Striatal and Prefrontal control of sequential decisions: a probabilistic theory based on reinforcement learning.

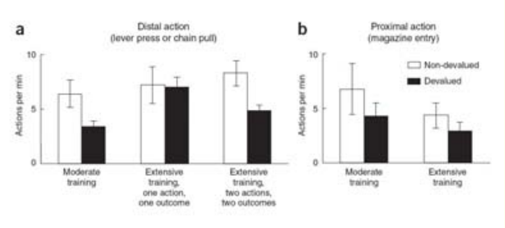
"Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control."
Daw, Niv, and Dayan, Dec 2005, *Nature NS*.

Ray Luo | Feb 2, 2006

Behavior as sequence of actions.

- Dopaminergic dorsolateral striatal system.
  - Reflexive, habitual.
  - Insensitive to devaluation of reward.
  - Heavily trained, distal to reward.
  - "Pavlovian conditioned."
- Prefrontal cortical system.
  - Planned, goal-directed.
  - Sensitive to devaluation of reward.
  - Complex tasks, proximal to reward.
  - "Instrumental conditioned."

Reward devaluation experiments.



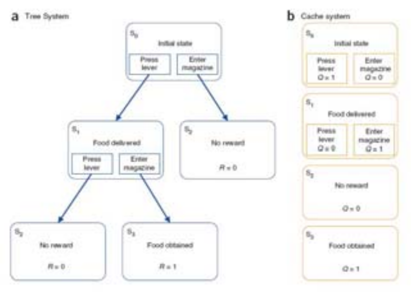
"Model"-based methods.

- State space reinforcement learning in a Markov decision process (for T).
- Adaptive dynamic programming:
  - Value iteration, policy iteration, transition model T(a, s, s') from environment.
  - Bellman $U(s) = R(s) + \max_a \Sigma_{s'} T(a, s, s') U(s')$.
- Exploration vs. exploitation.
- Tree search on the fly.
- "Prefrontal system."

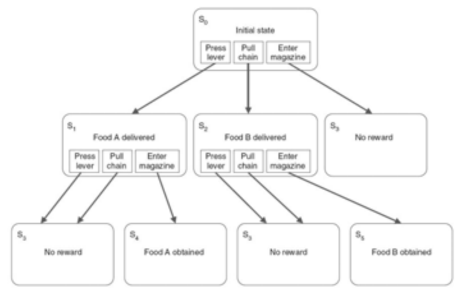
Model-"free" methods.

- Temporal-difference learning, going s->s':
  - $U(s) = U(s) + \alpha(R(s) + U(s') – U(s))$.
- Q-learning, going s->s':
  - $U(s) = \max_a Q(a, s)$.
  - $Q(a, s) = R(s) + \Sigma_{s'} T(a, s, s') \max_{a'} Q(a', s')$.
- Local effect of experience, but efficient.
- Caching / bootstrapping.
- "Striatal system."

One action, one outcome model.



2 actions, 2 rewards, tree search.



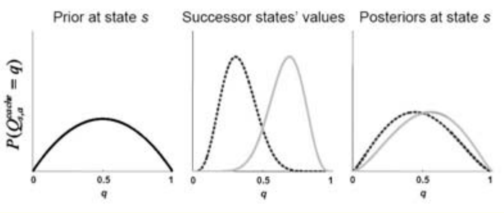
Where are the uncertainties?

- Probability distribution (beta instead of normal gamma) of value function for each state action pair.
- Bayesian Q learning for model- free algorithm.
  - Myopic VPI selection with mixture Q-update.
- Modeled based “noisy” Bayesian tree search for model- based algorithm.
  - Value distributions, Dirichlet successor priors.
- See supplement, Dearden 98, Mannor 04.

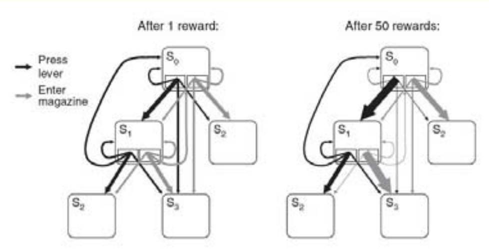
(Dubious if not false) assumptions.

- Brain switches between controller modules for the same action in different situations (Substitute: Pavlovian- >operant transfer?).
- Brain always makes the optimal decision in order to minimize uncertainty.
- Reinforcement learning system that doesn’t “learn,” i.e. the controllers are in place.
- Brain does statistical inference without us knowing about it.

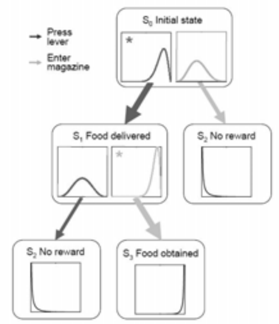
Q-value distribution updating.



Evolution of model by simulation.



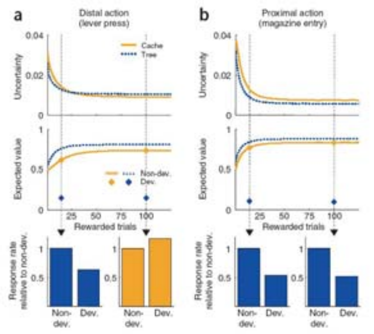
Tree search evolution in detail.



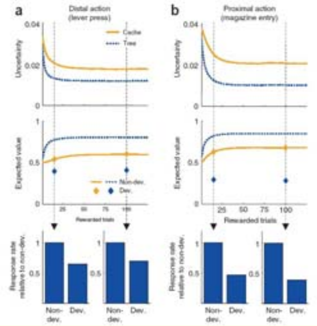
Uncertainties in action.

- Asymptotic value distribution variances greater for model-free method.
- Propagation of uncertainties up to distal level makes variances greater in model-based tree search ("nudge" more than once).
- More actions, more outcomes imply less data, so efficient use of data becomes key.
- Model-free method independent of devaluation.

Simulation of one action paradigm.



Simulation of two action paradigm.



Discussion 1.

- Cache system as limbic influence?
- Medial vs. lateral corticostriatal loops?
- Two types of exploration: explore unknown states and evaluate unevaluated states.
- Novelty reward favors exploration?
- Arbitration via infralimbic cortex and ACC?
- "Fast" striatum trains "slow" prefrontal cortex with reversals favoring PFC (Pasupathy)?

Discussion 2.

- Caching for "fan-out," search for "fan-in."
- Motivational gate for unconditioned stimulus?
- Instrumental outcome in new motivational state is key for "incentive" learning.
- Advantage as difference between Q value and (Pavlovian) state value: modifies reward error signals as found in (Bayer & Glimcher).
- Why model? Is the incentive associated with computational model reflected in tree search?