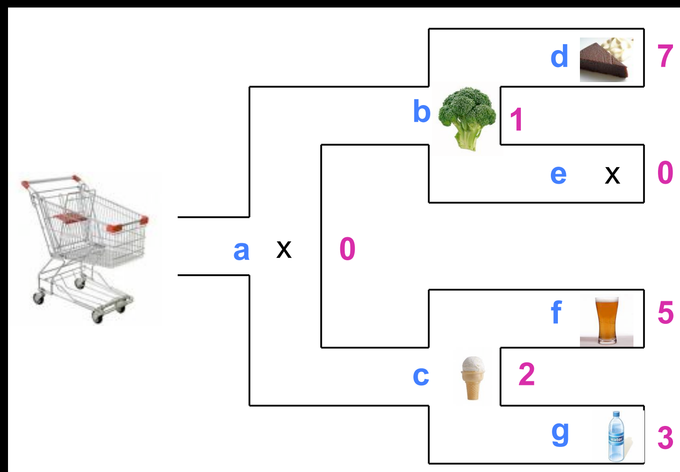
World: "You are in state 77. Your immediate reward is -7. You have 3 actions"

Robot: "I'll take action 3"

The task description requires no memory (*doesn't* mean that the decision maker does not use memory to solve the task!)



what can we compute here?



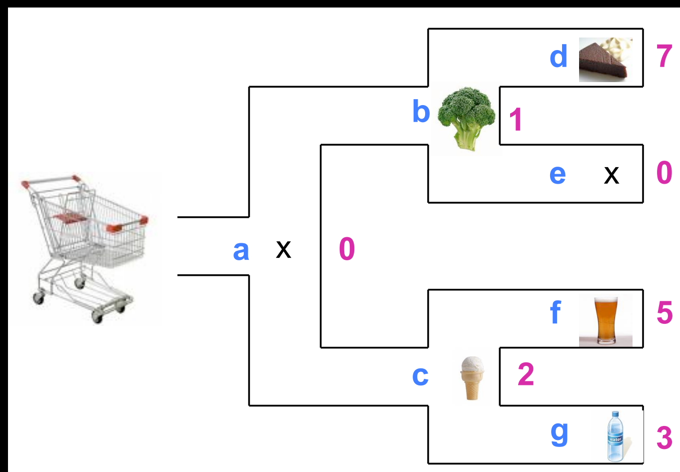
state values: $V(S) = E[\text{sum of future rewards}|S]$
actually: $V^{\pi}(S) = E[\text{sum of future rewards}|\pi,S]$

Key RL idea #1: Bellman's glorious equation

$$V^{\pi}(S) = \sum_a \pi_{s,a} \sum_{s'} T^a_{s\to s'} [R^a_{s\to s'} + V^{\pi}(S')]$$

In a Markov decision process, state values are recursive

but there's more: computing the value of actions



(policy dependent) State-Action values:

Q^{π}(action|state) = E[sum of future rewards|S,a,π]

- Q(left|a) = ? Q(right|a) = ?
- which action is better?

Key RL idea #1 (again): Bellman's glorious equation

$$Q(S,a) = \sum_{S'} T^a_{S \to S'}[R^a_{S \to S'} + V(S')]$$

But.. what if we don't know T, R?

model-free learning: sampling

World: "You are in state 34. Your immediate reward is 3. You have 2 actions"

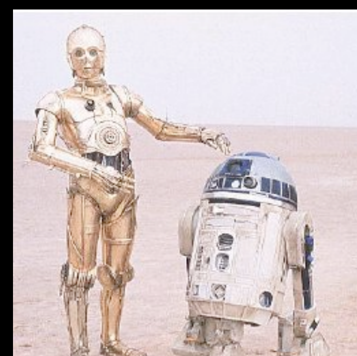
Robot: "I'll take action 1"

World: "You are in state 77. Your immediate reward is -7. You have 3 actions"

Robot: "I'll take action 3"

Take actions according to policy.

Treat experienced rewards and transitions as *samples*



Key RL idea #2: Model-free learning

$$V^{\pi}(S) = \sum_a \pi_{s,a} \sum_{s'} T^a_{s \rightarrow s'} [R^a_{s \rightarrow s'} + V^{\pi}(S')]$$

1. choose initial values $V_0(S)$
2. at time point t and state S_t behave according to π
3. observe S_{t+1} and $r(S_{t+1})$
4. compute prediction error $r(S_{t+1}) + V(S_{t+1}) - V(S_t)$
5. update $V(S_t)$ according to prediction error

learning of long-term values can be done using only local information and without a model of the environment

summary so far

Instrumental learning = learning optimal control

MDPs: class of stylized tasks

In a Markov process long term values can be defined that

- are self consistent (recursively defined)
- can be learned incrementally (dynamic programming)
- can be learned from experience even without a world model

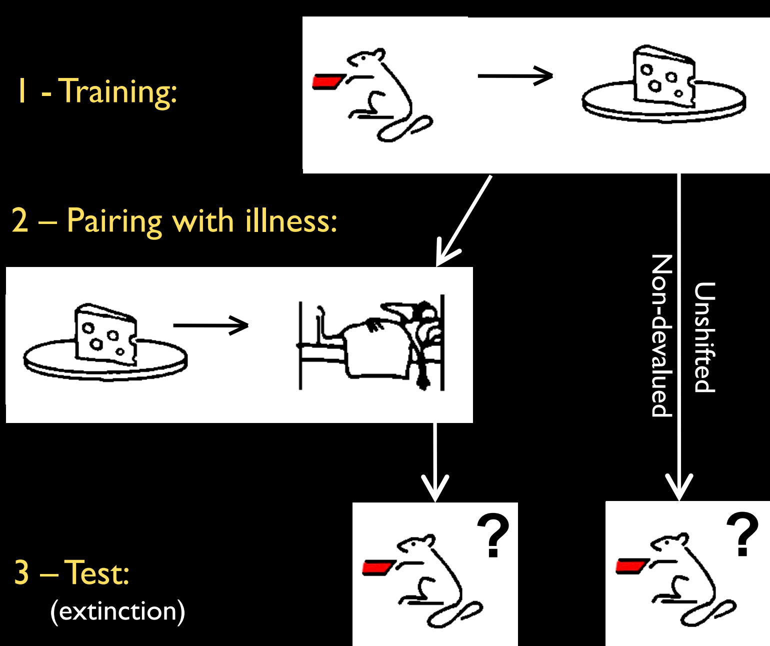
These values are helpful because they can help us improve the policy!

is animal learning model-based or model-free?

1 - Training:

2 – Pairing with illness:

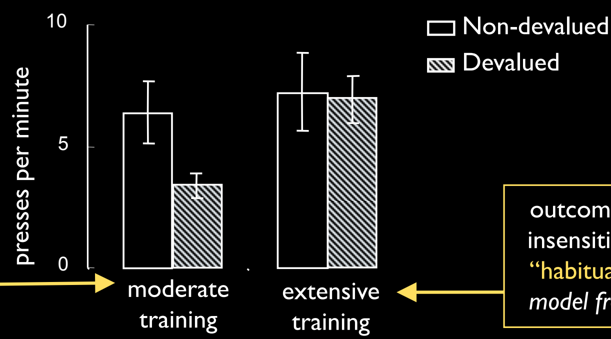
3 – Test:
(extinction)



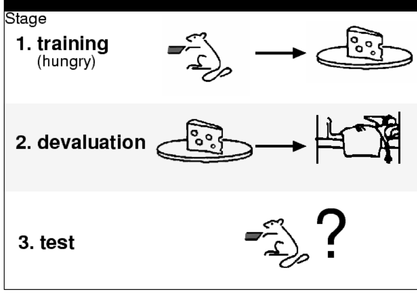
will animals work for food they don't want?

devaluation: results

outcome-sensitive
"goal directed"
model-based



outcome-insensitive
"habitual"
model free

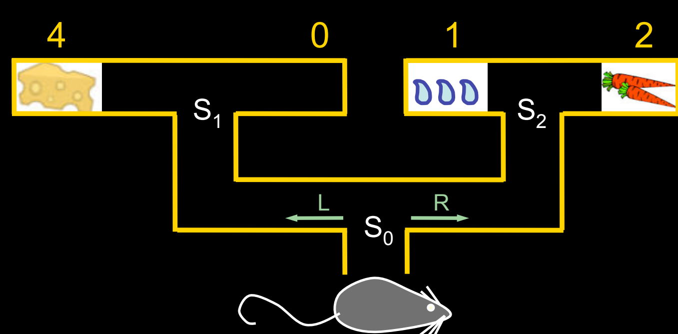


Animals will *sometimes* work for food they don't want!

→ in daily life: actions become automatic (habitual) with repetition

Holland (2004)

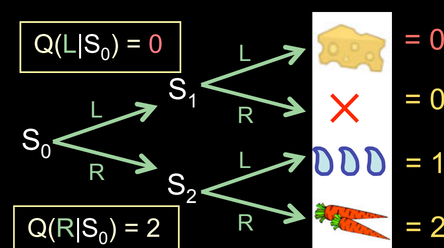
goal-directed actions as model-based reinforcement learning



learn model of task through experience

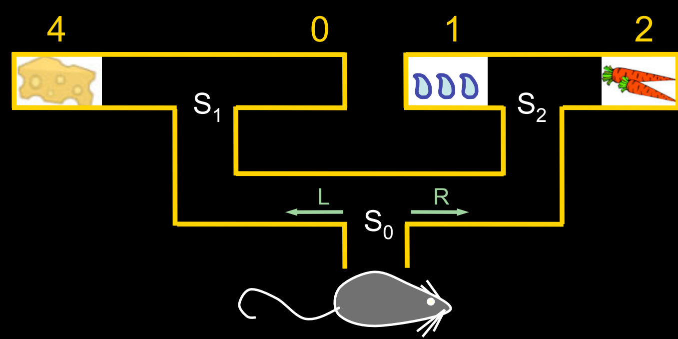
compute action values by "looking ahead" (mental simulation) in the map

computationally costly, but also flexible (immediately sensitive to change)



Daw et al. 2005

habitual actions as model-free reinforcement learning



- Shortcut: store values learn from past experience
  - then simply retrieve them to choose action
- Can learn these from prediction errors
  - incrementally, Rescorla-Wagner/TD learning
  - should depend on dopamine prediction-errors
  - this doesn't require building or searching a model

Stored:

$Q(S_0,L) = 4$
$Q(S_0,R) = 2$

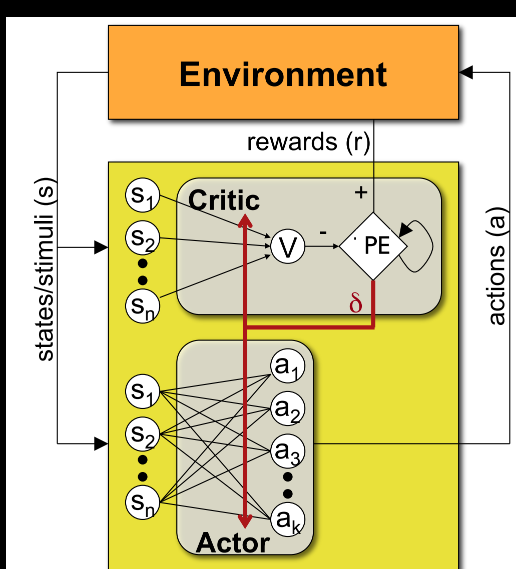
$Q(S_1,L) = 4$
$Q(S_1,R) = 0$

$Q(S_2,L) = 1$
$Q(S_2,R) = 2$

Daw et al. 2005

learning action values from prediction errors: Actor/Critic model

(N.B. skipped this in talk, but I left it here anyway)



Positive prediction error
Things are *better* than expected

→ update value of stimulus/state
→ update policy (probability of action)

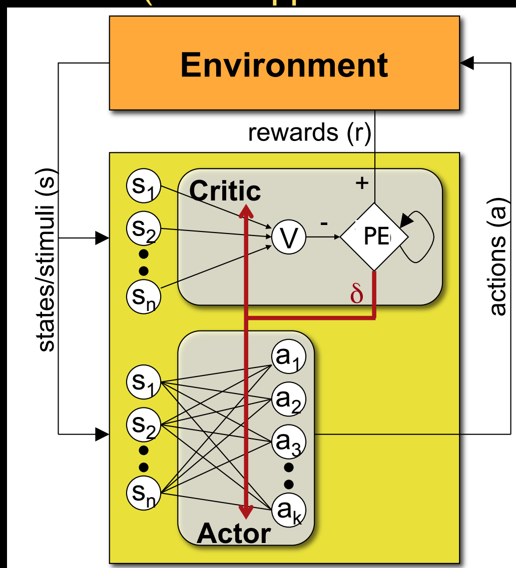
Negative prediction error
Things are *worse* than expected

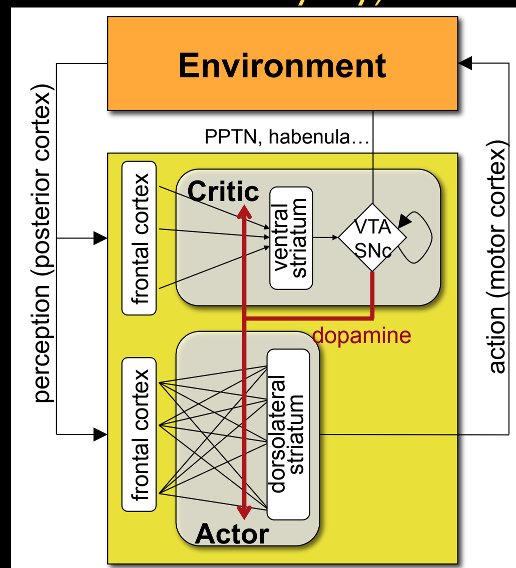
→ update value of state
→ update policy

Sutton (1978), Barto et al. (1983)

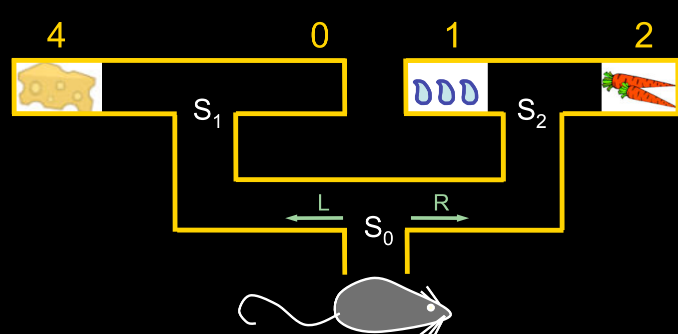
Actor/Critic: neural implementation

(N.B. skipped this in talk, but I left it here anyway)





habitual actions as model-free reinforcement learning



- choosing actions is easy so behavior is quick, reflexive (S-R)
- but needs a lot of experience to learn
- and inflexible, need relearning to adapt to any change (habitual)

Stored:

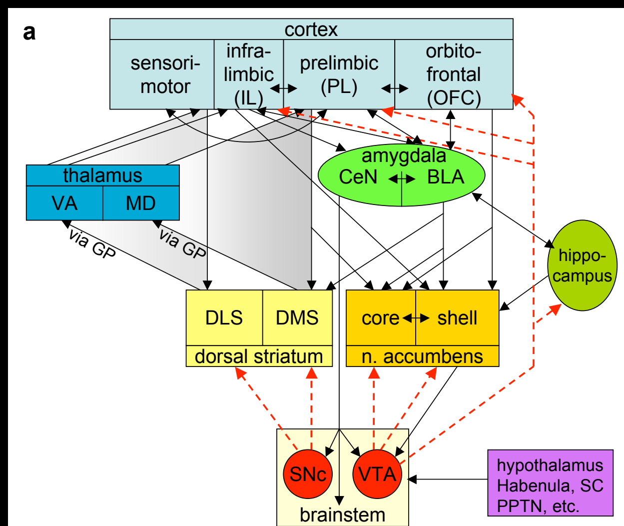
$Q(S_0,L) = 4$
$Q(S_0,R) = 2$

$Q(S_1,L) = 4$
$Q(S_1,R) = 0$

$Q(S_2,L) = 1$
$Q(S_2,R) = 2$

Daw et al. 2005

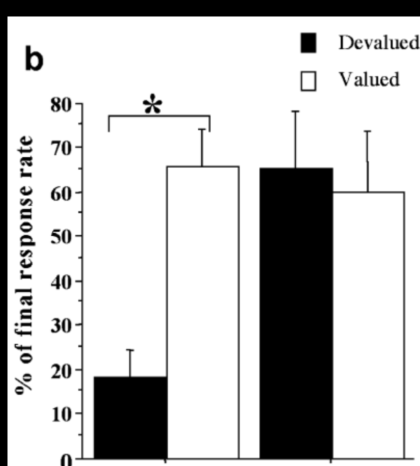
in the basal ganglia: two parallel routes to action selection



habits in the dorsolateral striatum

(N.B. skipped in talk)

overtrained rats



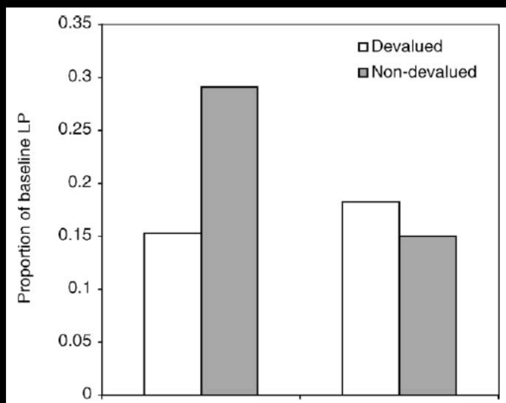
dorsolateral striatum lesion   control (sham lesion)

- ➡ animals with lesions to DLS *never develop habits* despite extensive training
- ➡ also treatments depleting dopamine in DLS
- ➡ also lesions to infralimbic division of PFC (same corticostriatal loop) or VA nucleus of thalamus

Yin, Knowlton, et al. (2004)

infralimbic cortex enables habitual responding (N.B. skipped in talk)

overtrained rats



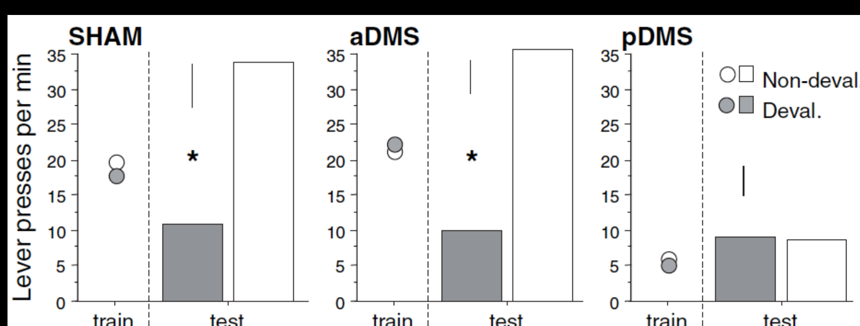
IL PFC inactivation (muscimol)

control

even after habits have been formed, devaluation sensitivity can be *reinstated* by temporary inactivation of IL PFC

Coutureau & Killcross, 2003

dorsomedial striatum necessary for goal-directed behavior (N.B. skipped in talk)

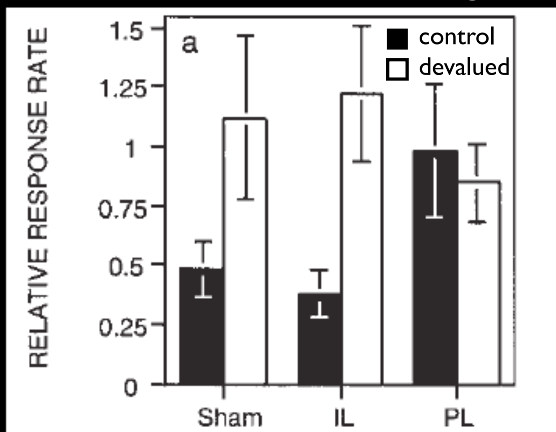


lesions of the posterior DMS (pDMS) cause animals to leverpress *habitually* even with only moderate training

Yin, Ostlund, et al., (2005)

prelimbic cortex also part of the goal-directed loop (N.B. skipped in talk)

moderate training



Prelimbic (PL) PFC lesions cause animals to leverpress *habitually* even with only moderate training

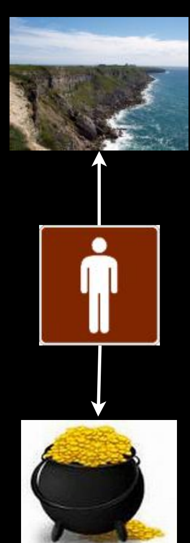
(also dorsomedial PFC and mediodorsal thalamus (same loop))

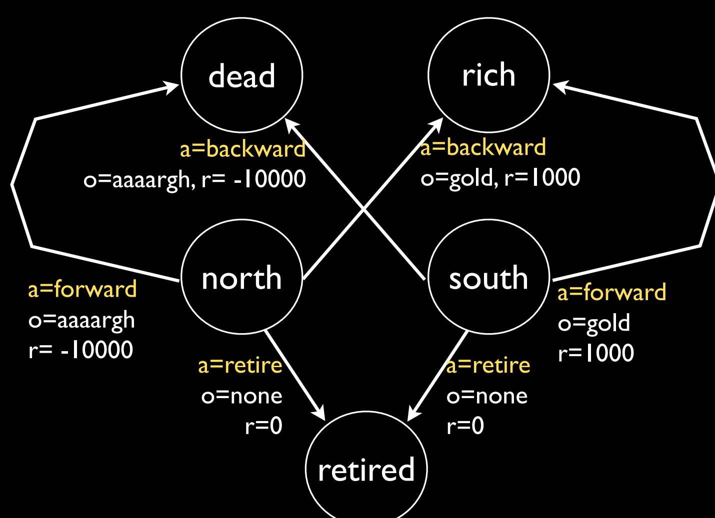
Killcross & Coutureau (2003)

summary so far...

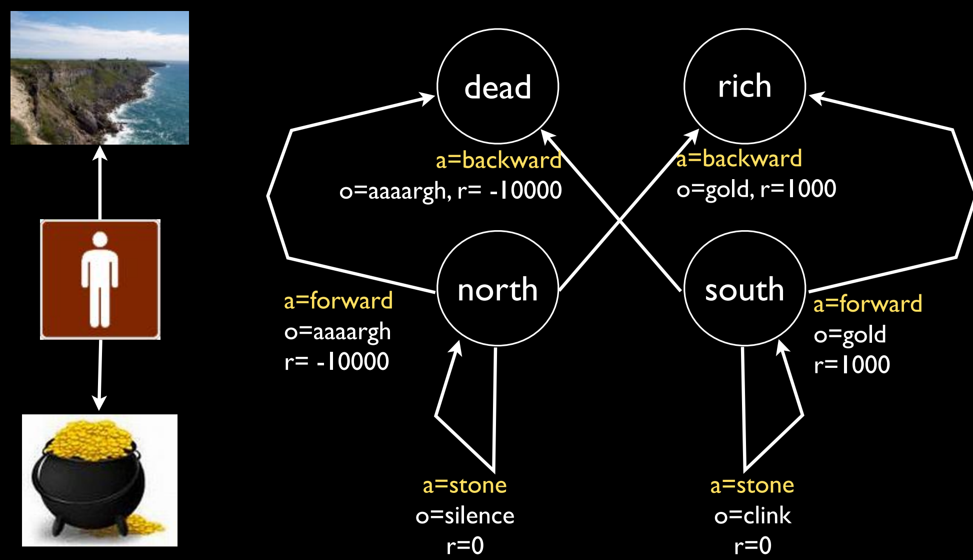
- Behavioral and neural evidence for two parallel decision making systems in the basal ganglia
- One system (IL→DLS→VA thalamus) learns stimulus values using dopamine prediction errors and supports habitual/model-free behavior
- One system (PL→DMS→MD thalamus) seems to use a more flexible "cognitive map" of the task to make decisions, supporting goal directed/model-based behavior

Act 5: between a cliff and a pot of gold (in the dark)

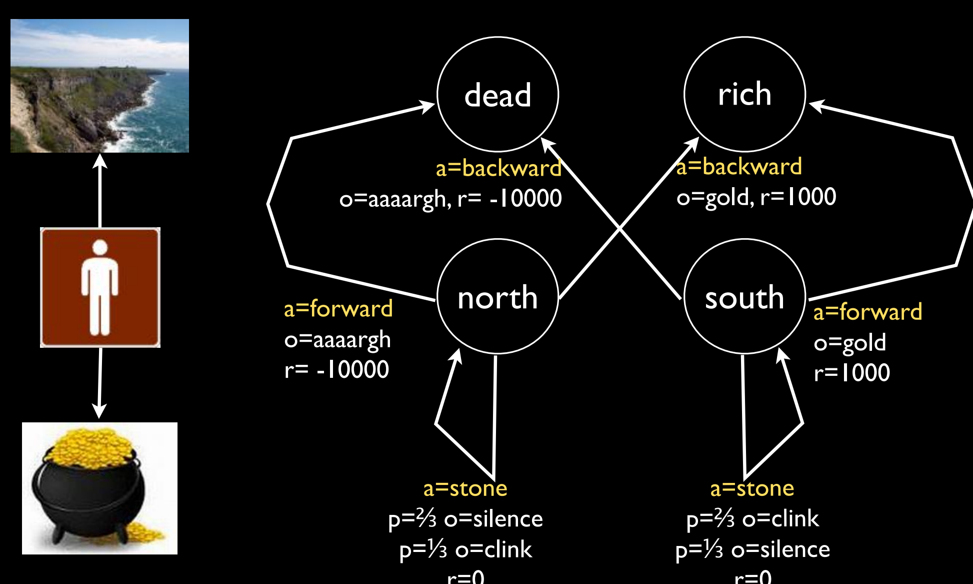




- what is the optimal policy?

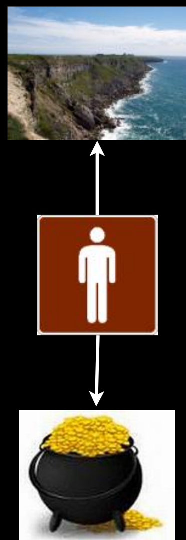


- information gathering action



- what to do in this case?
- integrate multiple observations across time

belief states in a POMDP

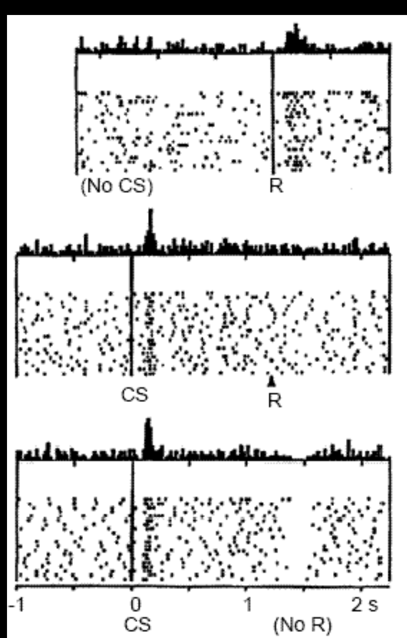


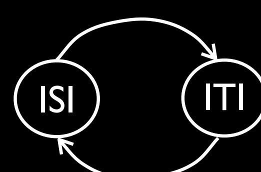
given a model of the environment (transition & observation functions, like previous diagram)

- infer hidden state using observations, model (and Bayes rule)
- this produces distribution over hidden states
  p(north | clink) ∝ p(clink | north) p(north)
- distribution is called "belief state"

- belief states form an MDP and so we can use RL machinery for learning! (Kaelbling et al 1995)

Belief states in the brain?





Schultz, Dayan & Montague 1997

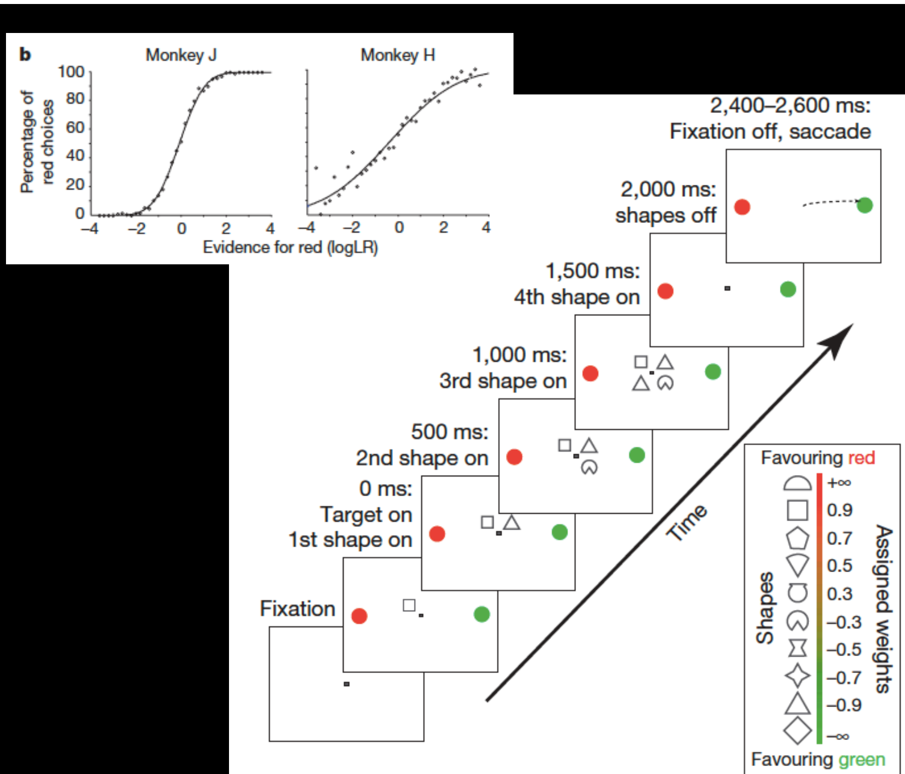
Belief states in the brain?

Vol 447|28 June 2007|doi:10.1038/nature05852 nature

ARTICLES

Probabilistic reasoning by neurons

Tianming Yang[1] & Michael N. Shadlen[1]



summary so far

- Belief states as framework for thinking about real world learning tasks: incorporating uncertainty about current state into RL
- separates inference of state (in perceptual areas?) from learning in basal ganglia (dopamine etc.)
- Note: confusing (or deliberate?) use of 'decision making'

additional reading

- Rescorla & Wagner (1972) - *A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement* - the original chapter that is so well cited (and well written!)
- Sutton & Barto (1990) - *Time derivative models of Pavlovian reinforcement* - shows step by step why TD learning is a suitable rule for modeling classical conditioning
- Rescorla (1988) - *Pavlovian conditioning: its not what you think it is* - a manifesto for studying big questions using simple behavior
- Niv & Schoenbaum (2008) - *Dialogues on prediction errors* - a guide for the perplexed
- Hare et al. (2008) - *Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors* - an elegant and careful study of values and prediction errors in humans
- Niv (2009) - *Reinforcement learning in the brain* - summary of what I talked about
- Dijksterhuis et al. (2006) - *On making the right choice: the "deliberation without attention" effect* - advice for decision making in real life