The Science of Learning and Memory

Lecture 6

Todd M. Gureckis
Department of Psychology
New York University
todd.gureckis@nyu.edu
2 Instrumental Conditioning
Pavlovian (Classical) Conditioning

- Learning predictions about biologically important events
- Outcomes (like food) are not contingent on behavior

Instrumental Conditioning

- Learning to take actions
- Outcomes are contingent on behavior
- aka “operant conditioning”
Pavlovian (Classical) Conditioning

- Learning predictions about biologically important events

Most dependable difference is about the contingency between action.

Instrumental Conditioning

Learning to take actions

Outcomes are contingent on behavior

aka “operant conditioning”

Many aspects of classical conditioning and instrumental are the similar (e.g., extinction, learning curves, etc...)
Edward Thorndike

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" (Thorndike, 1911)
Edward Thorndike

Law of Effect

First trial in puzzle box

After many trials in puzzle box

Stimuli inside of puzzle box

- Scratch at bars
- Push at ceiling
- Dig at door
- Howl
- Etc.
- Press Level

Stimuli inside of puzzle box

- Scratch at bars
- Push at ceiling
- Dig at door
- Howl
- Etc.
- Press Level
Edward Thorndike
Law of Effect

- According to Thorndike, the reinforcer “stamps in” the association between the situation and some action.
- Reinforcer not needed after training: behavior becomes habitual.
- No need to assume more intelligence of causal learning/insight.
- Automatic learning process once there is a goal (motivation).
Edward Thorndike

Law of Effect

Many determinants of responding:

- Drive/motivation - affects learning and performance
- Reward magnitude & contrast effects
- Delay to reward (long time general leads to slower learning)
- Frequency of reinforcement (continuous versus partial)

Instrumental conditioning as a form of adaptive control over the environment: animal behaves such as to bring about good things and avoid bad things
“Skinner Box”

developed the “free-operant” procedure

1. Hungry rat pressed a lever for food

2. Some programmed relationship between presses & rewards (“schedule”)

3. Main dependent variable: response rate (also timing, pattern, vigor), as a function of schedule
Superstition
Reinforcement (food) at fixed intervals leads to strange, stereotyped behavior like dancing and thrusting head into the corner of the cage, etc...

Humans do it!
Reinforcement (food) at fixed intervals leads to strange, stereotyped behavior like dancing and thrusting head into the corner of the cage, etc... Humans do it! Not just a clever experiment (the first *Psych Science* paper?!) Demonstrates contiguity between ongoing actions and rewards in driving instrumental learning (consistent with Thorndike’s hypothesis).
Superstition?

- Maybe not...
- Only terminal responses (pecking) increase at the end of the fixed time interval
- In-between (interim responses) were less conditioned but may reflect based aspects of the way organisms search for food when hungry (the superstitious behavior is just the way pigeons forage!)
- Pre-organized, species-specific patterns of foraging and feeding behaviors ... a search strategy!
- See Timberlake & Lucas, 1989; Staddon & Simmelhag, 1971
The concept of “Belongingness”

- Just as certain CS-US relationships are easier to acquire, certain behaviors are easier to condition instrumentally (remember Garcia & Koelling, 1966 finding on conditioning tastes/sickness/shock):
  - Harder to get cats to condition grooming to escape relative to playing with strings (Thorndike)
  - At zoo pigs trained to drop coin in a slot try kicking it rather than dropping with mouth
  - ...and raccoons keep playing with the coins and never dropping them in (Breland & Breland, 1961)
  - and yet instrumental conditioning remains incredibly powerful...

http://www.youtube.com/watch?v=Nc9xq-TVyHi
Learned Helplessness
Seligman & Maier, 1967

- Animals given painful shocks at random intervals, other group could perform action to avoid shock
- Subsequently put in new situation where could learn action to avoid a shock. Only the group that had control in the first part learned.
- Typical interpretation is that animals who lacked control learned there was nothing they could do to escape shocks (learning to learn?)
- Violation of “Law of Effect”!? Shows important of reward salience, interpretation of the task, and general knowledge states on instrumental conditioning
- A possible animal/learning model for certain forms of depression?
"Skinner Box"

“free-operant” procedure

1. Hungry rat presses a lever for food
2. Some programmed relationship between presses & rewards ("schedule")
3. Main dependent variable: response rate (also timing, pattern, vigor), as a function of schedule
4. (sometimes) there is a stimulus (S) that appear before the response that alerts the animal that the response is available/warranted

S
stimulus

R
response

O
outcome
Curious Stimulus Effects
(operate outside of reward to control behavior)

- **Habit slip:** rats in a maze will run right past a food bowl on their way to where they expect reward (the maze stimulus is controlling behavior in a powerful way) (Stoltz & Lott, 964)

- Pressing the 4th floor button when you work on the 8th floor

- **Protestant ethic effect:** pigeons learn first to perform an action for food while an empty food well is in their cage. Later free food placed in the well. Prefer to work for the food than get the open reward. (Neuringer, 1969)
Instrumental Responses

- **Shaping**: the process of slowing building up more complex behaviors through strategic reinforcement

- **Basically reward successive approximations of the desired behavior**

- To teach bird to peck at hopper for food, first reinforce movement toward the hopper, then for contact with the lever, the closer and closer to desired behavior

- **Autoshaping**: just light up response button and give reward at fixed intervals, eventually pigeon will automatically start tapping button and getting reward.

- **Chaining**: building up sequences of actions one after another. Usually best to start with things closer to final reward (last action in chain) and go backwards (something predicted from TD(\(\lambda\)))
Instrumental Conditioning and Motivation (we’ll have more to say about this next time)

Typical finding:

- Increase the # of pecks required for food, animals won’t work as hard
- Make rats less hungry, they don’t work as hard

Puzzle: Why don’t animals just work as hard as possible to get the reward?

(Foster et al 1997)
Schedules of Reinforcement

1 Ratio schedules

- Rewarded after N leverpresses

- Variable ratio (VR): N is randomized, with some average requirement (VR30: 1/30th of leverpresses are rewarded)

2 Interval schedules

- Reward becomes available T seconds after last one is retrieved

  - If T is fixed (say 30s) how should you behave?

- Variable interval (VI): T in randomized (VR30: every second there is a 1/30th chance reward will become available)

  - now how should you behave?
Schedules of Reinforcement
## Contingency Effects

<table>
<thead>
<tr>
<th>Response/CS</th>
<th>US/Reinforcer</th>
<th>No US/No Reinforcer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response/CS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No response/No CS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Contingency = \( p(\text{US}|\text{CS}) - P(\text{US}/\text{no CS}) \)

\[ = \frac{a}{a+b} - \frac{c}{c+d} \]

Each cell always visited in classical case (set up by experimenter)

Not necessarily the case in instrumental... have to sample or EXPLORE cells to gain information
Relating learning, choice, and the ‘Strength’ of responding

Concurrent schedules! --- Akin to an N-armed bandit in computer science:

Put out multiple levers with different reward schedules and let the animal choose between them.

Say one lever make food available, on average, 30 secs after you pick it up, the other on average, 60 seconds.

what should you do?
The Matching Law

Animals share responses between levers

1. Favor the richer option proportionally more

2. In the variable interval task, this is nearly optimal

The Matching Law

\[
\frac{C(A)}{C(A) + C(B)} = \frac{r(A)}{r(A) + r(B)}
\]

\(C(A)\): # of A choices

\(r(A)\): # of times A rewarded
The Matching Law

Matching law holds when magnitudes of rewards are manipulated (keeping arming intervals constant):

In this case, matching is the proportion of total reward received from a choice.

2. In the variable interval task, this is nearly optimal.

The Matching Law

% of A choices matches % rewards on A:

\[ \frac{C(A)}{C(A) + C(B)} = \frac{r(A)}{r(A) + r(B)} \]

- \( C(A) \): # of A choices
- \( r(A) \): total reward from A
The Matching Law

\[
\frac{C(A)}{C(A) + C(B)} = \frac{r(A)}{r(A) + r(B)}
\]

- Data is a list of choices (order invariant):
  - Ignores facts like pacing of responding, how fast/often they choose
  - Is it applicable to discrete trial problems (i.e., not free operant)?
  - Is the matching law a description or a process?

- Focuses on overall, steady-state responses:
  - prediction is about aggregate of statistics over a large number of repeated choices
  - doesn’t say anything about learning itself
Optimal foraging behavior

Ideal Free Distribution (Fretwell & Lucas, 1972; Seth, 2001)

Interestingly, animals distribute themselves among patches so as to maximize the gained resources.

Equilibrium achieved when no organism can profit from moving

Harper (1982)

Godin & Keenleyside (1984)
Even human groups!


(Well they undermatch a bit)
The Matching Law

\[ \frac{C(A)}{C(A) + C(B)} = \frac{r(A)}{r(A) + r(B)} \]

- So far discussion has been focused on interval schedules
- What about choice between ratio schedules?
  - Lever A pays off with 1/30 probability (VR30)
  - Lever B pays off with 1/60 probability (VR60)
- Now what should you do?
The Matching Law

\[
\frac{C(A)}{C(A) + C(B)} = \frac{r(A)}{r(A) + r(B)}
\]

- So far discussion has been focused on interval schedules.
- What about choice between ratio schedules?
  - Lever A pays off with 1/30 probability (VR30)
  - Lever B pays off with 1/60 probability (VR60)
- Now what should you do?

MAXIMIZE! Choose A, never B!
The Matching Law

\[
\frac{C(A)}{C(A) + C(B)} = \frac{r(A)}{r(A) + r(B)}
\]

- In the interval situation **maximizing is matching!**

- Any differential allocation of choice to option A will result in more reward for that option than for B. By that logic this should increase allocation to A until, in the limit, exclusively A is chosen.

- Thus, the matching law doesn’t necessarily predict matching **behavior**
The Matching Law - Summary

\[
\frac{C(A)}{C(A) + C(B)} = \frac{r(A)}{r(A) + r(B)}
\]

\* Variable interval task:

\* Common sense: divide choices between levers (because a neglected choice will eventually be baited)

\* Matching law predicts: divide choices between levers

\* Animals do - divide choices

\* Variable ratio task:

\* Common sense: choose richer option exclusively (it is a better bet)

\* Matching law predicts: choose richer option exclusively

\* Animals do - generally choose the better option exclusively
Explanations for matching


Anticipation by our discussion of group foraging.

1. People track average satisfaction or value per unit invested in each alternative

2. Melioration: based on these values, they shift behavior/choices to alternatives that provide a higher per-unit return

Can be tested in dynamic tasks there contingencies are linked to subject’s previous choices.
Matching as a Decision

- One argument for the “rationality” of matching is that the world is often dynamically changing. If so, then sticking with what you currently think is best is a bad idea.

- Even besides this, the existing of matching implies that the brain must somewhere compute something about the relative value of possible actions.

- Thus, matching can be used to study the neural representation of relatively value (c.f., Sugrue, Corrado, Newsome, 2004)

- Requires two components
  - A learning procedure
  - A choice rule (the matching law)
Rhesus monkeys make eye fixations to one of two locations which get rewarded at different rates.

Dynamic foraging task since the reward from each option “drifts” in time (Poisson arrival times, but one are baited are always available... like VI schedule, makes matching the optimal strategy)

Ever 200 or so trials, the reward magnitudes or rates change suddenly causing discontinuous shifts in the reward rates.
Optimal integration of the reward history would be incapable of this quick adaptation.

Instead, suggest some “leaky integrator” which is pooling information in time (recency biased).

Distinct from traditional matching law which integrates “globally”.
Matching Behavior and the Representation of Value in the Parietal Cortex

Leo P. Sugrue,* Greg S. Corrado, William T. Newsome

weight for last trial
weight for trial before that

trials into past
Matching Behavior and the Representation of Value in the Parietal Cortex

Leo P. Sugrue,* Greg S. Corrado, William T. Newsome

How do the monkeys do?

Near optimal setting of tau

The monkey’s match!
Matching Behavior and the Representation of Value in the Parietal Cortex

Leo P. Sugsue,* Greg S. Corrado, William T. Newsome

The model

Learning  Matching

choice
Recording from target-selective neuron is LIP

- Region involved in planning and executing saccades
Recording from target-selective neuron is LIP

Selected on the basis of delayed saccade task to find RF selective cells (like a forced choice version of matching game)
Recording from target-selective neuron is LIP

Cells modulated by trial-to-trial utility of target in the receptive field
Recording from target-selective neuron is LIP

Cells modulated by trial-to-trial utility of target in the receptive field
The take-home

- LIP neurons have a spatial map of sensory, motor, & utility function.

- They report dynamic utilities that change trial-by-trial with learning (not unlike that predicted by rescorla-wagner).

- Matching behavior, rather than being an oddity of irrational choice can be used to STUDY the neural representation of value.

- Dynamic task highlights how matching behavior and local learning is important for dealing with a dynamic and changing environment.
Other tidbit: The Premack principal

- We don’t ONLY go around chasing primary rewards

- In some cases, limiting animals access to a behavior they WANT to do naturally can act as reinforcement... known as the Premack principal

- Basically for kids, watching TV is a preferred activity to homework. By limiting TV access, TV can become a reward for doing homework

- Premack (1959) gave rats free access to running wheel and water. Spent 250 seconds running and 50 seconds drinking.

- Then, set up contingency where could only run after drank a certain amount. Rats learn contingency and increase water drinking to gain access to wheel.
What is learned in instrumental conditioning?

Thorndike

Situation → Response

reward

Skinner

what is the S? what about S\textsuperscript{D}?

Tolman

S → cognitive map → R
Remember Latent Learning?

**Fig. 4**

6-Unit Alley T-Maze


**Fig. 5**

Tolman’s point

- Previously, we talked about Tolman’s ideas/work as evidence that animals learn (in an unsupervised fashion) an “internal model” of their environment.

- A second key point we didn’t emphasize is that they **USE** this “cognitive map” to **PLAN BEHAVIORS**.
The modern debate: S-R or R-O

- **S-R theory:**
  - Parsimonious - same theory for Pavlovian conditioning (CS/CR pairing)
  - But it overlook the critical contingency in instrumental conditioning between the response and outcome

- **alternative: R-O theory (or A-O)**
  - proponents: Rescorla, Dickinson, et al.
  - same spirit as Tolman (know ‘map’ of contingencies and desires, and can use this to plan... maybe INFORMATION PROCESSING!)
A good test: Outcome devaluation
A good test: Outcome devaluation

Holland (2004) see also Dickinson article for examples
A good test: Outcome devaluation

Holland (2004) see also Dickinson article for examples

Suggests some measure of reward or outcome **expectancy** (shifts to habitual control with extensive training)
Goal directed vs. Habitual Behavior

“Cognitive” vs. “Associationist”
Goal directed vs. Habitual Behavior

“Cognitive” vs. “Associationist”

Be very afraid. The rats of NIHM are really smart!
Devaluation and Lesions

- Systematic patterns of behavior in devaluation studies suggest distinct neural pathways involved in either the goal-directed or habitual system.

- Animals with lesion to Dorsolateral Striatum (DLS) *never develop habits* despite extensive training.

- Also happens with drugs that deplete dopamine in DLS.

- Also happens with lesions to infralimbic division of PFC (same corticostriatal loop).

Yin et al. (2004) - Slides from Y. Niv
Systematic patterns of behavior in devaluation studies suggest distinct neural pathways involved in either the goal-directed or habitual system.

After habit have been formed, devaluation sensitivity can be reinstated by temporary inactivation of IL PFC.

Coutureau & Killcross (2003) - Slides from Y. Niv
Devaluation and Lesions

- Systematic patterns of behavior in devaluation studies suggest distinct neural pathways involved in either the goal-directed or habitual system.

- Prelimbic (PL) PFC lesions cause animals to leverpress \textit{habitually} even with only moderate training (along with dmPFC and mediodorsal thalamus which are in same loop).

Killcross & Coutureau (2003) - Slides from Y. Niv
What does it mean?
The same action (lever pressing) can arise from psychologically & neurally dissociable pathways:

- Moderately trained behavior is “goal-directed”: dependent on outcome representations, like planning with a cognitive map

- Overtrained behavior is “habitual”: not dependent on outcome representation, as predicted by S-R theory

- Which is right S-R or R-O? probably BOTH!

- Lesions suggest that two parallel systems interactively contribute to behavior in any situation
The distinction between the habit-based controller and the goal-directed controller can be directly related to the concept of “model-based” or “model-free” learning algorithms in computer science (particularly the subfield of reinforcement learning).

What does a “model-free” model mean?
Key Components of RL Systems

- The environment
- Reward function
- Policy
- Value function
- Model of the environment
The environment

- Modeled as a Markov Decision Process
- At its essence means that the system is defined by the one-step dynamics
- Simply put, the distribution of future states and rewards depend only on where you are now and what action you take

$$Pr\{s_{t+1} = s', r_{t+1} = r | s_t, a_t \}$$
Reward Function

\[ R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \]

- Typically a single number which indicates how good or bad the current state is.

- The overall goal of the agent is to maximize the discounted reward it gets over the long term.

- The parameter (gamma) determines how much weight is given to immediate version delayed rewards.

- Reward are the immediate, primary sensory feedback from a particular state, in contrast to value functions.
The rules for how an agent should act. A full set of stimulus-response rules or associations.

Represented as $\pi$, where $\pi(s, a)$ means the probability of selecting action $a$ in state $s$.

Policies can be explicit stochastic in nature or deterministic.

What we are trying to learn: a good policy for the environment we face.
Value Function

\[ V^\pi = E_\pi \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right\} \]

- The long run total that an agent can expect to make in the future starting from that state. So unlike rewards these are not immediate short-term values, but based on a longer temporal window.

- Since the goal is to maximize reward over the long term, in a sense you are trying to choose actions or states associated with higher values.

- The value function is essentially a stand-in for what you will get in the long run from a particular choice and is important in learning.
Model of Environment

Knowledge about the way the world works

Essentially means a representation of the Markov process itself such as the probabilities of moving from state a to state b given that you take action c

Not absolutely necessary (e.g., Monte Carlo methods or temporal difference learning can operate without an explicit model!!)

\[
\mathcal{P}_{ss'}^a = Pr\{s_{t+1} = s' | s_t = s, a_t = a\}
\]

\[
\mathcal{R}_{ss'}^a = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\}
\]
An Example: Behaving optimally in a known world

Rewards & State Transitions
An Example: Behaving optimally in a known world

Rewards & State Transitions

Agent’s Policy (\(\pi\))
An Example: Behaving optimally in a known world

Rewards & State Transitions

Agent's Policy ($\pi$)

Value Function ($V$)

$\gamma = 0$
Bellman Equation

\[ V^\pi(s) = \sum_a \pi(s, a) \sum_{s'} P_{ss'}^a \left[ R_{ss'}^a + \gamma V^\pi(s') \right] \]

If we know the function \( P_{ss'}^a \) and \( R_{ss'}^a \) (i.e. have perfect knowledge of the environment), we can easily solve for \( V^\pi(s) \) by simply solving the systems of equations under our policy.
Rewards & State Transitions

Agent's Policy ($\pi$)

Value Function ($V$)

\[
\begin{array}{cccccc}
-0.5 & 10 & -0.25 & 5 & -0.5 \\
-0.25 & 0 & 0 & 0 & -0.25 \\
-0.25 & 0 & 0 & 0 & -0.25 \\
-0.25 & 0 & 0 & 0 & -0.25 \\
-0.5 & -0.25 & -0.25 & -0.25 & -0.5 \\
\end{array}
\]
Rewards & State Transitions

Agent's Policy (\(\pi\))

Value Function (\(V\))

\[ \begin{array}{cccccc} 
3.3 & 8.8 & 4.4 & 5.3 & 1.5 \\
1.5 & 3.0 & 2.3 & 1.9 & 0.5 \\
0.1 & 0.7 & 0.7 & 0.4 & -0.4 \\
-1.0 & -0.4 & -0.4 & -0.6 & -1.2 \\
-1.9 & -1.3 & -1.2 & -1.4 & -2.0 \\
\end{array} \]
Rewards & State Transitions

Agent’s Policy (π)

Value Function (V)

\[ \gamma = 0.9 \]
Finding the optimal policy

\[ V^*(s) = \max_a \sum_{s'} P_{ss'}^a [R_{ss'}^a + \gamma V(s')] \]

Again, if we know the function \( P_{ss'}^a \) and \( R_{ss'}^a \) (i.e., have perfect knowledge of the environment), we can easily solve for \( V^\pi(s) \) by simply solving the systems of equations under our policy \( \pi \).
Rewards & State Transitions

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>+5</th>
<th></th>
<th>B’</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td>+10</td>
<td></td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>A’</td>
<td></td>
<td></td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>

Agent’s Policy ($\pi$)

Value Function ($V$)

<table>
<thead>
<tr>
<th></th>
<th>$\pi^*$</th>
<th>$\gamma = 0.9$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$V^*$</td>
<td></td>
</tr>
<tr>
<td>22.0</td>
<td>24.4</td>
<td>22.0</td>
<td>19.4</td>
</tr>
<tr>
<td>19.8</td>
<td>22.0</td>
<td>19.8</td>
<td>17.8</td>
</tr>
<tr>
<td>17.8</td>
<td>19.8</td>
<td>17.8</td>
<td>16.0</td>
</tr>
<tr>
<td>16.0</td>
<td>17.8</td>
<td>16.0</td>
<td>14.4</td>
</tr>
<tr>
<td>14.4</td>
<td>16.0</td>
<td>14.4</td>
<td>13.0</td>
</tr>
</tbody>
</table>
Evaluating the world when you don’t know anything about it

\[ V^*(s) = \max_s \sum P_{ss'} [R_{ss'} + \gamma V(s')] \]

\[ V(s_t) = V(s_t) + \alpha \{ r_{t+1} - \gamma V(s_{t+1}) - V(s_t) \} \]

- Run around in the world following your policy and just keep track of how much you’ve earned from each state
- Temporal Difference methods (incrementally bootstrap your estimates of the value of various states as you are exploring/learning)
Summary

- The goal of the RL agent is to maximize reward over the long term.

- The way this is implemented is through a function which determines the value of various “states” or “situations” under a certain policy.

- Once you know how to evaluate a policy, there are a number of ways to actually arrive at optimal policies.

- In many of the most interesting cases, it essentially reduces to something like: start out exploring a lot, then slowly become more and more biased by the values you’ve experienced.

- **CRITICALLY**, there is more than one way to solve the RL problem depending on if you represent the world model or not.
Key idea

from Daw, Niv, Dayan, 2005
Key idea

Model-based system

Model-free system

from Daw, Niv, Dayan, 2005
Strategy 1: Model Based RL

Learn model of task through experience (= cognitive map)
Compute Q values by “looking ahead” in the map
Computationally costly, but also flexible (immediately sensitive to change)
strategy II: model free RL

- Shortcut: store long-term values
  - then simply retrieve them to choose action
  - Q(S0,L) = 4
  - Q(S0,R) = 2

- Can learn these from experience
  - without building or searching a model
  - incrementally through prediction errors
  - dopamine dependent SARSA/Q-learning or Actor/Critic

Q(S1,L) = 4
Q(S1,R) = 0
Q(S2,L) = 1
Q(S2,R) = 2

from Yael Niv
strategy II: model free RL

- 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:
- $Q(S_0, L) = 4$
- $Q(S_0, R) = 2$
- $Q(S_1, L) = 4$
- $Q(S_1, R) = 0$
- $Q(S_2, L) = 1$
- $Q(S_2, R) = 2$
two big questions

- Why should the brain use two different strategies/controllers in parallel?
- If it uses two: how can it arbitrate between the two when they disagree (new decision making problem…)

from Yael Niv
answers

- each system is best in different situations (use each one when it is most suitable/most accurate)
  - goal-directed (forward search) - good with limited training, close to the reward (don’t have to search ahead too far)
  - habitual (cache) - good after much experience, distance from reward not so important

- arbitration: trust the system that is more confident in its recommendation
  - different sources of uncertainty in the two systems
  - don’t always choose the highest value
  - uncertainty is different from risk

from Yael Niv
Intermittent Summary

- Instrumental behavior is not a simple unitary phenomenon: the same behavior can result from different neural and computational origins.

- The distinction between habit and goal-directed instrumental behavior can be understood, computationally, as a distinction between model-based and model-free approaches to RL (maximizing reward).

- The different neural mechanisms work in parallel to support behavior: cooperation and competition.

- Useful tests: outcome devaluation, contingency degradation.

from Yael Niv
The take-home for today

- Instrumental conditioning is a form of adaptive control over the environment: the animal behaves such as to bring about good things and avoid bad things.

- Way more complex than classical conditioning. True understanding involves theories of choice, motivation, reward valuation, planning, impulse control, stimulus processing and generalization, theories of response rate, etc. In other words, it is a mess.

- That said, next time we will try to clean up that mess by talking about modern theories that have made considerable progress explaining many of the effects we just covered.

- Despite the confusingness that is instrumental learning, the situation is much more relevant to society (people are free to act, we want to understand how they decide to act and forces in their environment that influence this). Ultimately we have no choice but to understand the mess.
Readings for next time


More coming soon. Check the website Friday afternoon.
References for Slides


Lecture notes from Yael Niv (http://www.princeton.edu/~yael/PSY338/index.html), Peter Dayan, and Nathaniel Daw The interweb.