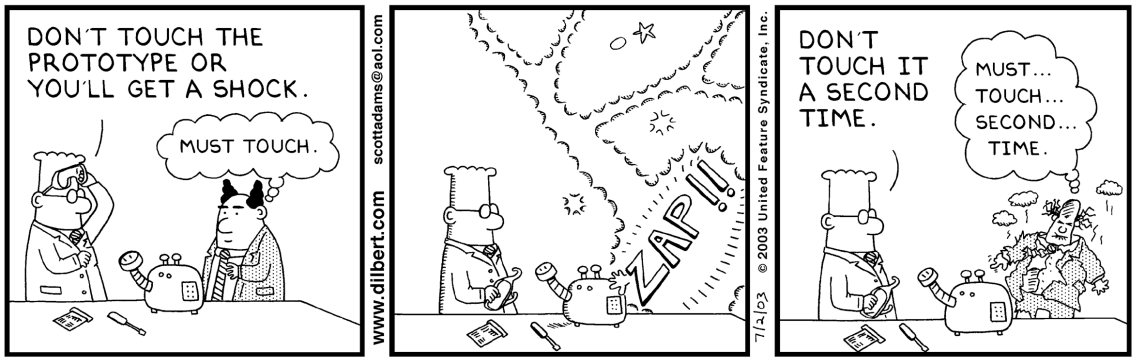
Instrumental Conditioning II: Modeling Action Selection

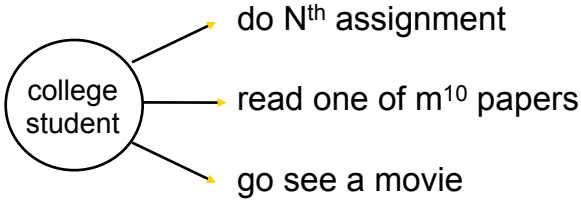


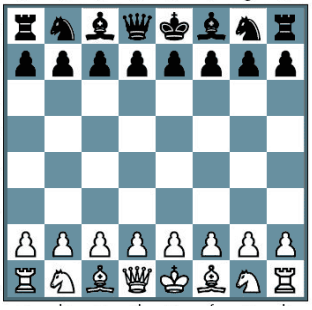
PSY/NEU338: Animal learning and decision making:
Psychological, computational and neural perspectives

how to model instrumental conditioning?

Marr's levels, again:

- The problem: find the best behavioral policy (what to do in what situation)
  best in terms of?
- cashing this out: the credit assignment problem
- Algorithms: Reinforcement learning (RL)

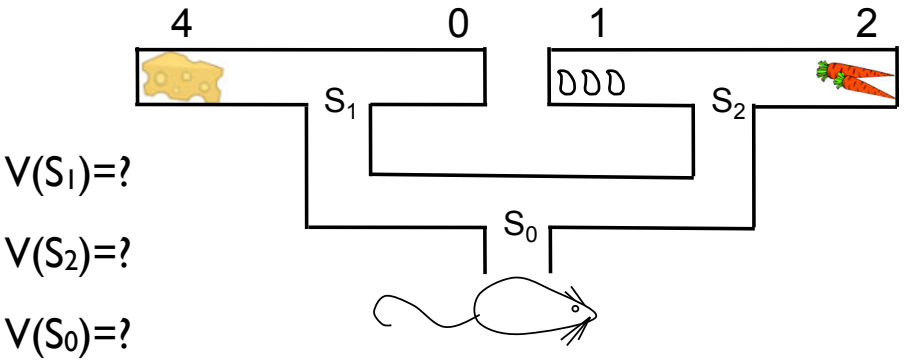




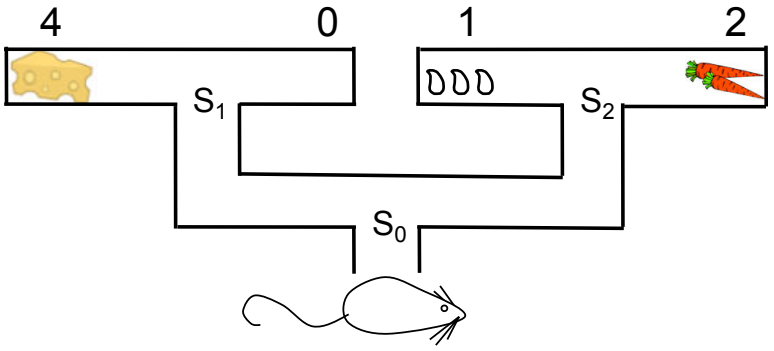
how to model instrumental conditioning?

Marr's levels, again:

- The problem: find the best behavioral policy (what to do in what situation)
- An algorithm: Actor/Critic learning



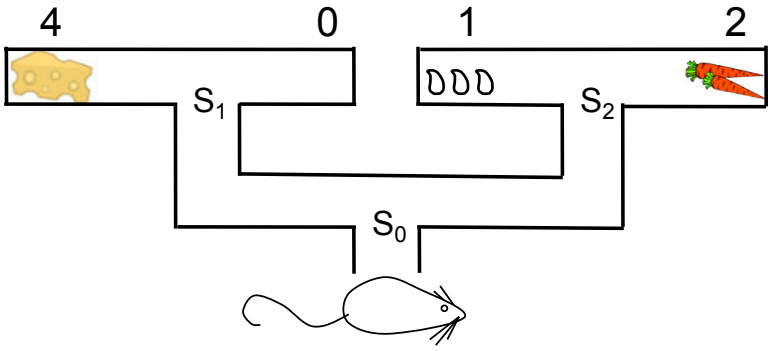
modeling instrumental conditioning



What is the value of S_2 under a random policy?

A) 2
B) 1.5
C) 1
D) It depends

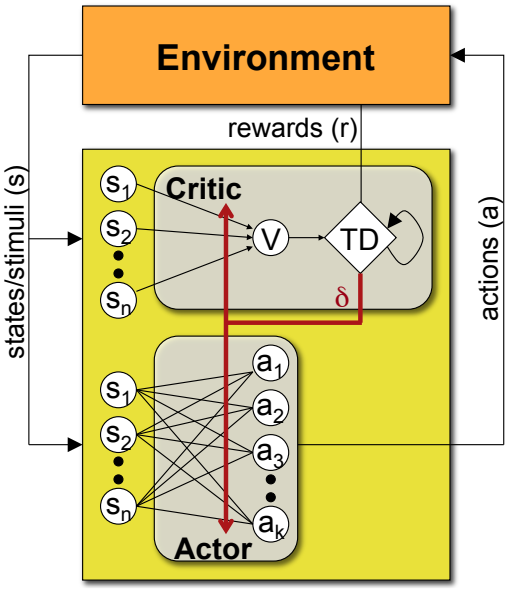
modeling instrumental conditioning



What will happen when the rat goes right at S_2?

A) he will experience a negative prediction error
B) he will not have a prediction error as everything is predictable
C) he will experience a positive prediction error
D) It depends

Actor/Critic: a player and a coach



Positive prediction error
Things are *better* than expected

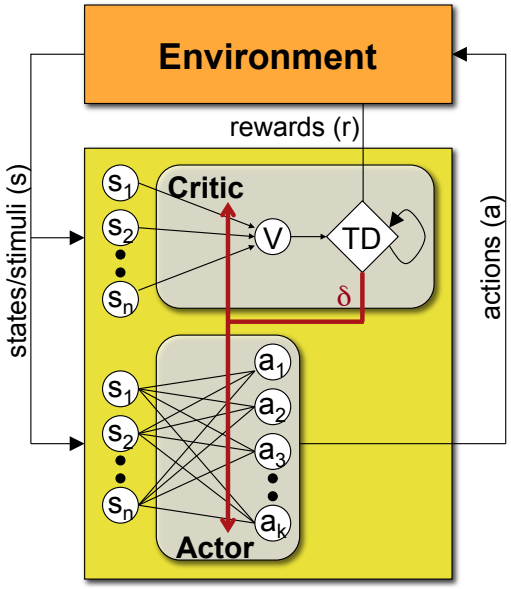
→ update value of state
$V(S) \leftarrow V(S) + \eta_v \delta$
→ update policy (probability of action)
$\pi(S,a) \leftarrow \pi(S,a) + \eta_\pi \delta$

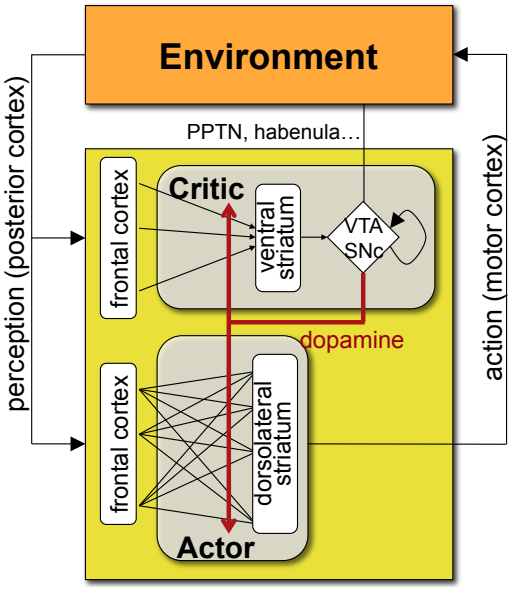
Negative prediction error
Things are *worse* than expected

→ update value of state
→ update policy

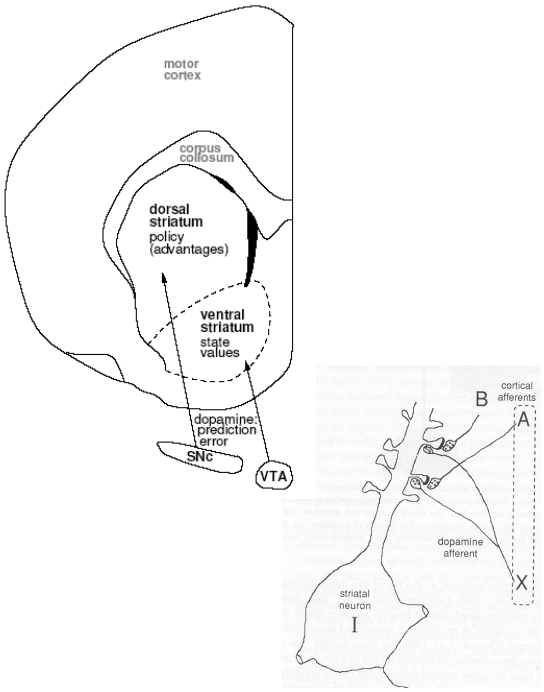
Sutton (1978), Barto et al. (1983)

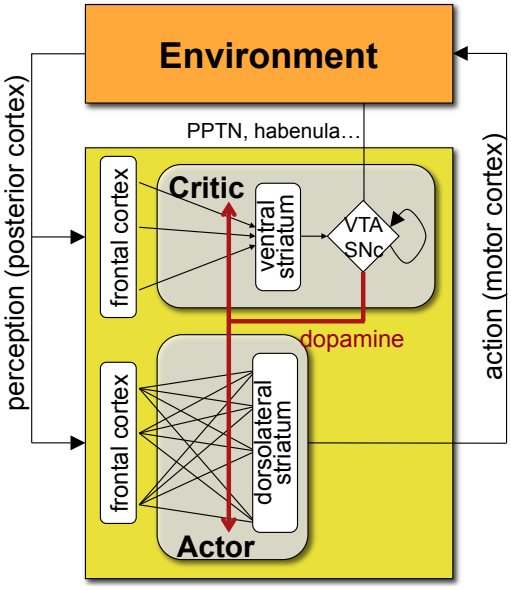
Actor/Critic: neural implementation





Actor/Critic: neural implementation





Actor/Critic: a player and a coach

prediction errors can act as surrogate rewards
(solve the credit assignment problem)

$\delta_t = r_t + V(S_{t+1}) - V(S_t)$

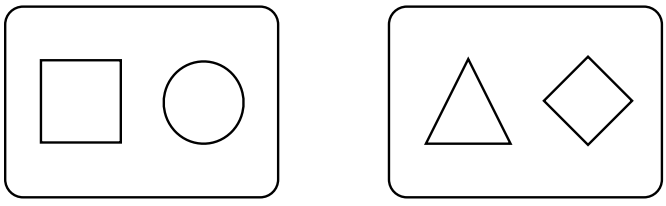
in absence of reward:

$\delta_t = V(S_{t+1}) - V(S_t)$

(can "coach" Actor)

evidence from fMRI: O'Doherty et al 2004

- in each trial two stimuli: one with high probability of reward (60%) and one with low (30%) probability of reward



- subjects choose one stimulus and receive outcome
- two types of trials: juice rewards and neutral rewards
- why was the experiment designed this way (hint: think of prediction errors)
- in another condition: no action selection, subjects only indicate the side the 'computer' has selected (Pavlovian conditioning)

pop quiz

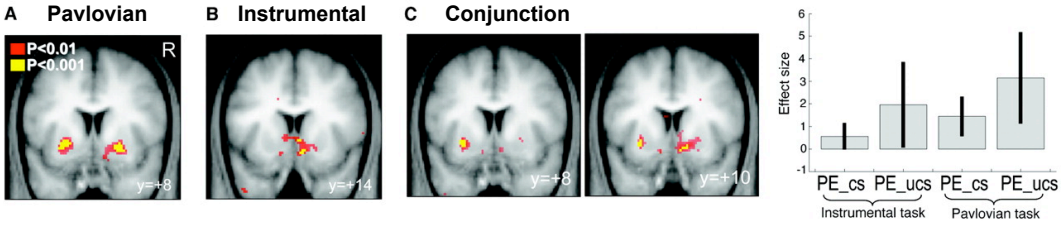
- what does fMRI BOLD response measure?
  A. the level of oxygen in the blood in an area
  B. the amount of activation in nearby neurons
  C. the amount of input to nearby neurons
  D. the concentration of water molecules in an area

pop quiz

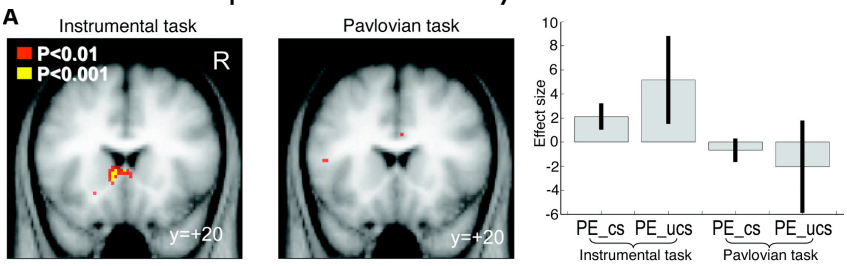
- where are correlates of prediction errors commonly seen in the brain in fMRI?
  A. the VTA and SNc, where dopamine neurons are
  B. all over the brain as dopamine neurons have brain-wide projections
  C. the ventral striatum, a major recipient of dopamine projections
  D. it depends on the study

evidence from fMRI: O'Doherty et al 2004

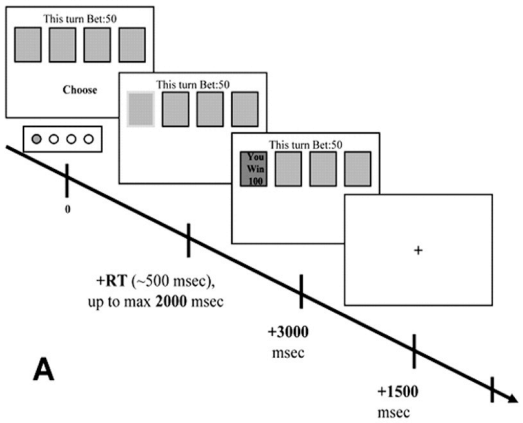
ventral striatum: correlated with prediction error in both conditions

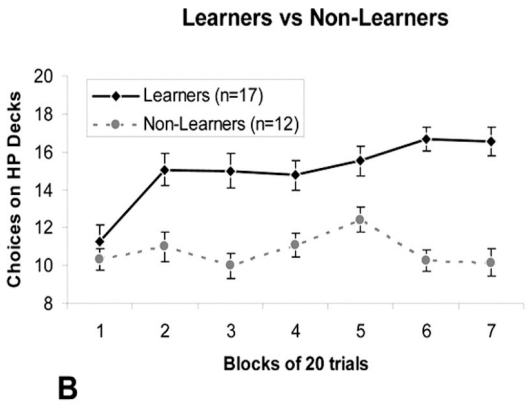


Dorsal striatum: prediction error only in instrumental task



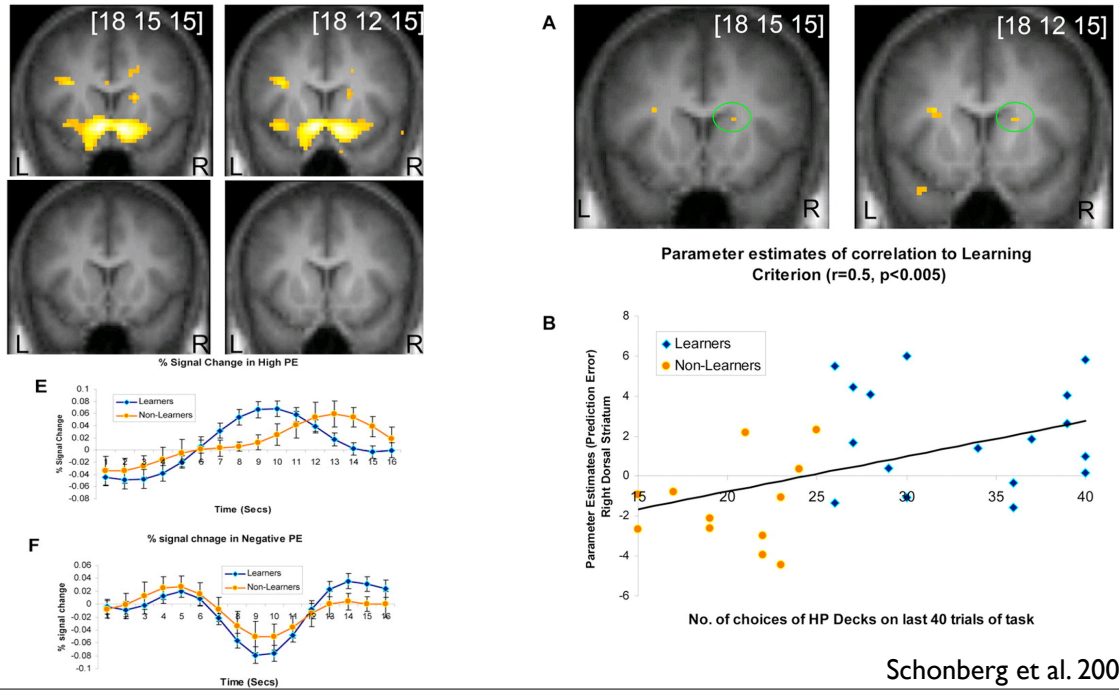
do prediction errors really influence learning?





Schonberg et al. 2007

do prediction errors really influence learning?



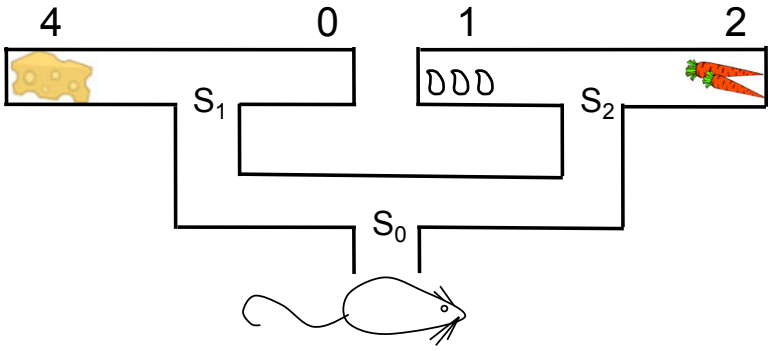
Schonberg et al. 2007

summary so far...

- Modeling instrumental conditioning: reinforcement learning uses predictive values to inform choice
- Computationally, prediction errors save the day again: can solve the credit assignment problem
- In the brain: evidence for division between prediction learning and policy learning (Actor/Critic)

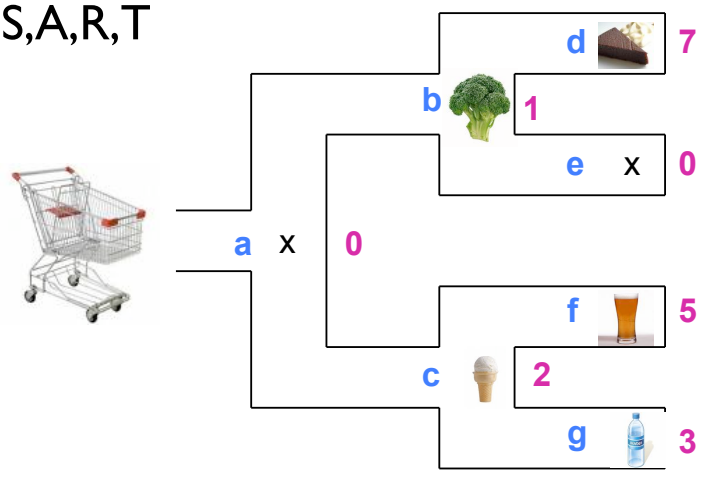
how to model instrumental conditioning?

- The problem: find the best behavioral policy (what to do in what situation)
- A bit more formally: Markov decision process (S,A,R,T)



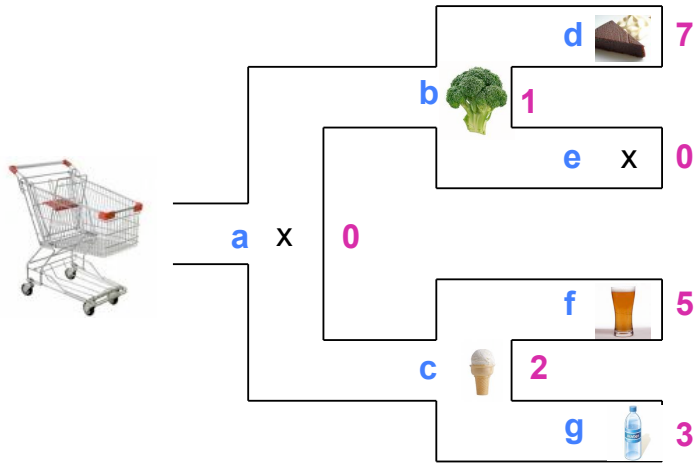
more formally: MDPs

S,A,R,T



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

The Markov property



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

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!)