Instrumental Conditioning VI:
There is more than one kind of learning

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

outline

• what goes into instrumental associations?
• goal directed versus habitual behavior
• neural dissociations between habitual and goal-directed behavior
• how does all this fit in with reinforcement learning?
Thorndike:  
\[ S \rightarrow R \]

reinforcer

Skinner:  
what is the S?

Tolman:  
\[ S \rightarrow \text{cognitive map} \rightarrow R \]

“The stimuli are not connected by just simple one-to-one switches to the outgoing responses. Rather, the incoming impulses are usually worked over and elaborated in the central control room into a tentative, cognitive-like map of the environment. And it is this tentative map, indicating routes and paths and environmental relationships, which finally determines what responses, if any, the animal will finally release.”

Tolman:  
\[ S \rightarrow \text{cognitive map} \rightarrow R \]
Maze task

- train rats to find food in a maze
- second group: exposed to maze but without food
- compare the groups in subsequent test with food
- what do you think will happen?
- what does this demonstrate?

Maze task: Latent learning

Blodgett (1929)
another example: shortcuts

training:

test:

result:

Tolman et al (1946)

summary so far...

• Even the humble rat can learn & internally represent spatial structure, and use it to plan flexibly
• Tolman relates this to all of society
• Note that spatial tasks are really complicated & hard to control
• Next: search for modern versions of these effects
• Key question: is S-R model ever relevant? and what is there beyond it? (especially important given what we know about RL)
the modern debate: S-R vs R-O

- S-R theory:
  - parsimonious - same theory for Pavlovian conditioning (CS associated with CR) and instrumental conditioning (stimulus associated with response)
  - but: the critical contingency in instrumental conditioning is that of the response and the outcome...

- alternative: R-O theory (also called A-O)
  - among proponents: Rescorla, Dickinson
  - same spirit as Tolman (know ‘map’ of contingencies and desires, can put 2+2 together)

How would you test this?

outcome devaluation

1 - Training:

2 - Pairing with illness:

2 - Motivational shift:

3 - Test: (extinction)

Q1: why test without rewards?
Q2: what do you think will happen?
Q3: what would Tolman/Thorndike guess?

will animals work for food they don’t want?
Animals will sometimes work for food they don’t want!

→ in daily life: actions become automatic (habitual) with repetition

Holland (2004)

outline

• what goes into instrumental associations?
• goal directed versus habitual behavior
• neural dissociations between habitual and goal-directed behavior
• how does all this fit in with reinforcement learning?
animals with lesions to DLS never develop habits despite extensive training

→ also treatments depleting dopamine in DLS

→ also lesions to infralimbic division of PFC (same corticostriatal loop)


dervaluation: results from lesions I

overtrained rats

dorsolateral striatum lesion control (sham lesion)

% of final response rate

Devalued Valued

after habits have been formed, devaluation sensitivity can be reinstated by temporary inactivation of IL PFC

Coutureau & Killcross (2003) 14
devaluation: results from lesions III

Lesions of the pDMS cause animals to leverpress habitually even with only moderate training.

Yin, Ostlund, Knowlton & Balleine (2005)

devaluation: results from lesions IV

Prelimbic (PL) PFC lesions cause animals to leverpress habitually even with only moderate training (also dorsomedial PFC and mediodorsal thalamus (same loop)).

complex picture of behavioral control

neural dissociation between goal-directed and habitual controllers

inspired by Balleine (2005)

what does all this mean?

• The same action (leverpressing) can arise from two psychologically & neurally dissociable pathways
  1. moderately trained behavior is “goal-directed”: dependent on outcome representation, like cognitive map (also associated with hippocampus - literal or abstract map of environment)
  2. overtrained behavior is “habitual”: apparently not dependent on outcome representation, like S-R
• S-R habits really do exist, they just don’t describe all of animal behavior
• Lesions suggest two parallel systems, in that the intact one can apparently support behavior at any stage
behavior is not always consistent:
leverpressing is habitual and continues for unwanted food…
...at same time nosepoking is reduced (explanations?)

why are nosepokes always sensitive to devaluation?

- Balleine & Dickinson: 3rd system - Pavlovian behavior is
directly sensitive to outcome value
- But: doesn’t make sense... the Pavlovian system has
information that it is withholding from the instrumental
system?
- Also.. true for purely instrumental chain
- And anyway, it seems that all the information is around all
the time, so why is behavior not always goal-directed?
**outline**

- what goes into instrumental associations?
- goal directed versus habitual behavior
- neural dissociations between habitual and goal-directed behavior
- how does all this fit in with reinforcement learning?

**back to RL framework for decisions**

3 states: "no food", "food in mag", "eating"
2 actions: "press lever", "poke nose"
immediate reward is 1 in state "eating" and 0 otherwise

need to know long term consequences of actions $Q(S,a)$ in order to choose the best one

*how can these be learned?*
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 1: “model-based” RL

no food
press lever
food in mag
poke nose
eating

nose poke

press lever

food in mag

poke nose
eating

r = 1

I predict sales to be nothing for two years and then take a sudden surge.

The surge was added so I could get the business case approved. The two-year lag gives me time to get promoted.

23

24
strategy II: “model-free” RL

• Shortcut: store long-term values
  – then simply retrieve them to choose action

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

\[
\begin{align*}
Q(S_0, L) &= 4 \\
Q(S_0, R) &= 2 \\
Q(S_1, L) &= 4 \\
Q(S_1, R) &= 0 \\
Q(S_2, L) &= 1 \\
Q(S_2, R) &= 2
\end{align*}
\]

• 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)
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…)

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
  • compare to: always choose the highest value
back to animals pressing a lever for a devalued food, but not nose-poking to get it: can you explain this?

summary

• instrumental behavior is not a simple unitary phenomenon: the same behavior can result from different neural and computational origins

• different neural mechanisms work in parallel to support behavior: cooperation and competition

• useful tests: outcome devaluation, contingency degradation