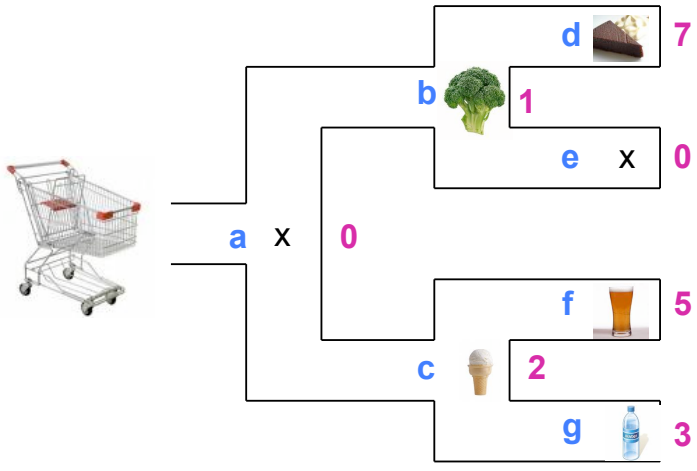
Instrumental Conditioning IV: actor/critic in our brain?

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

Markov Decision Processes



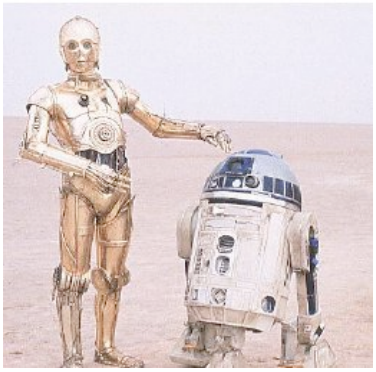
- The idea: given the current situation, history does not matter
- $P(S_{t+1}|S_1,S_2,\ldots,S_t,a_1,a_2,\ldots,a_t) = P(S_{t+1}|S_t,a_t)$
- $P(r_t|S_1,S_2,\ldots,S_t,a_1,a_2,\ldots,a_t) = P(r_t|S_t,a_t)$

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!)



learning a policy for MDPs

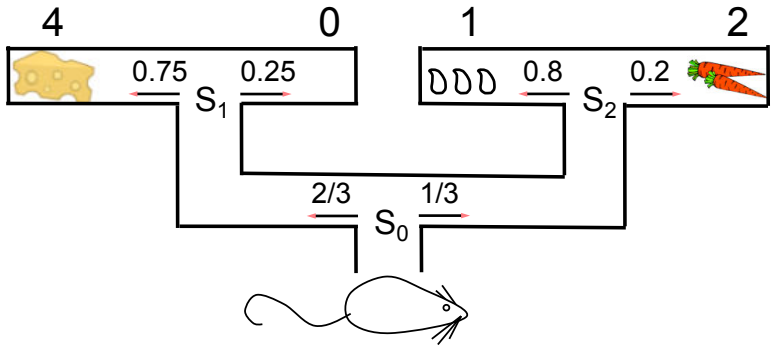
(policy dependent) State values:

$V^{\pi}(S)$ = E[sum of future rewards|S,π]

$V(S_0)$ = ?

A. 4.2
B. 2.4
C. 2
D. 2.8

can this value help choose actions?

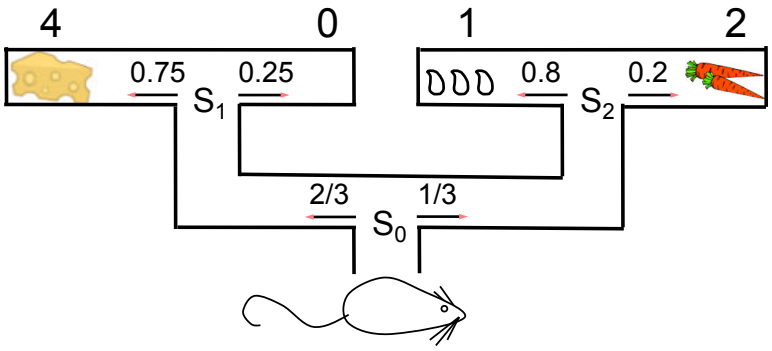


computing the value of actions

(policy dependent) State-Action values:

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

- Q(L | S_0) = ?
- Q(R | S_0) = ?
- which action is better?



learning optimal policies

Optimal policy: in terms of future rewards; a policy that obtains the largest possible amount of reward overall

How to learn an optimal policy?

OPTION 1: "batch" algorithm:

- behave according to current policy
- estimate Q values based on experience
- improve policy based on these Q values
- repeat

learning optimal policies

Optimal policy: in terms of future rewards; a policy that obtains the largest possible amount of reward overall

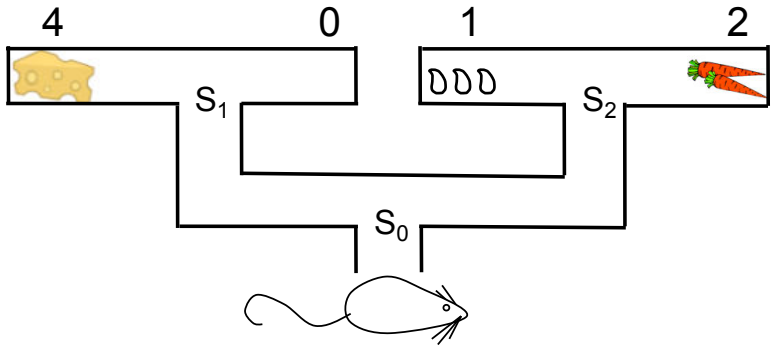
How to learn an optimal policy?

OPTION 2: "online" algorithm:

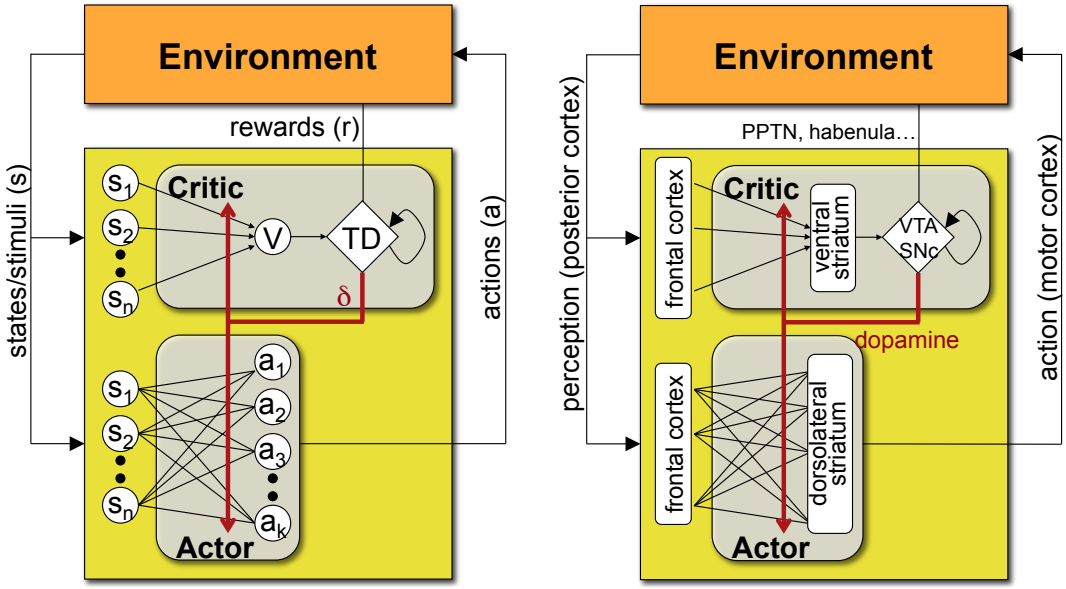
- behave according to current Q values
- calculate prediction error after every action
- update Q value based on prediction error
- repeat

SARSA versus Q-learning

- $\delta_{t+1} = r_t + Q(a_{t+1} \mid S_{t+1}) - Q(a_t \mid S_t)$ (SARSA)
  or
  $\delta_{t+1} = r_t + \max_a Q(a \mid S_{t+1}) - Q(a_t \mid S_t)$ (Q-learning)
- $Q(a_t \mid S_t)^{new} = Q(a_t \mid S_t)^{old} + \eta\delta_{t+1}$
- choose actions according to *softmax*: $p(a|S) \propto e^{\beta Q(S,a)}$



compare to:



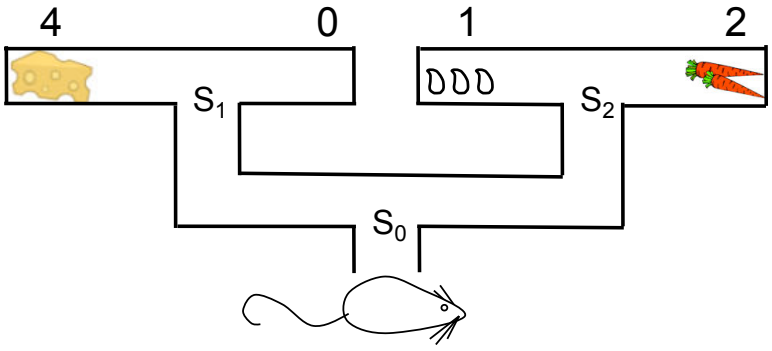
summary so far...

- Modeling instrumental conditioning (action selection): several models have been proposed (Q learning, SARSA, Actor/Critic)
- in all cases reinforcement learning uses predictive values to inform choice
- remember: all this works only in MDPs (but many problems can be represented as MDPs or approximated by MDPs)

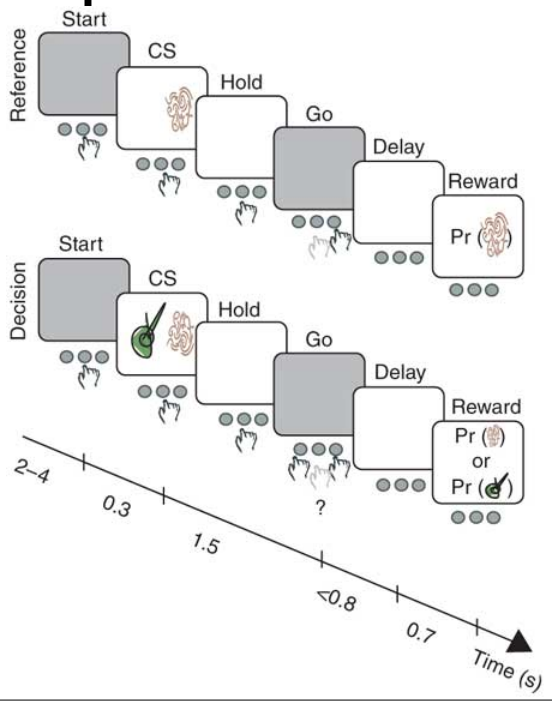
does the brain really use actor/critic learning?

how can we tell the models apart?

- $\delta_{t+1} = r_t + Q(a_{t+1} \mid S_{t+1}) - Q(a_t \mid S_t)$ (SARSA)
- $\delta_{t+1} = r_t + \max_a Q(a \mid S_{t+1}) - Q(a_t \mid S_t)$ (Q-learning)
- $\delta_{t+1} = r_t + V(S_{t+1}) - V(S_t)$ (Actor/Critic)

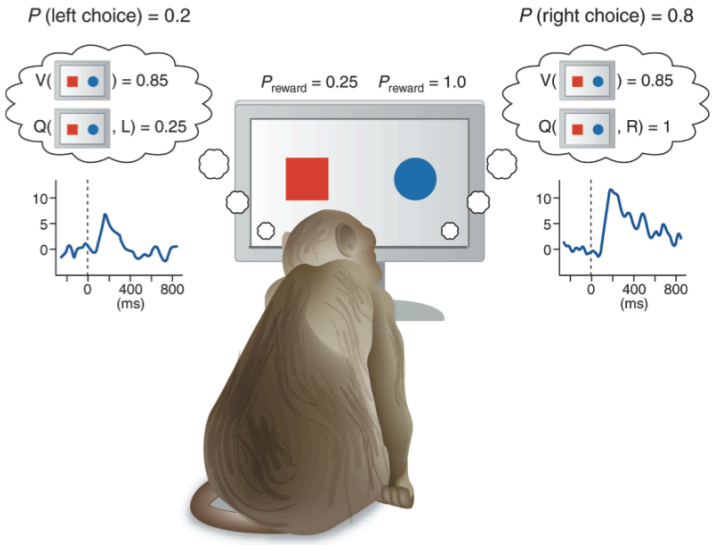


what do dopamine prediction errors represent at trial onset?



Morris et al. 2006 13

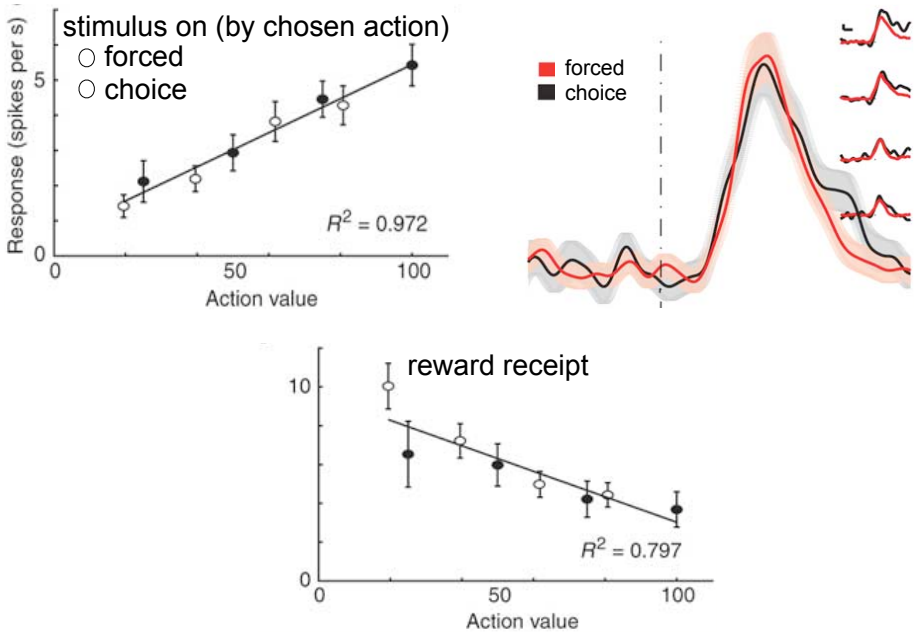
what do dopamine prediction errors represent at trial onset?



Katie Ris

Morris et al. 2006 14

what do dopamine prediction errors represent at trial onset?



Morris et al. 2006

summary so far...

- In the brain: evidence for division between prediction learning and policy learning (Actor/Critic)
- But: nature of prediction errors themselves suggests otherwise
- Not only do the models inform us about the brain, but the brain can inform us about the models!
- But… what about modeling free operant behavior?