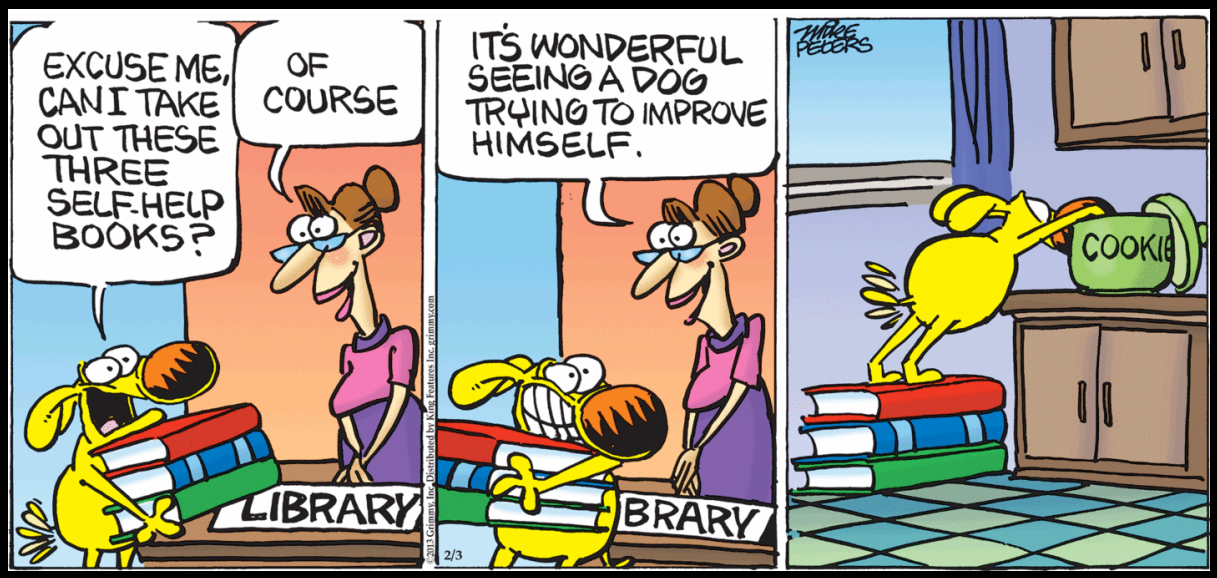
There is more than one kind of learning

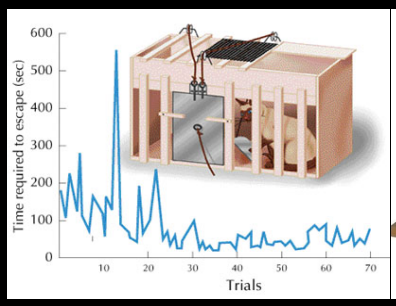


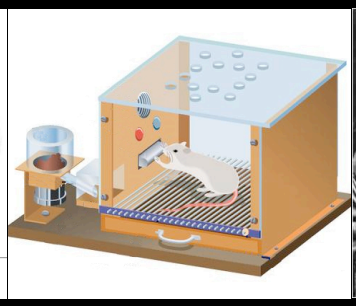
PSY/NEU338: from animal learning to changing people's minds

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 → R

reinforcer

Skinner:
what is the S?

R O

Tolman:

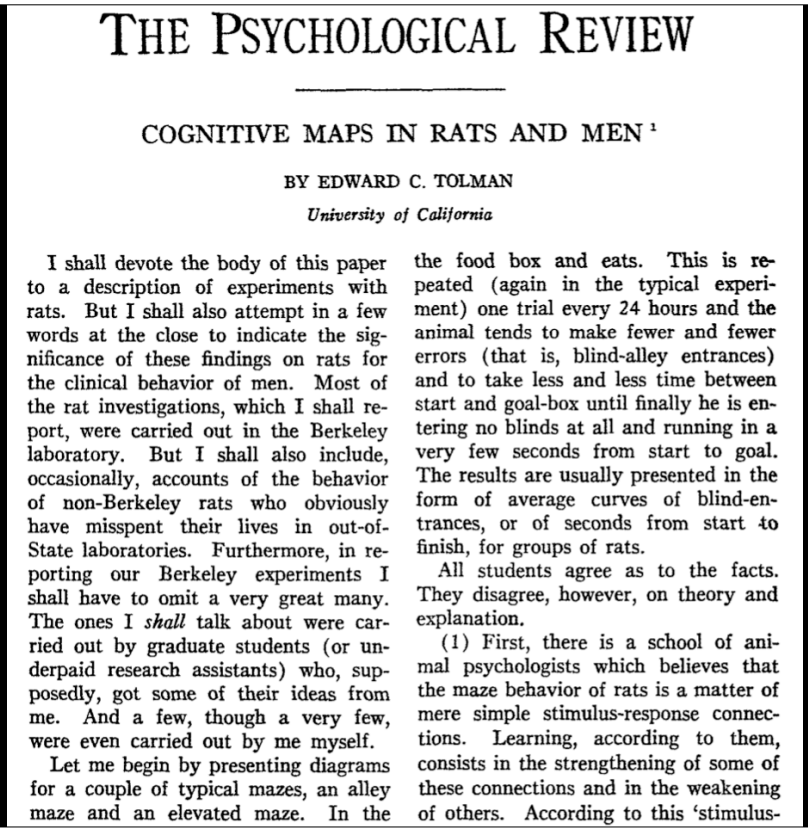
S → cognitive map → R

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 → cognitive map → R



THE PSYCHOLOGICAL REVIEW

COGNITIVE MAPS IN RATS AND MEN [1]

BY EDWARD C. TOLMAN

University of California

I shall devote the body of this paper to a description of experiments with rats. But I shall also attempt in a few words at the close to indicate the significance of these findings on rats for the clinical behavior of men. Most of the rat investigations, which I shall report, were carried out in the Berkeley laboratory. But I shall also include, occasionally, accounts of the behavior of non-Berkeley rats who obviously have misspent their lives in out-of-State laboratories. Furthermore, in reporting our Berkeley experiments I shall have to omit a very great many. The ones I *shall* talk about were carried out by graduate students (or underpaid research assistants) who, supposedly, got some of their ideas from me. And a few, though a very few, were even carried out by me myself.

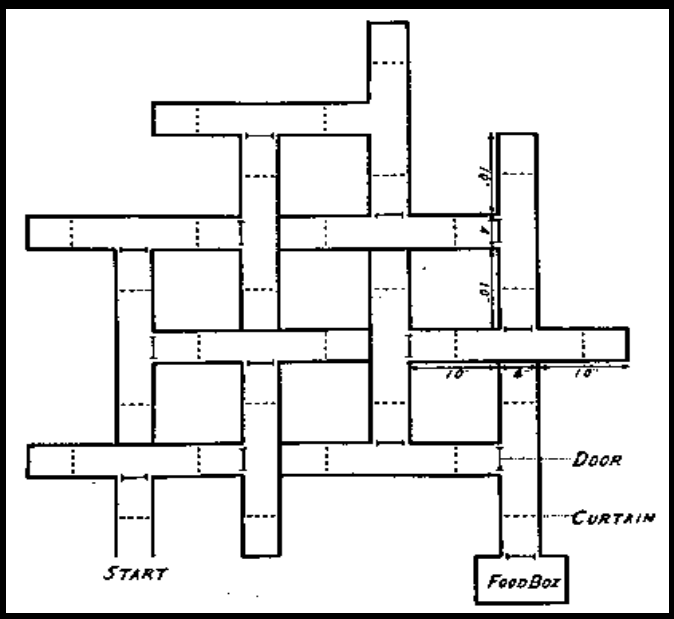
Let me begin by presenting diagrams for a couple of typical mazes, an alley maze and an elevated maze. In the

the food box and eats. This is repeated (again in the typical experiment) one trial every 24 hours and the animal tends to make fewer and fewer errors (that is, blind-alley entrances) and to take less and less time between start and goal-box until finally he is entering no blinds at all and running in a very few seconds from start to goal. The results are usually presented in the form of average curves of blind-entrances, or of seconds from start to finish, for groups of rats.

All students agree as to the facts. They disagree, however, on theory and explanation.

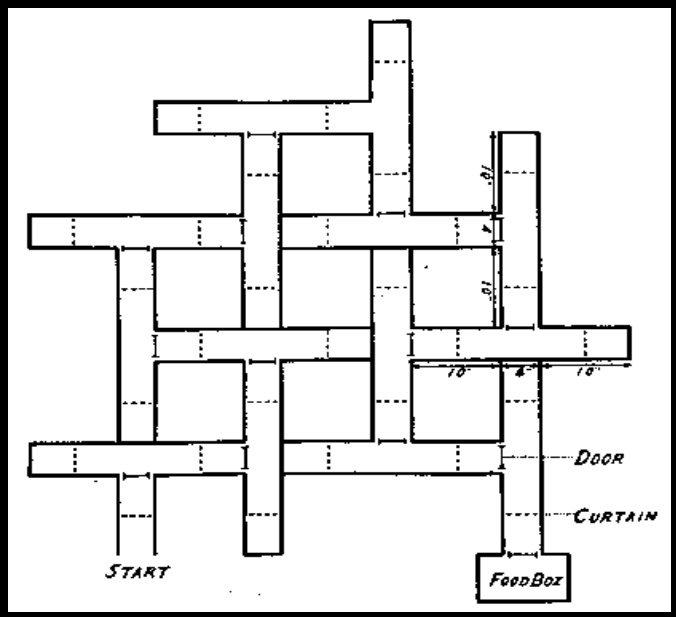
(1) First, there is a school of animal psychologists which believes that the maze behavior of rats is a matter of mere simple stimulus-response connections. Learning, according to them, consists in the strengthening of some of these connections and in the weakening of others. According to this 'stimulus-

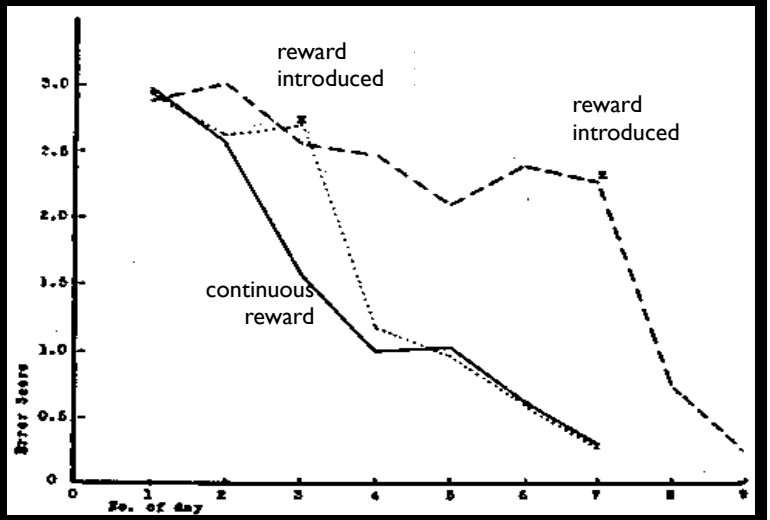
Maze task



- 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

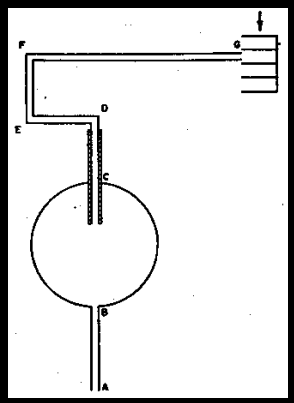




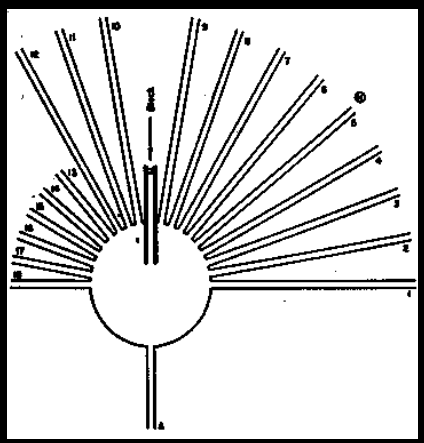
Blodgett (1929) 7

another example: shortcuts

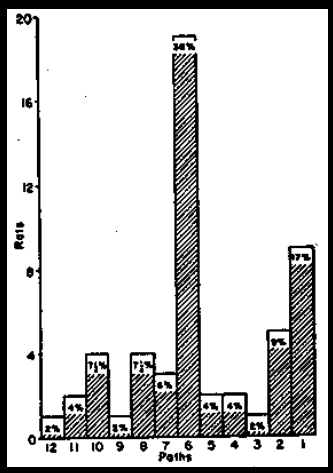
training:



test:



result:



Tolman et al (1946) 8

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 *rat*iocinations 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 their displaced hates ranges all the way from discrimination against minorities to world conflagrations.

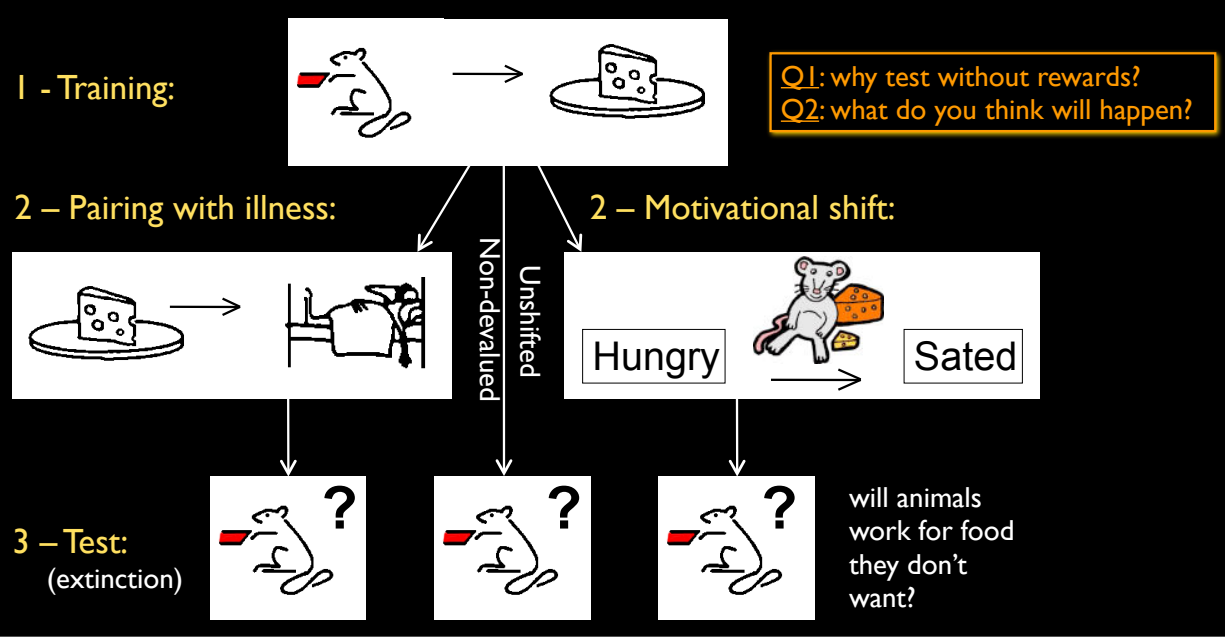
What is the name of Heaven and Psychology can we do about it? My only answer is to preach again the virtues of reason-of, that is, broad cognitive maps. And to suggest that the child-trainers and the world-planners of the future can only, if at all, bring about the presence of the required rationality (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

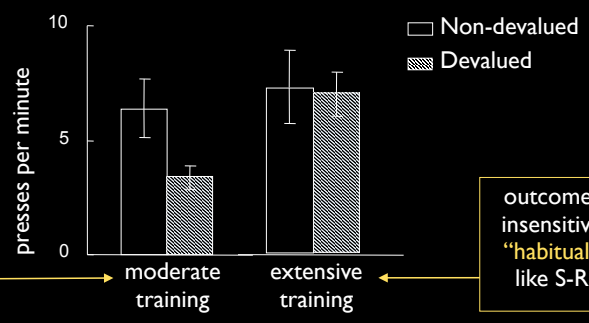
- S-R theory:
  - parsimonious - same theory for Pavlovian conditioning (CS associated with CR) and instrumental conditioning
  - Fits well with reinforcement learning (action values, actor/critic)
  - but: the critical contingency in instrumental conditioning is that of the action and the outcome…
-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

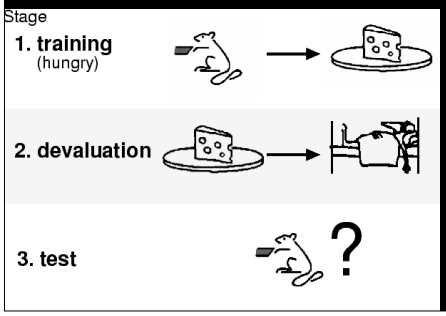


devaluation: results



outcome-sensitive "goal directed" like R-O

outcome-insensitive "habitual" like S-R



Animals will *sometimes* work for food they don't want!

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

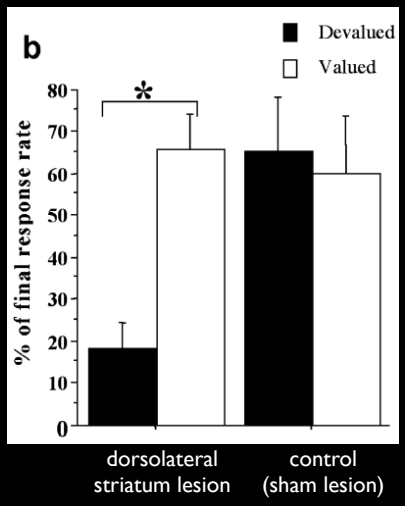
Holland (2004)

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

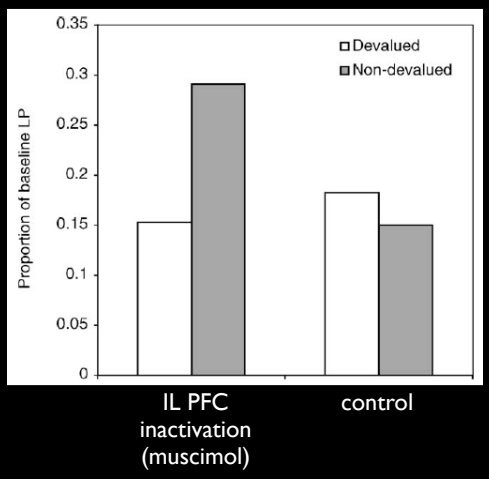
→ also treatments depleting dopamine in DLS

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

Yin et al (2004)

devaluation: results from lesions II

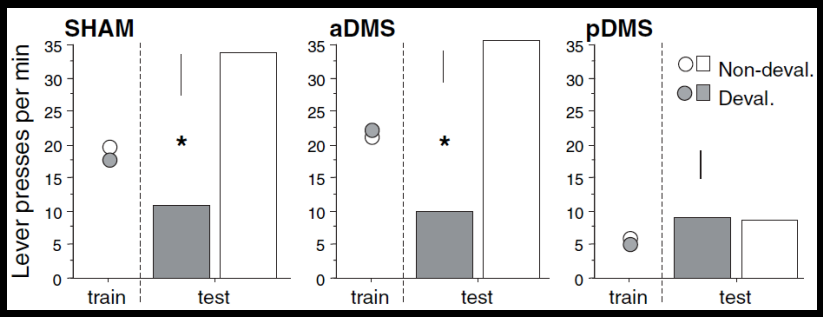
overtrained rats

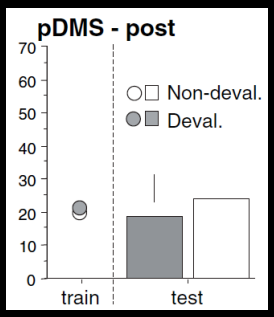


after habits have been formed, devaluation sensitivity can be *reinstated* by temporary inactivation of IL PFC

Coutureau & Killcross (2003) 17

devaluation: results from lesions III



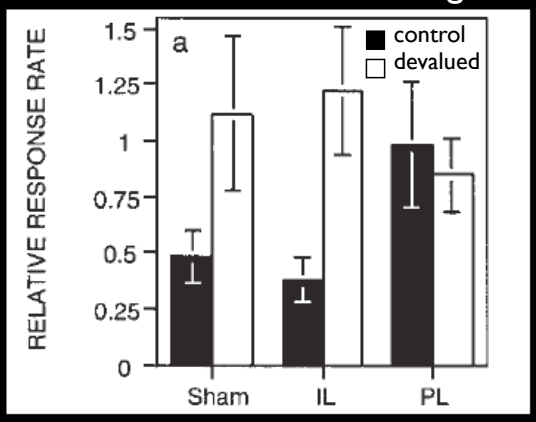


lesions of the pDMS cause animals to leverpress *habitually* even with only moderate training

Yin, Ostlund, Knowlton & Balleine (2005) 18

devaluation: results from lesions IV

moderate training

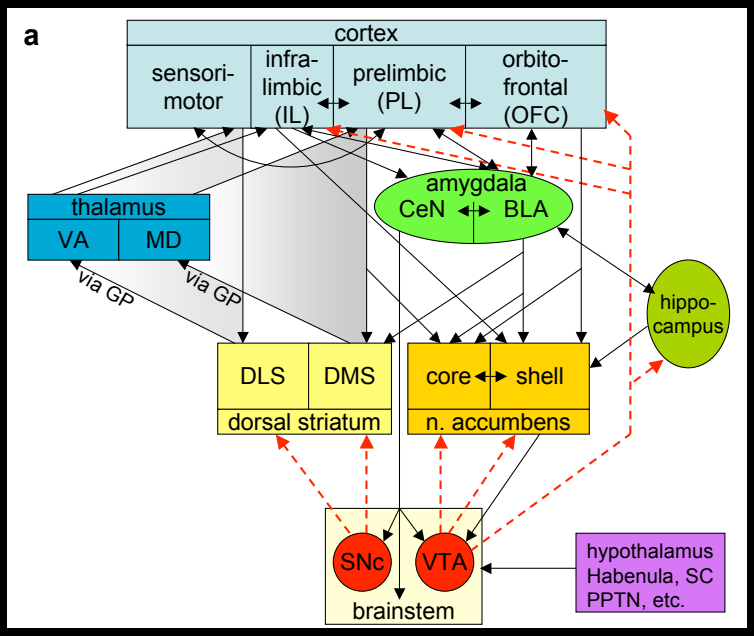


Prelimbic (PL) PFC lesions cause animals to leverpress *habitually* even with only moderate training

(also dorsomedial PFC and mediodorsal thalamus (same loop))

Killcross & Coutureau (2003)

Behavioral control in parallel loops



inspired by Balleine (2005)

what does all this mean?

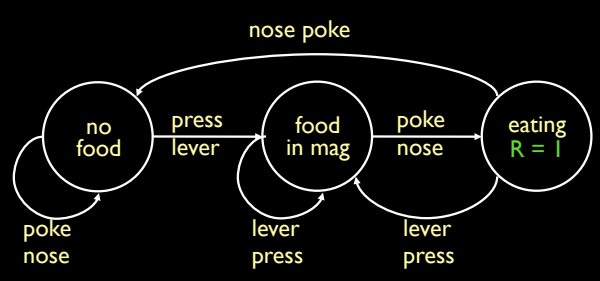
- The same action (leverpressing) can arise from two psychologically & neurally dissociable pathways
  1. moderately trained behavior is "goal-directed": dependent on outcome representation, like cognitive map (also associated with hippocampus - literal or abstract map of environment)
  2. overtrained 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

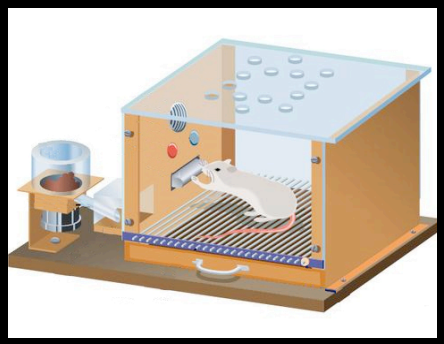
5 min break

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



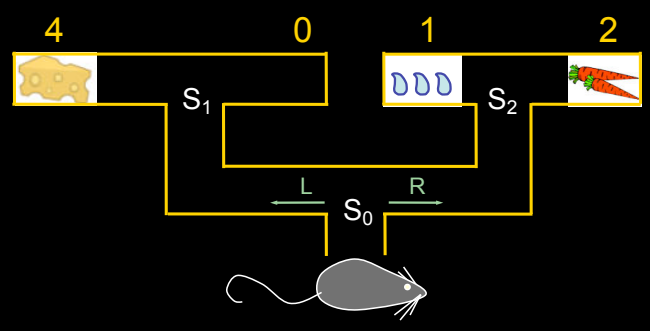


3 states: "no food", "food in mag", "eating"
2 actions: "press lever", "poke nose"
immediate reward is 1 in state "eating" and 0 otherwise

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

how can these be learned?

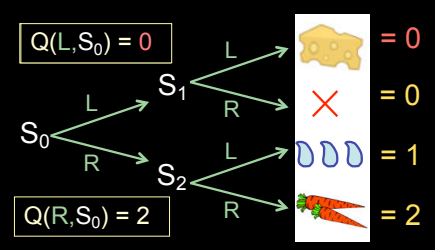
strategy I: "model-based" RL



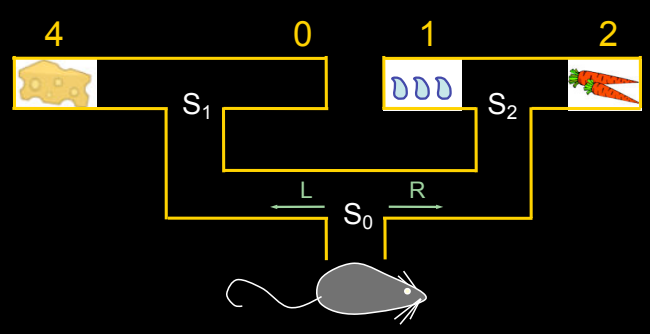
learn model of task through experience (= cognitive map)

compute Q values by "looking ahead" (mentally simulating) in the map

computationally costly, but also flexible (immediately sensitive to change)



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

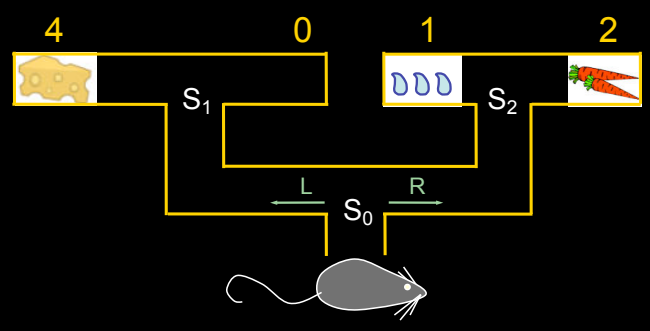
Stored:

$Q(S_0,L) = 4$
$Q(S_0,R) = 2$

$Q(S_1,L) = 4$
$Q(S_1,R) = 0$

$Q(S_2,L) = 1$
$Q(S_2,R) = 2$

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:

$Q(S_0,L) = 4$
$Q(S_0,R) = 2$

$Q(S_1,L) = 4$
$Q(S_1,R) = 0$

$Q(S_2,L) = 1$
$Q(S_2,R) = 2$

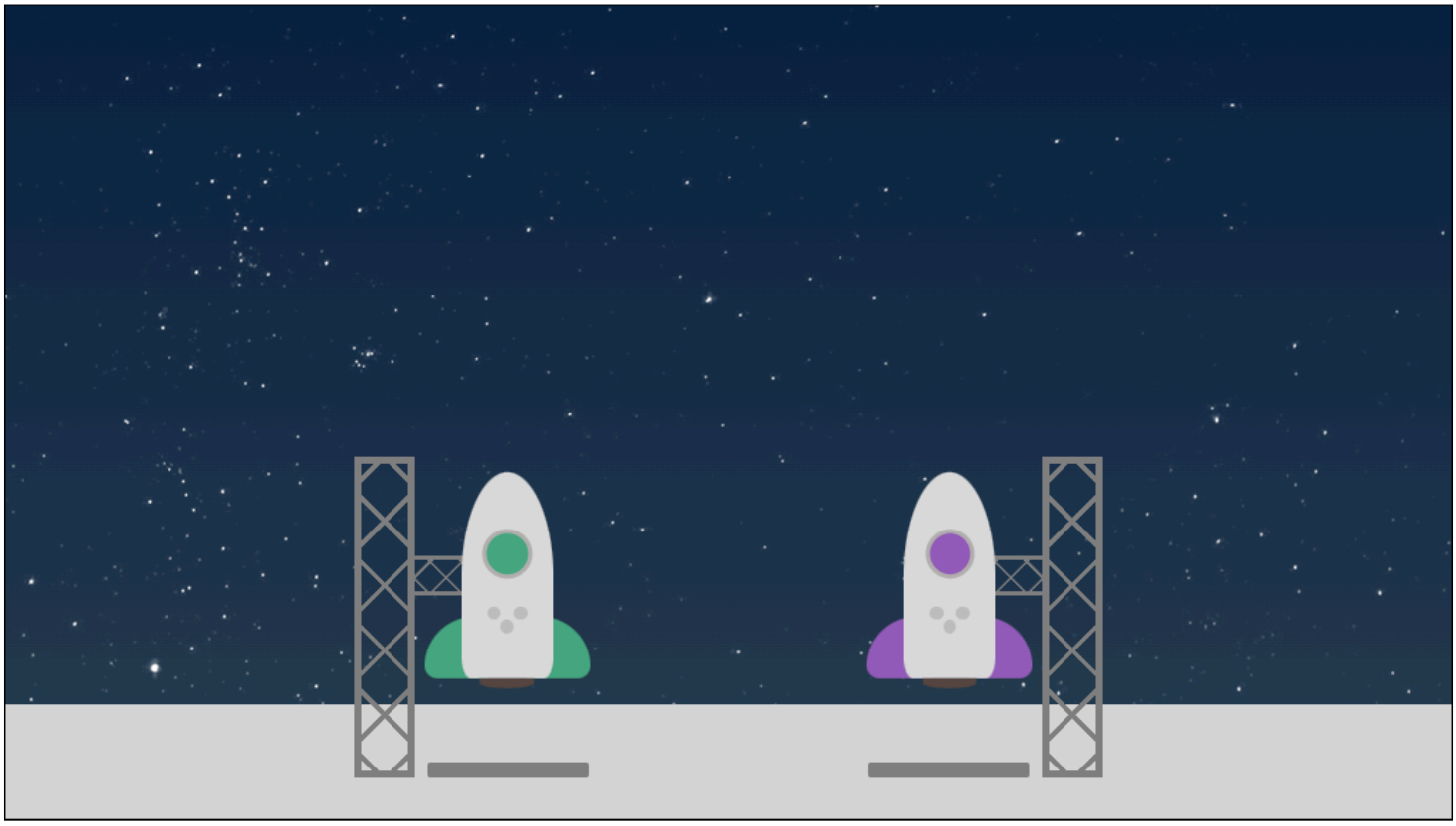
Why should the brain use two different strategies/controllers in parallel?

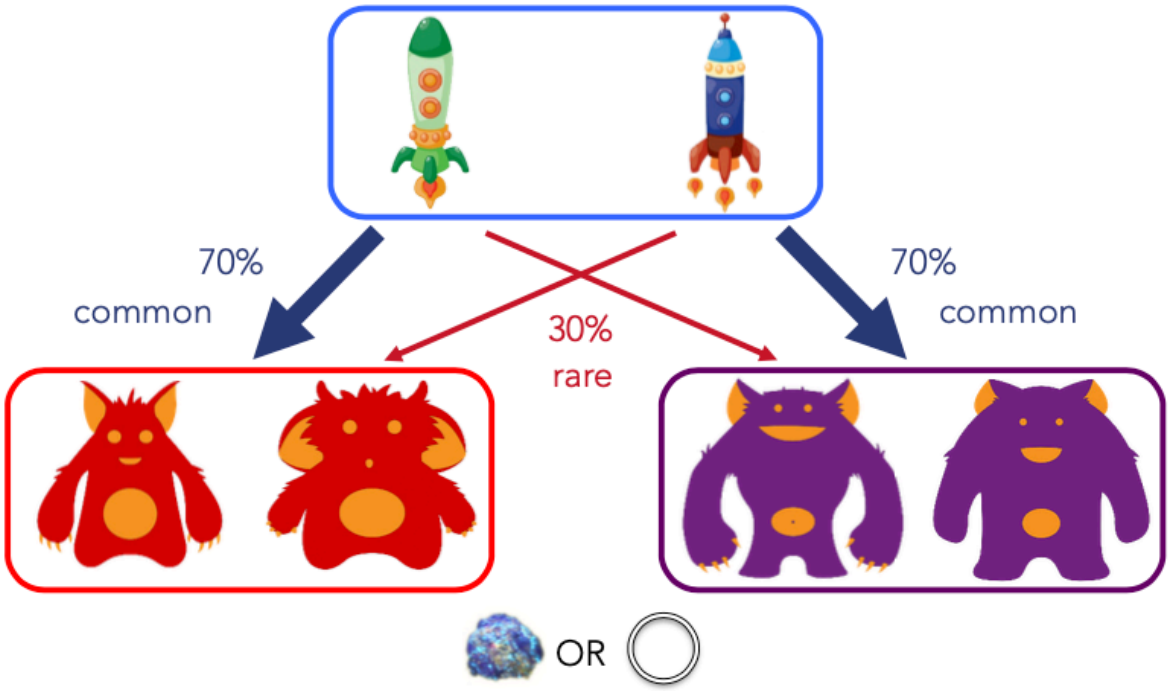
each system is best in different situations (use each one when it is most suitable/most accurate)

- 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

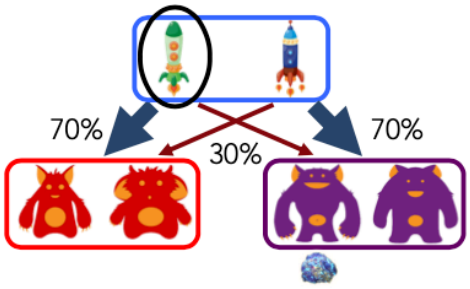
- instrumental behavior is not a simple unitary phenomenon: the same behavior can result from different *neural* and *computational* origins
- Model-based and model-free RL map on to goal-directed and habitual behavior, respectively
- Is this also true for humans? How do WE make choices?

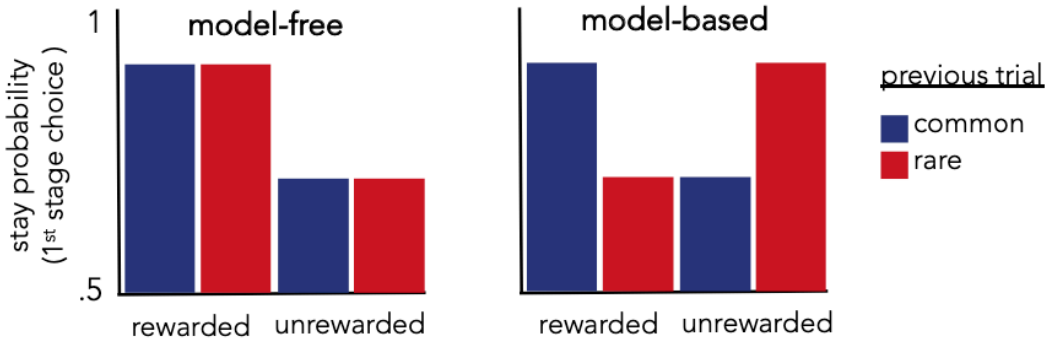




Slide from Cate Hartley (NYU), based on Daw, Gershman, Seymour, et al 2011

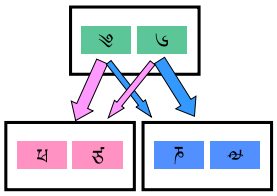
dissociating learning strategies

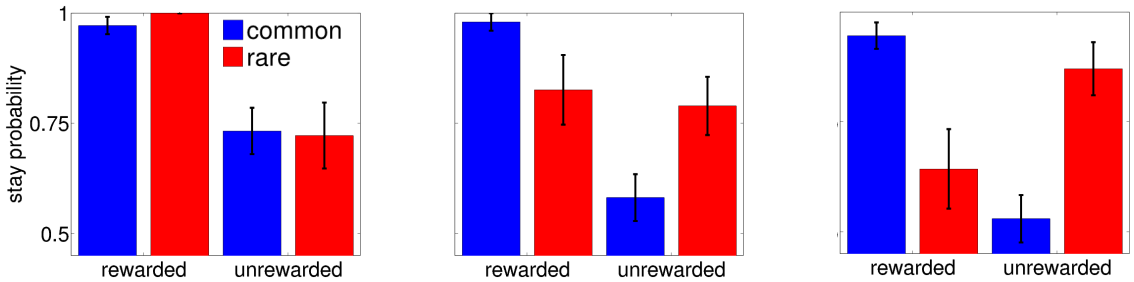




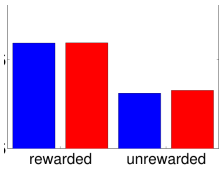
Slide from Cate Hartley, NYU

Three example participants

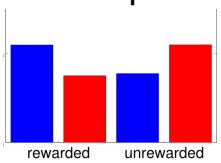




model free habits

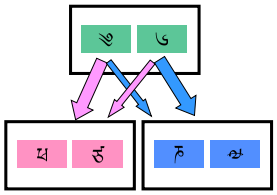


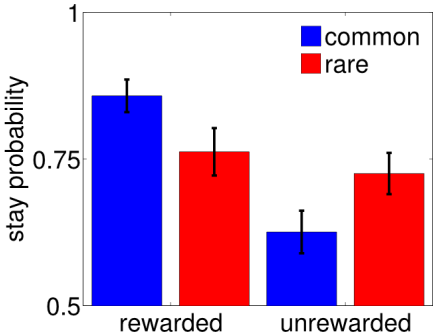
model based planning



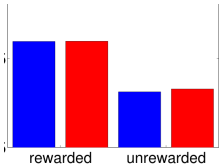
Daw, Gershman, Seymour, et al 2011

Group results

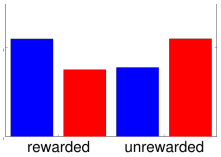




model free habits



model based planning



Daw, Gershman, Seymour, et al 2011

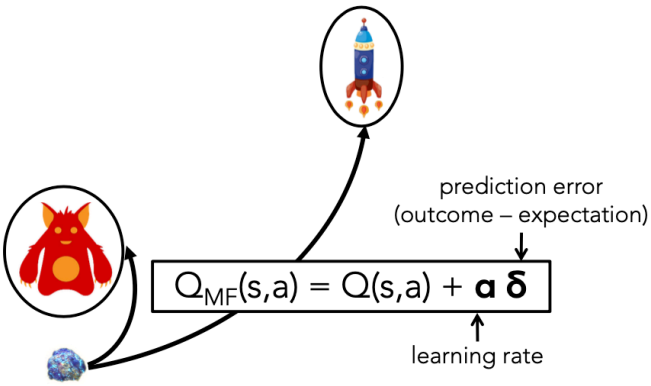
use of habits varies situationally:

- overtraining
- stress
- distraction
- motivation, fatigue, belief

and between individuals:

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

model-free value estimation



model-based value estimation

transition probability best choice

$$Q_{MB}(s_1,a_i) = P(s_2 \mid a_i) * \max(Q(s_2,a)) + P(s_3 \mid a_i) * \max(Q(s_3,a))$$

Total value: $Q(S_1, a) = (1 - w) \cdot Q_{MF}(S_1, a) + w \cdot Q_{MB}(S_1, a)$

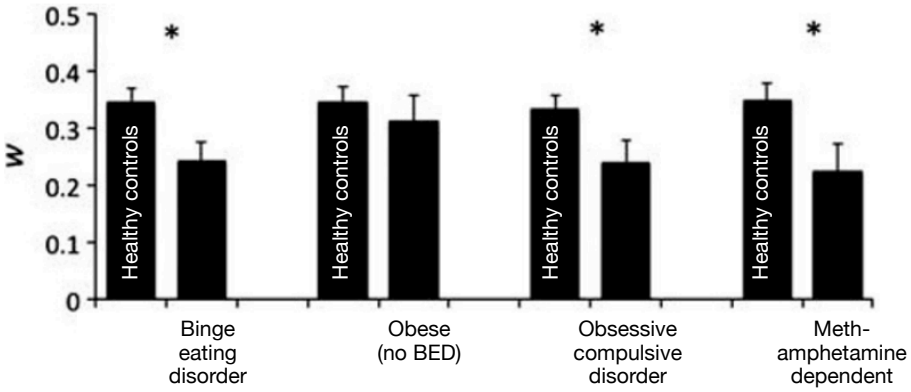
Slide from Cate Hartley, NYU

ORIGINAL ARTICLE

Disorders of compulsivity: a common bias towards learning habits

V Voon[1,2,3], K Derbyshire[4,5], C Rück[6], MA Irvine[1], Y Worbe[2], J Enander[6], LRN Schreiber[4], C Gillan[2,7], NA Fineberg[8], BJ Sahakian[1,2], TW Robbins[2,7], NA Harrison[9], J Wood[3], ND Daw[10], P Dayan[11], JE Grant[5] and ET Bullmore[1,2,3,12]

Valerie Voon



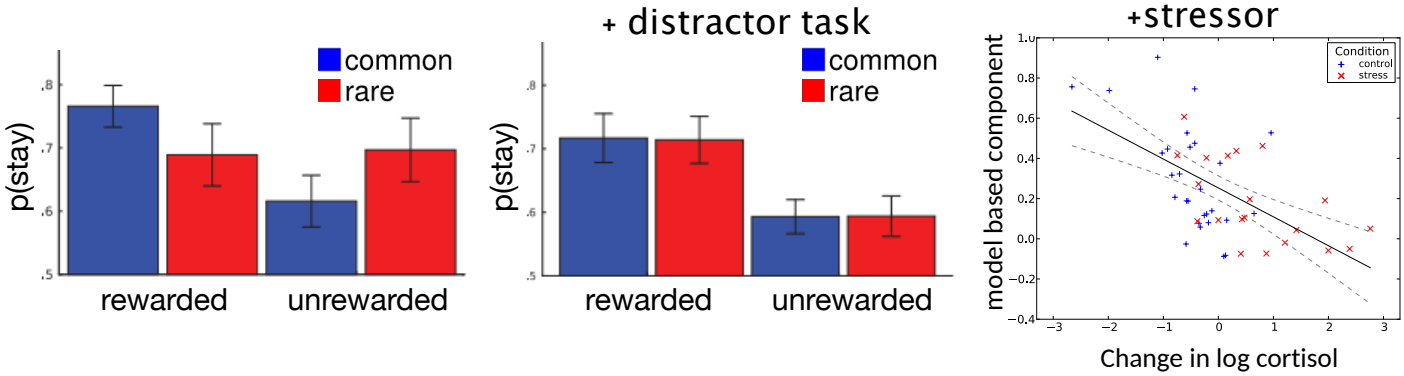
Voon et al., Molecular Psychiatry, 2014

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

A. Ross Otto[1], Samuel J. Gershman[2], Arthur B. Markman[1], and Nathaniel D. Daw[3]

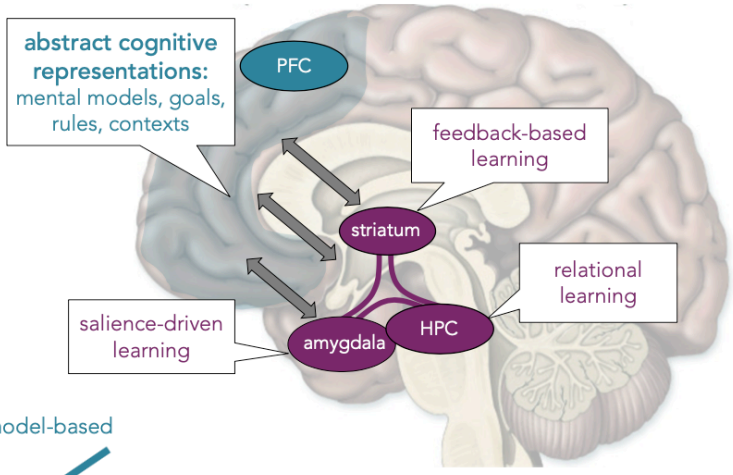
Working-memory capacity protects model-based learning from stress

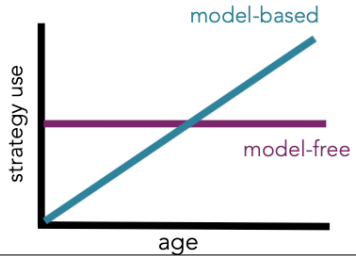
A. Ross Otto[a,1], Candace M. Raio[b], Alice Chiang[b], Elizabeth A. Phelps[a,b,c], and Nathaniel D. Daw[a,b]



(Otto et al Psych Science, 2013; Otto et al PNAS, 2013)

Developmental changes in learning strategies

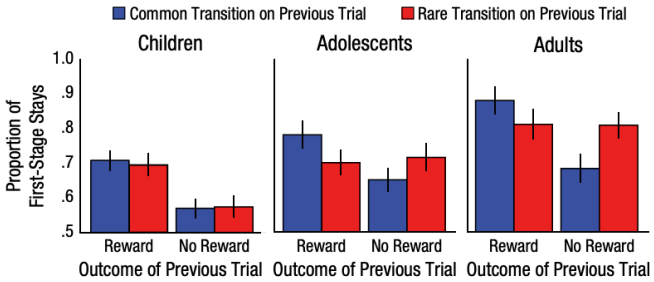


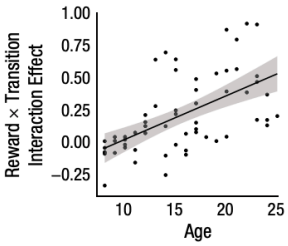


Slide from Cate Hartley, NYU

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

Johannes H. Decker[1], A. Ross Otto[2], Nathaniel D. Daw[3,4], and Catherine A. Hartley[1]





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