There is more than one kind of learning

PSY/NEU338: from animal learning to changing people’s minds

Where were we?

- Instrumental conditioning: action selection can be modeled using our models of Pavlovian conditioning — error-correcting learning rules on the values of actions (SARSA, Q learning)
- Can also use Pavlovian prediction errors to learn action propensities directly (Actor/Critic)
- These models can even explain free-operant (not trial-based) behavior: choice of response rates, and how motivation affects these
- In the brain: dopamine prediction errors train actor and critic in the striatum
what goes into an association?

Thorndike:  
\[ S \rightarrow R \]
reinforcer

Skinner:  
what is the S?

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

Edward Tolman (1886-1959)

“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:  
\[ S \xrightarrow{\text{cognitive map}} R \]
**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

trainig:

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 (read his 1948 paper!)
• Note that spatial tasks are really complicated & hard to control
• Next: more modern versions of these findings

• Key question: is S-R model ever relevant? and what is there beyond it? (especially important given what we know about RL)

This completes my report of experiments. And now, at last, I come to the humanly significant and exciting problem: namely, what are the conditions which favor narrow strip-maps and what are those which tend to favor broad comprehensive maps not only in rats but also in men?

There is considerable evidence scattered throughout the literature bearing on this question both for rats and for men. Some of this evidence was obtained in Berkeley and some of it elsewhere. I have not time to present it in any detail. I can merely summarize it by saying that narrow strip maps rather than broad comprehensive maps seem to be induced: (1) by a damaged brain, (2) by an inadequate array of environmentally presented cues, (3) by an overdose of repetitions on the original trained-on path and (4) by the presence of too strongly motivational or of too strongly frustrating conditions.

My argument will be brief, cavalier, and dogmatic. For I am not myself a clinician or a social psychologist. What I am going to say must be considered, therefore, simply as in the nature of a rat psychologist’s ratiocinations offered free.

Over and over again men are blinded by too violent motivations and too intense frustrations into blind and unintelligent and in the end desperately dangerous hates of outsiders. And the expression of these displaced hates ranges all the way from discrimination against minorities to world conflagrations.

What is the name of Heaven and Psychology can we do about it? My only answer is to preach again the virtues of reason—of, that is, broad cognitive maps. And to suggest that the child-trainers and the world-planners of the future can only, if at all, bring about the presence of the required rationality (i.e., comprehensive maps) if they see to it that nobody’s children are too over-motivated or too frustrated. Only then can these children learn to look before and after, learn to see that there are often round-about and safer paths to their quite proper goals—learn, that is, to realize that the well-beings of White and of Negro, of Catholic and of Protestant, of Christian and of Jew, of American and of Russian (and even of males and females) are mutually interdependent.
the modern(-ish) debate: S-R vs A-O

• S-R theory:
  • parsimonious - same theory for Pavlovian conditioning (CS associated with CR) and instrumental conditioning
  • Fits well with reinforcement learning (action values, actor/critic)
  • but: the critical contingency in instrumental conditioning is that of the action and the outcome…

• alternative: A-O theory (also called R-O)
  • among proponents: Rescorla, Dickinson, same spirit as Tolman
  • critical ingredients are knowing ‘map’ of A→O contingencies (transitions) and desires (which Os are valuable), and then put 2+2 together to plan actions.

How can we 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?

Non-devalued  
Unshifted

Hungry  Sated

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) 13

outline

• goal directed versus habitual behavior
• neural dissociations between habitual and goal-directed behavior
• how does all this fit in with reinforcement learning?
devaluation: results from lesions I

overtrained rats

→ 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)

Yin et al. (2004)
after habits have been formed, devaluation sensitivity can be reinstated by temporary inactivation of IL PFC

Coutureau & Killcross (2003)

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

Yin, Ostlund, Knowlton & Balleine (2005)
devaluation: results from lesions IV

Moderate training

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


Behavioral control in parallel loops

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**

5 min break
outline

- 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

There are 3 states: “no food”, “food in mag”, “eating”.
2 actions: “press lever”, “poke nose”
Immediate reward is 1 in state “eating” and 0 otherwise.

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

How can these be learned?
strategy I: “model-based” RL

learn model of task through experience (= cognitive map)
compute Q values by “looking ahead” (mentally simulating) in the map
computationally costly, but also flexible (immediately sensitive to change)

Q(S_1, L) = 4
Q(S_1, R) = 2
Q(S_0, L) = 0
Q(S_0, R) = 1
Q(S_2, L) = 1
Q(S_2, R) = 2

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

Stored:
Q(S_1, L) = 4
Q(S_1, R) = 0
Q(S_0, L) = 4
Q(S_0, R) = 2
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)

Why should the brain use two different strategies/controllers in parallel?

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
summary so far

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

• Model-based and model-free RL map on to goal-directed and habitual behavior, respectively

• Is this also true for humans? How do WE make choices?
dissociating learning strategies
Three example participants

Daw, Gershman, Seymour, et al 2011

Group results

Daw, Gershman, Seymour, et al 2011
use of habits varies *situationally*:
- overtraining
- stress
- distraction
- motivation, fatigue, belief

and *between individuals*:
- capacity (IQ, working memory, age)
- genetics (e.g. dopamine genes)
- drugs, neurological & psychiatric disease

\[
Q(S_1, a) = (1 - w) \cdot Q_{MF}(S_1, a) + w \cdot Q_{MB}(S_1, a)
\]
Disorders of compulsivity: a common bias towards learning habits

V. Voon1,2,3, K. Derbyshire5, C. Rück3, M.A. Irvine1, Y. Worbe2, J. Erander2, L.R.N. Schreiber6, C. Gillan7,8, N.A. Fineberg9, B.J. Sahakian1,2, T.W. Robbins1,2, N.A. Harrison1, J. Wood1, N.D. Daw10, P. Dayan11, J.E. Grant12, and E.T. Bullmore1,2,3,12

The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive

A. Ross Otto1, Samuel J. Gershman2, Arthur B. Markman1, and Nathaniel D. Daw3

Working-memory capacity protects model-based learning from stress

A. Ross Otto1, Candace M. Raio3, Alice Chiang3, Elizabeth A. Phelps1,2, and Nathaniel D. Daw1,3

Developmental changes in learning strategies

From Creatures of Habit to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based Reinforcement Learning

Johannes H. Decker¹, A. Ross Otto², Nathaniel D. Daw³, and Catherine A. Hartley¹

- Common Transition on Previous Trial
- Rare Transition on Previous Trial
- Reward
- No Reward

Outcome of Previous Trial

Proportion of First-Stage Stays

0.00
0.25
0.50
0.75
1.00
-0.25
0.00
0.25
0.50
0.75
1.00

Age

10
15
20
25
Implications for daily life

• OMG so many.

• Your habits are not dumb (your basal ganglia know you really well, from a lifetime of experience)

• Goal directed behavior is (mentally) effortful. Give it the resources it needs. But know that it is also fallible in complex/stressful situations (thank the BG for habits!)

On Making the Right Choice: The Deliberation-Without-Attention Effect

Ap Dijksterhuis,* Maarten W. Bos, Loran F. Nordgren, Rick B. van Baaren

Contrary to conventional wisdom, it is not always advantageous to engage in thorough conscious deliberation before choosing. On the basis of recent insights into the characteristics of conscious and unconscious thought, we tested the hypothesis that simple choices (such as between different towels or different sets of oven mitts) indeed produce better results after conscious thought, but that choices in complex matters (such as between different houses or different cars) should be left to unconscious thought. Named the “deliberation-without-attention” hypothesis, it was confirmed in four studies on consumer choice, both in the laboratory as well as among actual shoppers, that purchases of complex products were viewed more favorably when decisions had been made in the absence of attentive deliberation.
Primer

Model-based decision making and model-free learning

Nicole Drummond and Yael Niv

Free will is anything but free. With it comes the onus of choice: not only what to do, but which inner voice to