Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control

Nathaniel D. Daw, Yael Niv, & Peter Dayan (2005)

I. Dual Systems?
    A. Existing Hypothesis: There are two distinct and parallel pathways for action selection and they are served by either the prefrontal cortex or the striatum and its dopaminergic (DA) afferents.
    B. Convention states:
        i. Dorsolateral striatum + DA afferents = habitual or reflexive control
        ii. Prefrontal cortex = reflective or cognitive action planning
    C. 2 BIG Questions:
        i. Why >1 controlling system?
        ii. What happens when they conflict?
    D. Two major classes of reinforcement learning
        i. "Model-free"
            1. "Cache" system: creating associations between an action or situation with it's long term value
            2. Computationally simple BUT inflexible
            3. Inflexibility causes this system to be insensitive to devaluation of an outcome
            4. Associated with DA neuron activity and striatal projections (e.g. temporal-difference learning)
        ii. "Model-based"
            1. "Tree search" system: anticipating the immediate outcome of each action in a sequence to develop predictions about the long term value of an action or situation
            2. Costly in terms of time, memory, and potential error BUT flexible
            3. Flexibility arises from the ability to make short term predictions about consequences of actions or situations, which allows it to be robust in the face of changes in circumstance
            4. Associated with prefrontal cortex
        iii. Each of these systems makes their own approximations to overcome statistical and computational challenges which results in differential accuracy profiles
    E. **Proposal:** The existence of two systems is justified and the conflict resolution explained because of the differential accuracy achieved with each of these systems (i.e. model-based vs. model-free)

II. Results
    A. Post-training reinforcer devaluation
        i. *Typical reinforcement learning paradigm in a rat*: Train a hungry rat to perform a series of actions to obtain a reward (i.e. food pellet)
        ii. *Devaluing a reinforcer*: reduce the value of the reward prior to a learning trial
            1. *In rats*: feed them the reinforcing food before the trial or pair the food with illness, so the rat no longer desires to the reinforcing food.
        iii. Hypothesis regarding behavior after outcome devaluation by system type:
            1. "Cache" system: continue to perform learned actions
                a. By definition this "habitual" behavior does not take into account outcomes
                b. Rat thinks, "Press lever = good"

2. "Tree" system: do not perform learned actions
    a. This "goal-directed" system allows for prediction of the immediate outcome of an action and thereby the long term outcome of a series of actions
    b. Rat thinks, "Press lever = get food → Don't want food → Don't press lever."

iv. Evidence from Behavioral Experiments
1. Rats exhibit differential behavior demonstrating both profiles of devaluation in varying circumstances
    a. Moderately trained lever presses = devaluation sensitive (tree)
    b. Extensively trained lever presses = devaluation insensitive (cache)
2. Block DA input to dorsolateral areas of the striatum
    a. Preserves learning
    b. Over-learned lever pressing = devaluation sensitive (tree)
3. Two factors interfere with transition to caching with extensive training
    a. Complexity of action choice: increased complexity = devaluation sensitivity persists (tree)
    b. Proximity of action to reward: closer proximity = devaluation sensitivity persists (tree) [note: evidence not as strong for this]
4. Lesions to a variety of structures can interrupt tree-search process→ eliminate devaluation sensitivity for even moderately trained behaviors

B. Theory Sketch

i. Separate and parallel reinforcement learners
1. Lesion studies demonstrate that each system can work for the other even in a situation where that system is not expected to be the dominant one

ii. Optimal control = maximizing the probability of achieving a desired outcome
1. Value function: Calculating the value of taking each action at each state accounting for the probability of a reward later being earned, when starting from a particular action in a particular state.

iii. A controller achieves dominance in determining value of an action based on the amount of uncertainty of the value each controller calculates
1. The value provided by the controller with the least uncertainty "wins"
2. Probability of choosing an action is proportional to the "winning" value
3. Uncertainty exists in each system because both begin ignorant with little experience and as they gain experience the task and thereby long term values can change

iv. Tree search system
1. Uses experience to estimate state transition and rewards (the structure of the trees)
2. Iterative search through trees to determine long term reward probability estimates
3. Computationally demanding; increasing noise with each search step

v. Cache system
1. Estimates long term values from experience- no tree construction
2. Bootstrapping
3. Calculation straight forward; little computational noise

C. Simulations

i. Quantitative results consistent with qualitative expectations

ii. Even with matched initial uncertainty, model- based (tree-search) learning was more certain early in training
iii. Past observations gradually have less bearing on present value estimates because the systems expect that action values may change
iv. Asymptotic nature of uncertainty has a greater effect on cache system
v. Asymptotic nature of uncertainty driving the effects of complexity and proximity on devaluation

III. Discussion
- A. They claim incorporating uncertainty into their systems has allowed them to give a unifying account of the literature on controller competition
    - i. Both systems are pursuing rational results but there are situations in which it is more appropriate to use one controller over the other
    - ii. They claim theories that conceptualize learning as one system get stuck on explaining the lesion studies
- B. Neural substrates
    - i. Authors acknowledge that for simplicity they have assumed these systems to be separate, however, multiple sources suggest that the interaction of the two systems is a more likely scenario
        1. Biological evidence that the neural substrates of the two systems intertwine
        2. Computationally it is more efficient to use a combination of both systems
    - ii. They propose viewing the competition between model-based and model-free control as between dorsomedial and dorsolateral corticostriatal loops
    - iii. Limited evidence for the uncertainty-based arbitration; a few suggestions to date
        1. Cholinergic and noradrenergic neuromodulation involved in uncertainty
        2. Arbitration candidates:
            - a. Infralimbic cortex
            - b. Anterior cingulate cortex
- C. Experimental Considerations
    - i. Neuronal recordings
        1. Recent evidence could support either striatal or prefrontal control in monkeys performing an over-learned associative learning task with reversals
        2. Devaluation challenge or changing task circumstances could help sort this out
        3. May also require a better defined neural organization for the tree search system to tease this apart
    - ii. Other considerations:
        1. Increase cognitive demand
        2. Introduce unexpected changes in task contingencies
        3. Change task structure in subtle ways
        4. Study ‘Pavlovian’ association tasks and ‘conditioned reinforcement’ tasks with a view to demonstrating interaction among the two systems; use lesion studies
        5. Casts ‘incentive learning’ in a new light; less need for past experience in model-based system

IV. Methods
- A. Background: They modeled the tasks with Markov decision processes (MDPs)
    - i. Agent started without knowing exact MDP (uncertainty)

ii. MDPs did not have static scalar utilities (due to devaluation treatments which changed some outcome utilities)

iii. Assumption: rewards were binary; probability of reward in terminal state was 1

B. Formal model

i. *State-action value function (Q)*: the expected probability that reward will be delivered if the agent takes a particular action in a particular state and continues to choose optimally from there

ii. Q is derived by calculation of both transition state value and probability of reward delivery

iii. This model tracks uncertainty whereas standard reinforcement models do not

1. Bayesian version used which tracks a posterior distribution of the state-action value function in addition to the expected value

a. Bayesian tree-search system calculates the posterior distribution over the MDP based on prior experience.

b. Bayesian caching system calculates the posterior distribution over Q_{cache} for each action and state and updates this as it encounters subsequent states.

2. Arbitration between the two systems based on variance (smaller variance wins!)

iv. Note: Posterior uncertainty describes how uncertain the probability of reward is NOT the inherent randomness of the reward delivery, i.e. one can precisely know that reward delivery will happen with 50% probability.

V. Supplementary Methods (courtesy of Matt)

A. Eq. 1: Bellman Equation

i. Action values in each state are defined in terms of values of subsequent states

ii. Basis for all of reinforcement learning and (more generally) control theory.

B. Model-based learning uses data more efficiently than caching

i. Caching: knowledge only propagates backward one step at a time, e.g. Agent must learn about states/actions in a series of behaviors before it can know anything about earlier states/actions in the series

ii. Model-based learning: learns each transition separately then it combines this knowledge to make a prediction about value (and it's probability distribution in this case) which allows a decision to be made

C. Computational noise in the model-based system is implemented via the nu parameter (based on pruning)

i. Without this assumption, the model-based system would win all the time

D. Competition between the two systems is down to a tradeoff between:

i. construction of a model uses data more efficiently than caching (increasing the certainty of model-based estimates)

ii. making inferences from the model requires computational shortcuts (i.e. pruning) that reduce the certainty of model-based estimates

Discussion Questions:

1) Is it possible to conceptualize this evidence from the perspective of one system? If so, how might that single system model account for the lesion data?

2) This article correlates the two systems with different areas of the brain. How might the implementation of these models work mechanistically in the brain?
3) How might these systems map on to Sloman's associative system and rule-based system?