Learning and Memory

Lecture 6/14

Cate Hartley and Todd Gureckis
Department of Psychology
New York University
{cate, 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

Instrumental Conditioning

- Learning to take actions
- Outcomes are contingent on behavior

aka "operant conditioning"

Most dependable difference is about the contingency between action.

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

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 learn to groom in order 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-IVyHI
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 importance of prior knowledge states on instrumental conditioning
- A possible 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

<table>
<thead>
<tr>
<th>S</th>
<th>R</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>stimulus</td>
<td>response</td>
<td>outcome</td>
</tr>
</tbody>
</table>
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, 1964)

- 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

- Fixed ratio
- Variable ratio
- Fixed interval
- Variable interval
## Contingency Effects

<table>
<thead>
<tr>
<th></th>
<th>US/Reinforcer</th>
<th>No US/No Reinforcer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response/CS</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>No response/No CS</td>
<td>$c$</td>
<td>$d$</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 press it, 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

% 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

$r(A)$: # 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 elsewhere.

Harper (1982)

Godin & Keenleyside (1984)
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 where contingencies are linked to subject’s previous choices.
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 once baited are always available... like VI schedule, makes matching the optimal strategy).

Every 200 or so trials, the reward magnitudes or rates change suddenly causing discontinuous shifts in the reward rates.
Matching Behavior and the Representation of Value in the Parietal Cortex

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

Learning

- 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”

A Global Matching Law

\[
\frac{I_{\text{red}}}{I_{\text{red}} + I_{\text{green}}} = \frac{C_{\text{red}}}{C_{\text{red}} + C_{\text{green}}}
\]

B Local Matching Law

\[
\frac{\hat{I}_{\text{red}}}{\hat{I}_{\text{red}} + \hat{I}_{\text{green}}} = pC_{\text{red}}
\]
Matching Behavior and the Representation of Value in the Parietal Cortex

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

The model

![Diagram showing the model of matching behavior and representation of value in the parietal cortex. The diagram includes leaky integrators for red and green reward histories, local income, local fractional income, and probability of choice.]
Recording from target-selective neuron in LIP

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

Selected on the basis of delayed saccade task to find RF selective cells (like a forced choice version of matching game)
Matching Behavior and the Representation of Value in the Parietal Cortex
Leo P. Sugrue,* Greg S. Corrado, William T. Newsome

Recording from target-selective neuron in LIP

Cells modulated by trial-to-trial utility of target in the receptive field
Recording from target-selective neuron in 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 $S^D$?

Tolman

$S$ → cognitive map → $R$
Remember Latent Learning?


(From H. C. Bledgett, The effect of the introduction of reward upon the maze performance of rats. *Univ. Calif. Publ. Psychol.*, 1929, 4, No. 8, p. 120.)
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

Phase 1 - Training:

Phase 2 - Pairing with illness:

Phase 2 - Motivational shift:

Phase 3 - Test: (extinction)

Unshifted

Non-devalued

Hungry

Sated

TRENDS in Cognitive Sciences
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 NIMH 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
Devaluation and Lesions

- 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.

![Graph showing moderate training](Image)

- Prelimbic (PL) PFC lesions cause animals to leverpress *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.