ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
CAMPUS DI CESENA

From reinforcement learning to decision-making: goals and habits in the brain

Cognition and Neuroscience
Academic year 2024/2025

Francesca Starita

francesca.starita2@unibo.it

Learning objectives

- **Outline the four generations of research in decision-making**: cognitive maps vs. stimulus-response (Generation 0), goal-directed vs. habitual actions (Generation 1 in experimental psychology and Generation 2 in cognitive neuroscience), and model-based vs. model-free computational analyses (Generation 3).
- **Compare and contrast stimulus-response theories and field theories (cognitive maps)** in the context of how animals learn to solve mazes (Generation 0).
- **Describe Tolman's maze experiments** and their contribution to the understanding of latent learning and cognitive maps.
- **Define latent learning**
- **Explain the concept of a cognitive map**
- **Understand the implications of Generation 0 studies**

Learning objectives

- **Explain how cognitive maps were operationalized for learning to choose appropriate actions in nonspatial domains**, leading to the concepts of goal-directed and habitual behavior (Generation 1).
- **Differentiate between goal-directed and habitual behavior/actions** based on their characteristics, such as deliberation, computational cost, flexibility, and influence of outcome value.
- **Identify the two key criteria for an action to be considered goal-directed**: knowledge of the response-outcome relationship and the motivational relevance of the outcome.
- **Describe the experimental methods used to test whether a behavior is goal-directed**, including reinforcer devaluation and contingency degradation.
- **Explain how extensive training (overtraining) can lead to habitual behavior**.
- **Understand the neural dissociation between goal-directed and habitual behavior in the rodent striatum**, with the dorsomedial striatum (DMS) supporting goal-directed behavior and the dorsolateral striatum (DLS) supporting habitual behavior.

Learning objectives

- **Describe how animal paradigms were adapted to study goal-directed and habitual actions in the human brain** using fMRI (Generation 2).
- **Explain the methods and results of human studies** investigating the neural substrates of goal-directed behavior, particularly the role of the medial orbitofrontal cortex (OFC).
- **Explain the methods and results of human studies** investigating the neural substrates of habitual behavior, particularly the role of the posterior dorsolateral striatum.
- **Understand the computational formalization of goal-directed actions as model-based and habitual actions as model-free** (Generation 3).
- **Differentiate between model-based and model-free algorithms** in terms of how they select actions and respond to changes in the environment.
- **Describe sequential two-choice Markov decision tasks** and their use in studying the influence of model-free and model-based control on behavior.
- **Discuss the idea of an integrated computational and neural architecture** for decision-making, where model-based and model-free systems interact.

Multiple systems contribute to learning and controlling behavior in animals

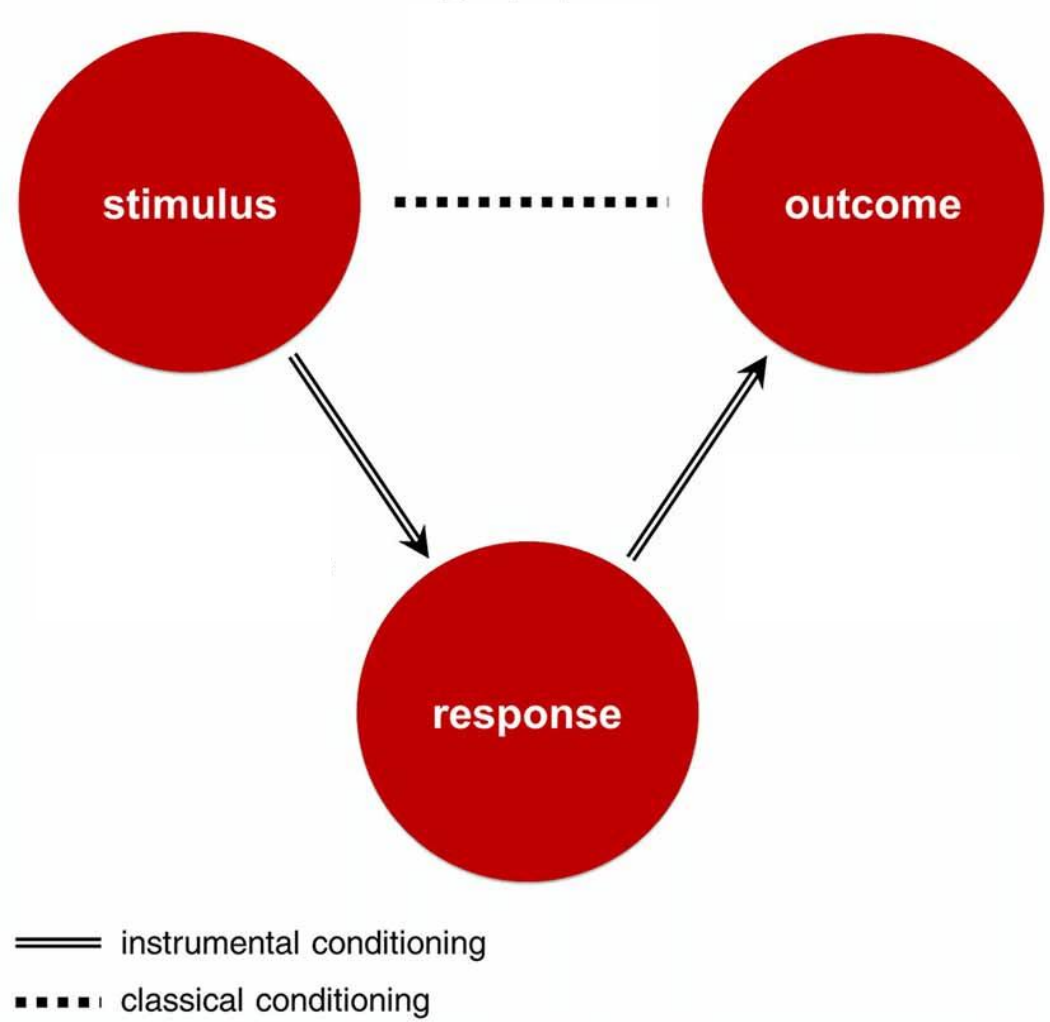
Three learning systems enable organisms to draw on previous experience to **make predictions** about the world and to **select behaviors** appropriate to those predictions:

1. a **Pavlovian system** that learns to predict biologically significant events so as to trigger appropriate responses;

Instrumental system that comprises

2. a **habitual system** that learns to repeat previously successful actions;
3. a **goal-directed system** that evaluates actions on the basis of their specific anticipated consequences.

Predictions are for control



Instrumental learning (or operant conditioning) involves associating an action with an outcome

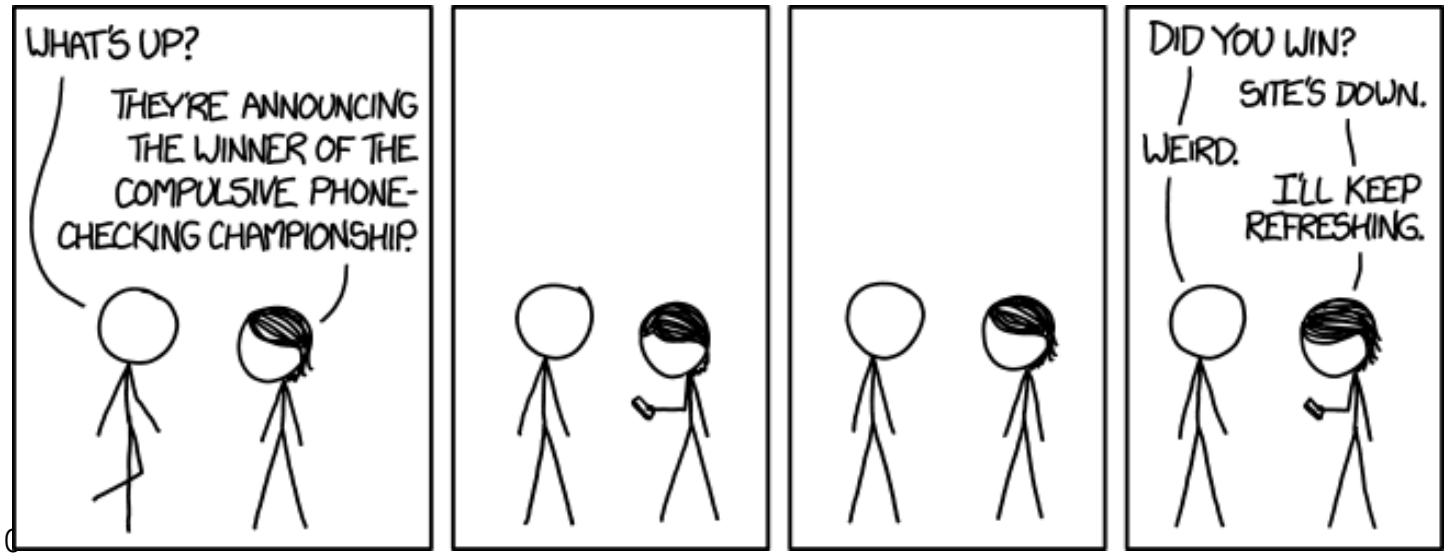
Thorndike's Law of effect

"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**. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond." (Thorndike, 1911)

We act to produce outcomes that are desirable or to avoid those that are harmful or aversive

But how do we select (decide) which is the appropriate action to take?

- Are we flexible in the actions we take?
- Do we choose the action to take with the goal in mind or do we automatically select actions based on previous (rewarded) experiences?
- Are our choices directed by the goal we want to achieve or are they automatically/habitually triggered based on our past (rewarded) experiences?



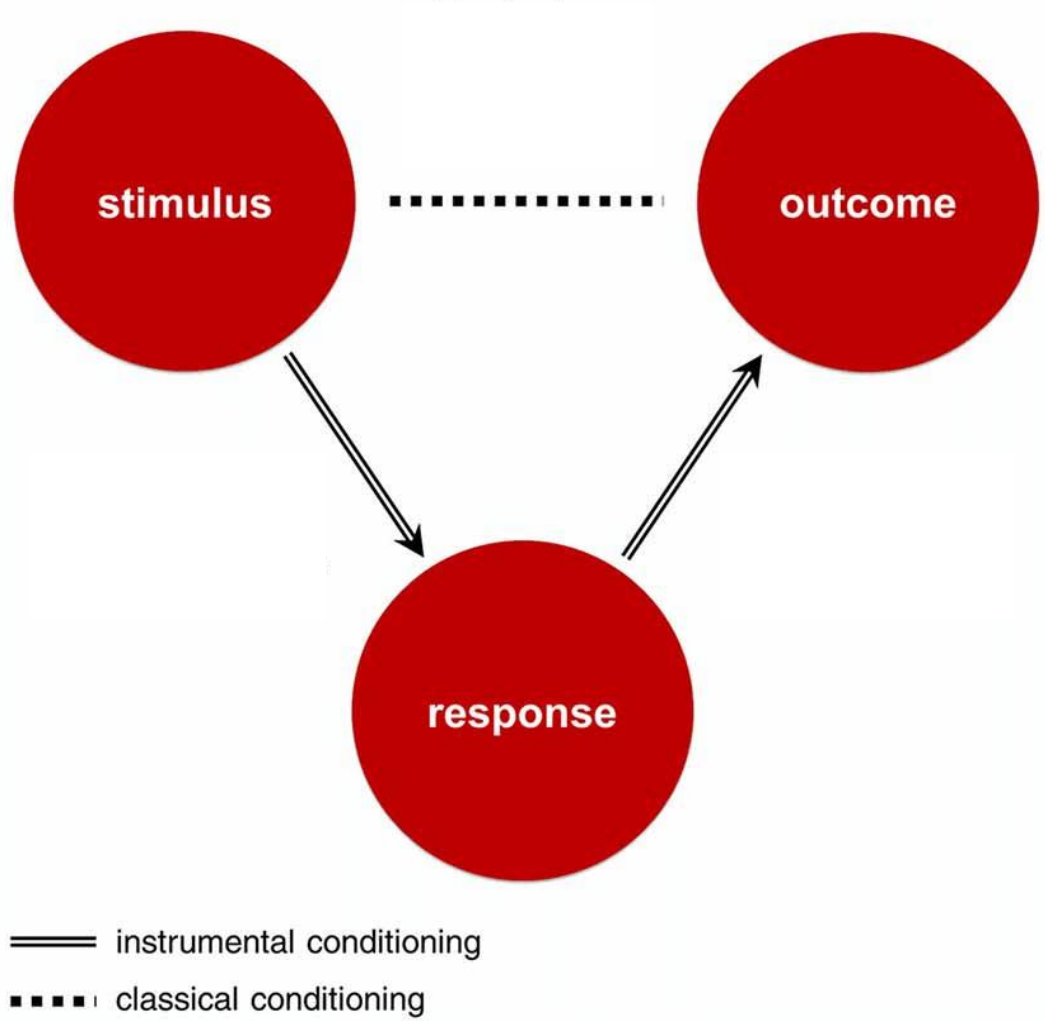
Multiple systems contribute to learning and controlling behavior in animals

- Are we flexible in the actions we take?
- Do we choose the action to take with the goal in mind or do we automatically select actions based on previous (rewarded) experiences?
- Are our choices directed by the goal we want to achieve or are they automatically/habitually triggered based on our past (rewarded) experiences?

Instrumental system that comprises

2. a **habitual system** that learns to repeat previously successful actions;
3. a **goal-directed system** that evaluates actions on the basis of their specific anticipated consequences.

Predictions are for control



Cell PRESS

Neuron
Review

Goals and Habits in the Brain

Ray J. Dolan[1,*] and Peter Dayan[2]
[1]Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London WC1 3BG, UK
[2]Gatsby Computational Neuroscience Unit, University College London, London WC1N 3AR, UK
*Correspondence: r.dolan@ucl.ac.uk
http://dx.doi.org/10.1016/j.neuron.2013.09.007

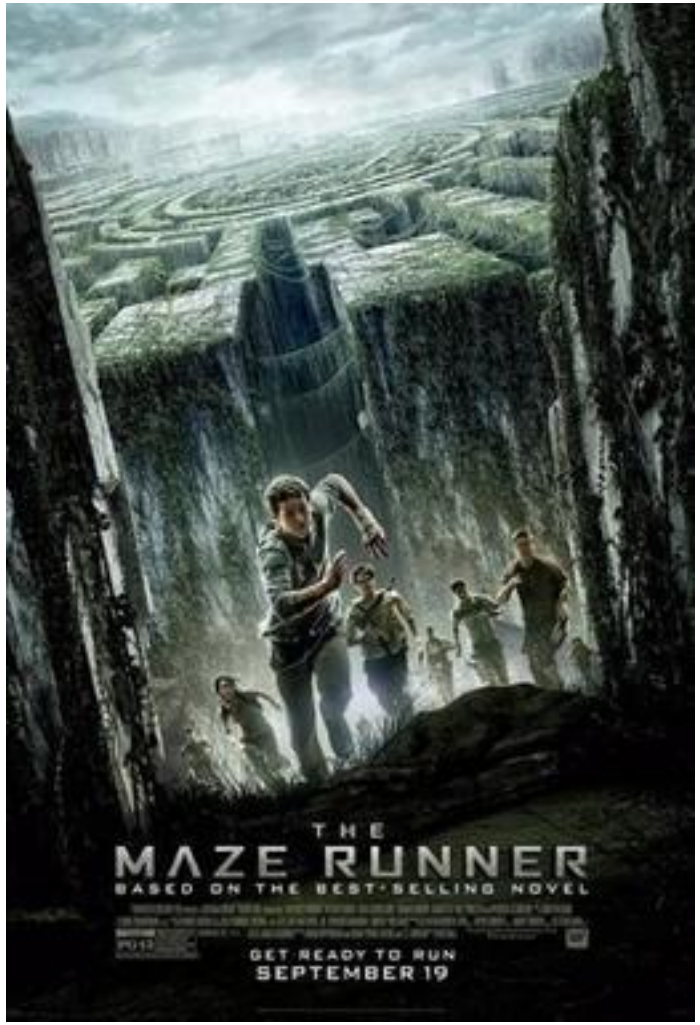
- **Generation 0**: cognitive maps vs stimulus-response [experimental psychology]
- **Generation 1**: goal-directed vs habitual actions [experimental psychology]
- **Generation 2**: goal-directed vs habitual actions in the human brain [cognitive neuroscience]
- **Generation 3**: model-based vs model-free computational analyses [computational neuroscience]

Generation 0

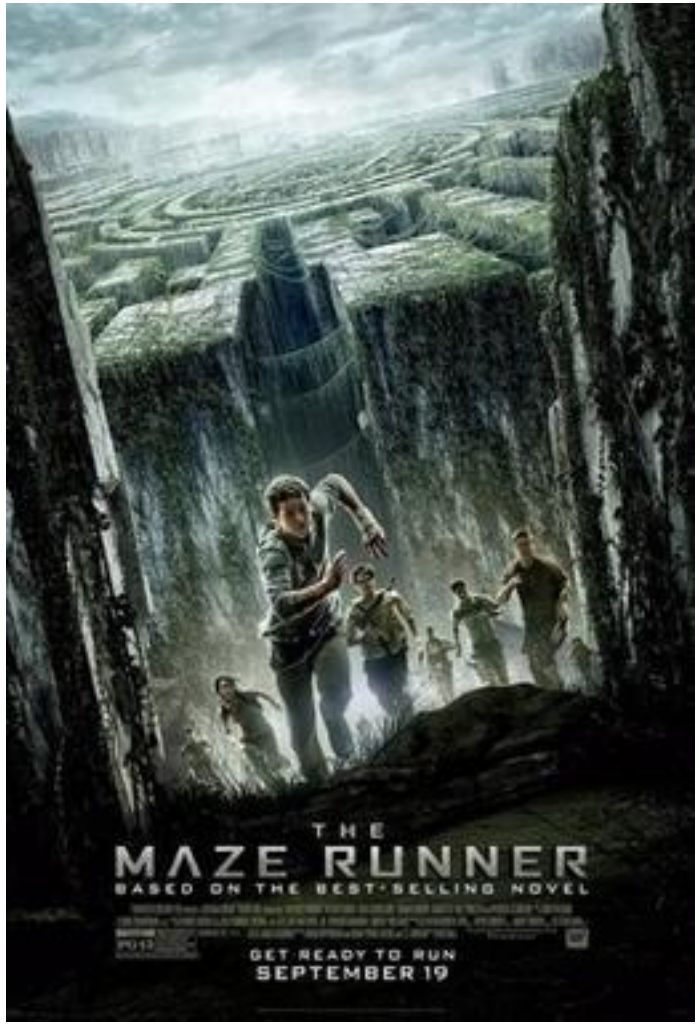
cognitive maps vs stimulus-response

[experimental psychology]

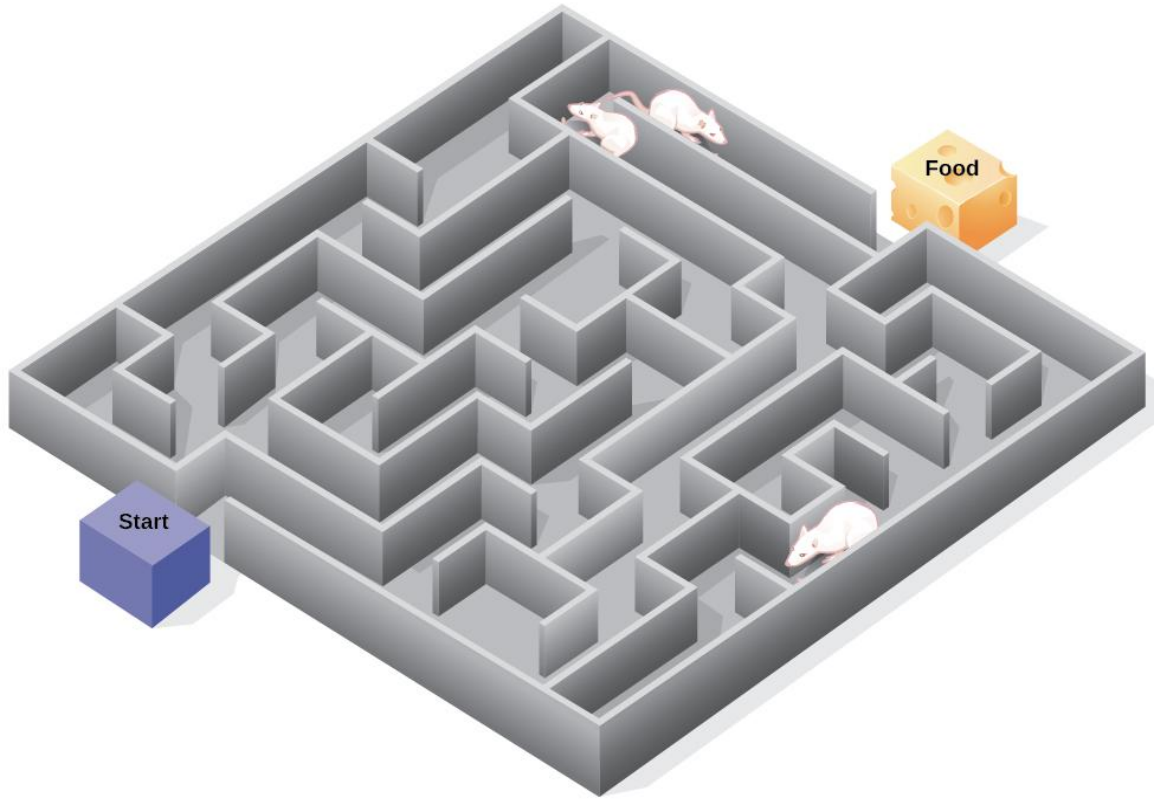
Generation 0: cognitive maps vs stimulus-response



Generation 0: cognitive maps vs stimulus-response



How does the animal learn to solve the maze?

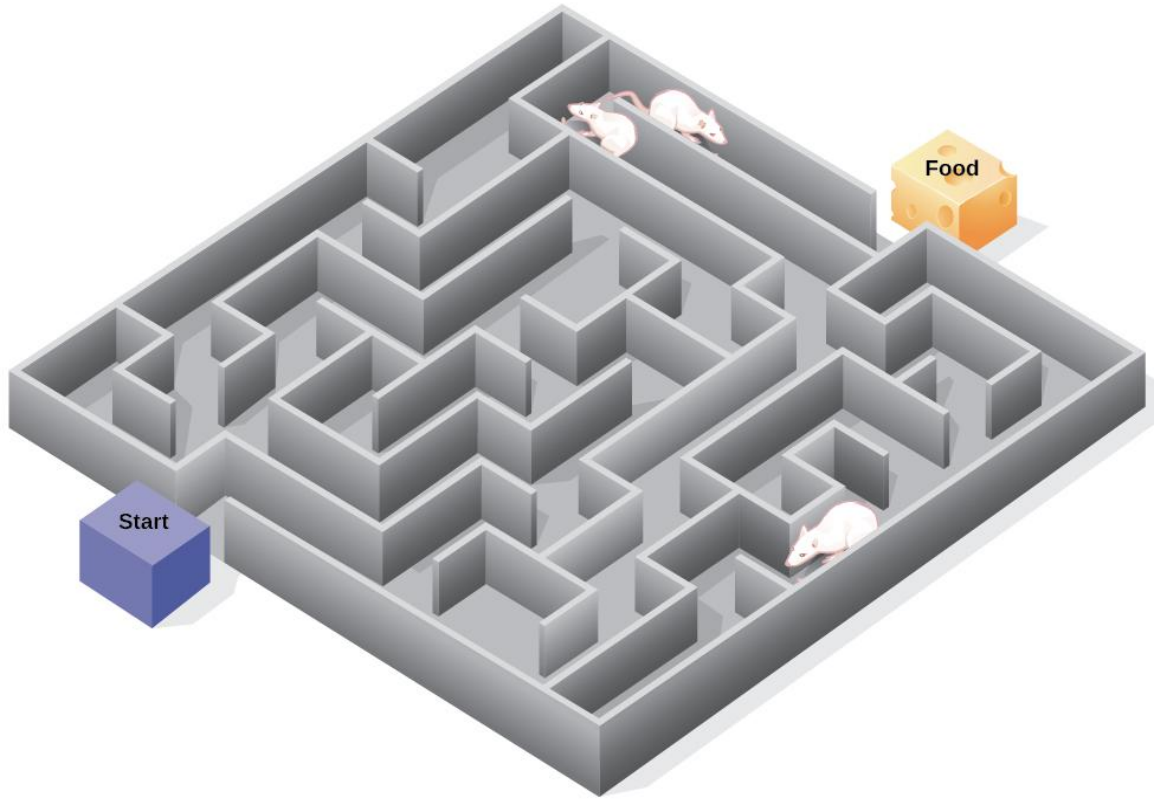


Generation 0: cognitive maps vs stimulus-response

Stimulus-response (S-R) theories

- the bedrock of psychology in the first half of the 20th century
- Solving the maze is a matter of **individual stimulus-response one-to-one connections**
- Learning depends on strengthening of some connections and weakening of others
- the animal helplessly responds to a succession of external and internal stimuli that callout the actions to take (e.g. turnings) and the like that follows

How does the animal learn to solve the maze?

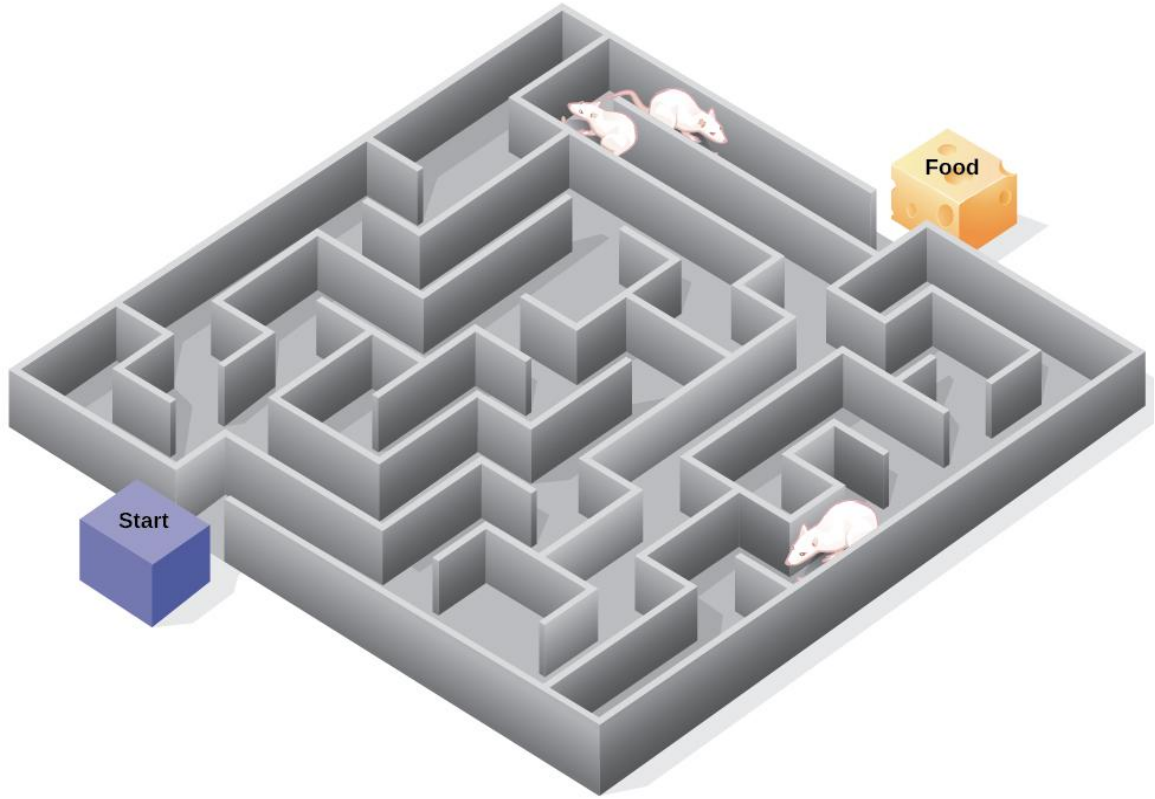


Generation 0: cognitive maps vs stimulus-response

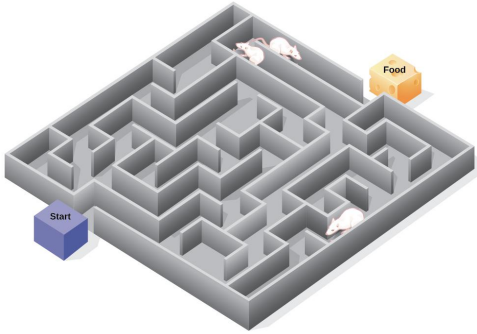
How does the animal learn to solve the maze?

Field theories

- Solving the maze is a matter of **creating a mental/cognitive map that includes multiple sets of connections**
- The mental map then guides what responses the animal will perform
- The mental map acts as a representational template that enables an animal to find the best possible action at a particular state



Generation 0: cognitive maps vs stimulus-response



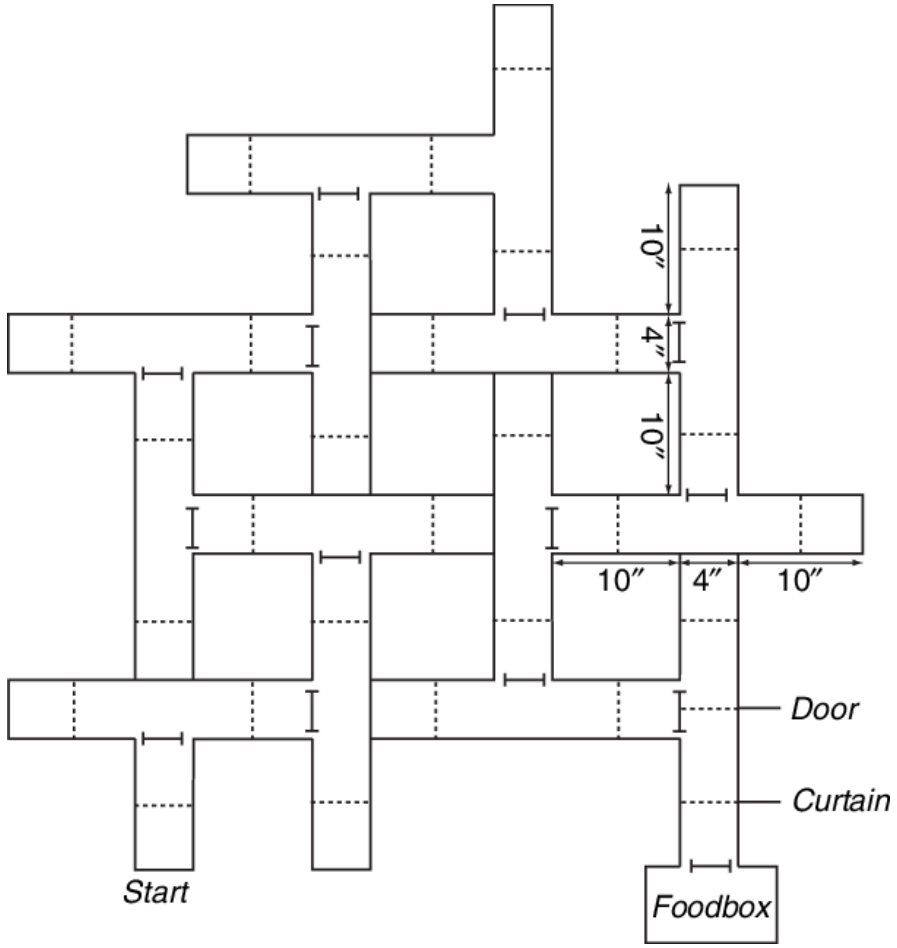
Tolman's maze

The maze had lots of doors and curtains to make it difficult for the rats to master.

Doors swung both directions, which prevented the rat from seeing most of the junctions as it approached. This forced the rat to go through the door to discover what was on the other side.

Curtains hung down and prevented the rat from getting a long distance perspective and it also meant that they could not see a wall at the end of a wrong turn until they had already made a choice and moved in that direction.

The rat was always in a small area, unable to see beyond the next door or curtain, so learning the maze was a formidable task.



Maze used by Tolman and Honzik (1930) to study latent learning in rats. From Tolman EC (1948) Cognitive maps in rats and men. Psychol. Rev. 55: 189-208

Generation 0: cognitive maps vs stimulus-response

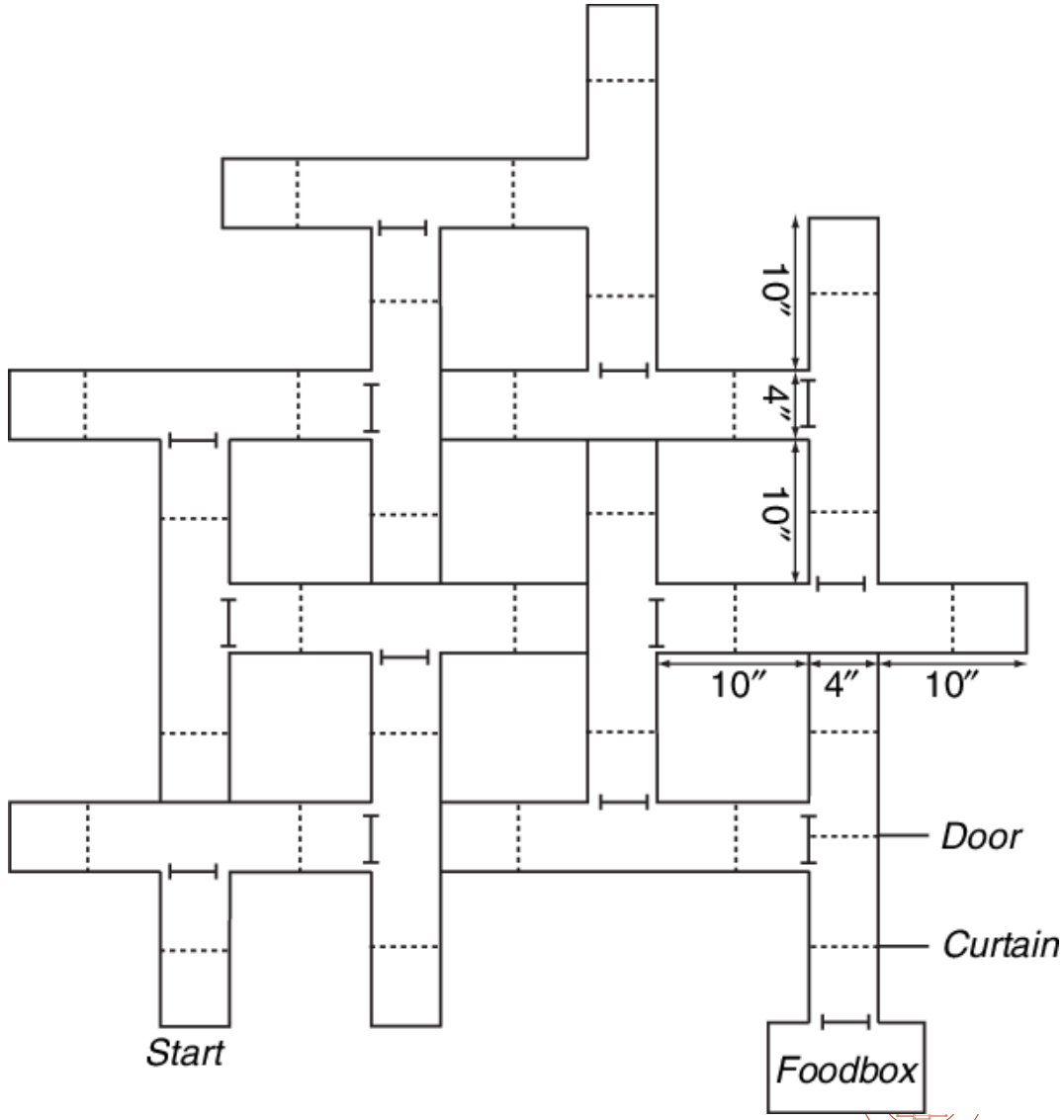
Experiment

Hungry rats have to find their way out through a maze

Group 1: no reward for solving the maze

Group 2: food reward for solving the maze

Which group completed the maze faster?



Maze used by Tolman and Honzik (1930) to study latent learning in rats. From Tolman EC (1948) Cognitive maps in rats and men. Psychol. Rev. 55: 189-208

Generation 0: cognitive maps vs stimulus-response

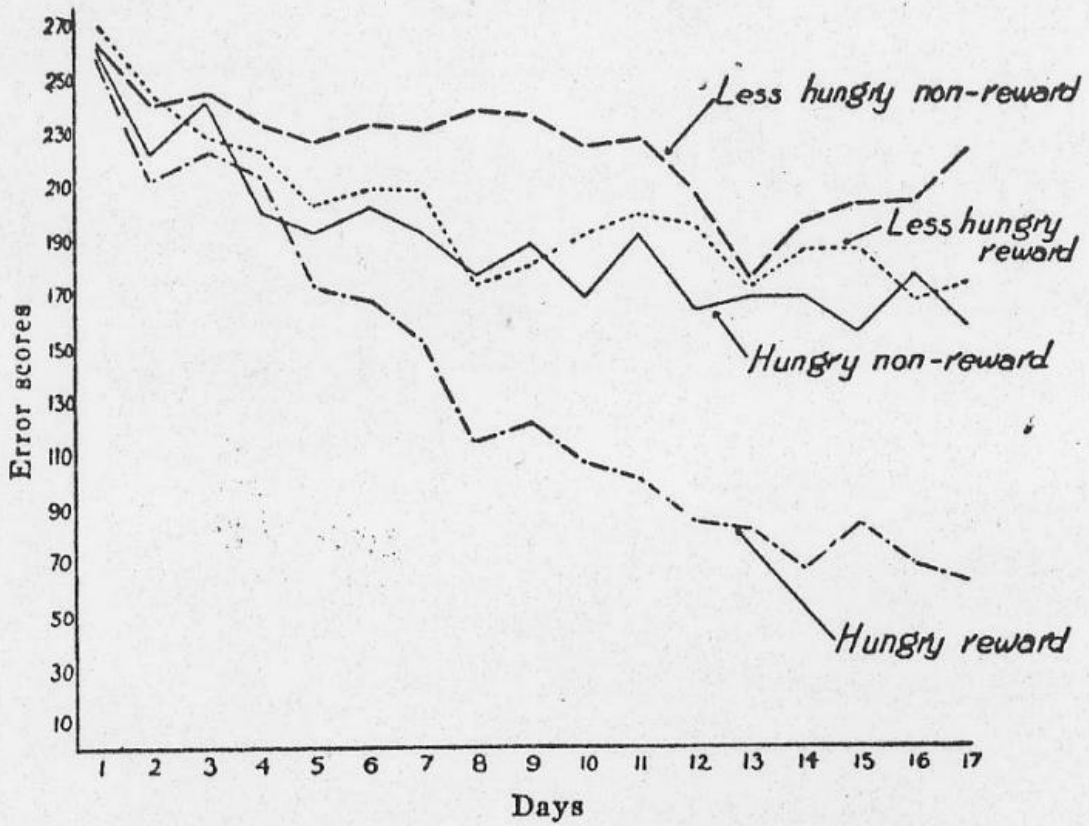
Experiment

Hungry rats have to find their way out through a maze

Group 1: no reward for solving the maze

Group 2: food reward for solving the maze

Which group completed the maze faster?



Error curves for four groups, 36 rats.

FIG. 3

(From E. C. Tolman and C. H. Honzik, Degrees of hunger, reward and non-reward, and maze learning in rats. *Univ. Calif. Publ. Psychol.*, 1930, 4, No. 16, p. 246. A maze identical with the alley maze shown in Fig. 1 was used.)

Generation 0: cognitive maps vs stimulus-response

Experiment

Hungry rats have to find their way out through a maze

Group 1: no reward for solving the maze

Group 2: food reward for solving the maze

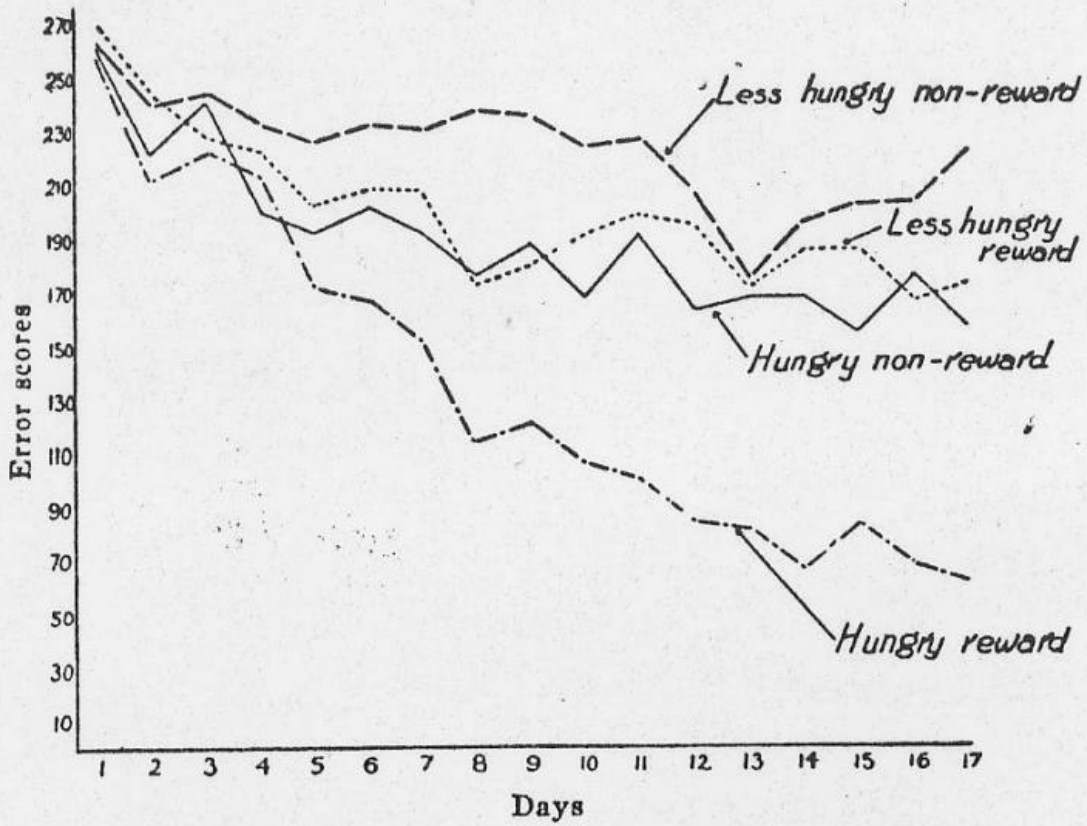
Which group completed the maze faster?

Group 2

Reward & motivation are crucial to learn

But...

Have the other groups learnt anything at all?



Error curves for four groups, 36 rats.

FIG. 3

(From E. C. Tolman and C. H. Honzik, Degrees of hunger, reward and non-reward, and maze learning in rats. *Univ. Calif. Publ. Psychol.*, 1930, 4, No. 16, p. 246. A maze identical with the alley maze shown in Fig. 1 was used.)

Generation 0: cognitive maps vs stimulus-response

Experiment

Hungry rats have to find their way out through a maze

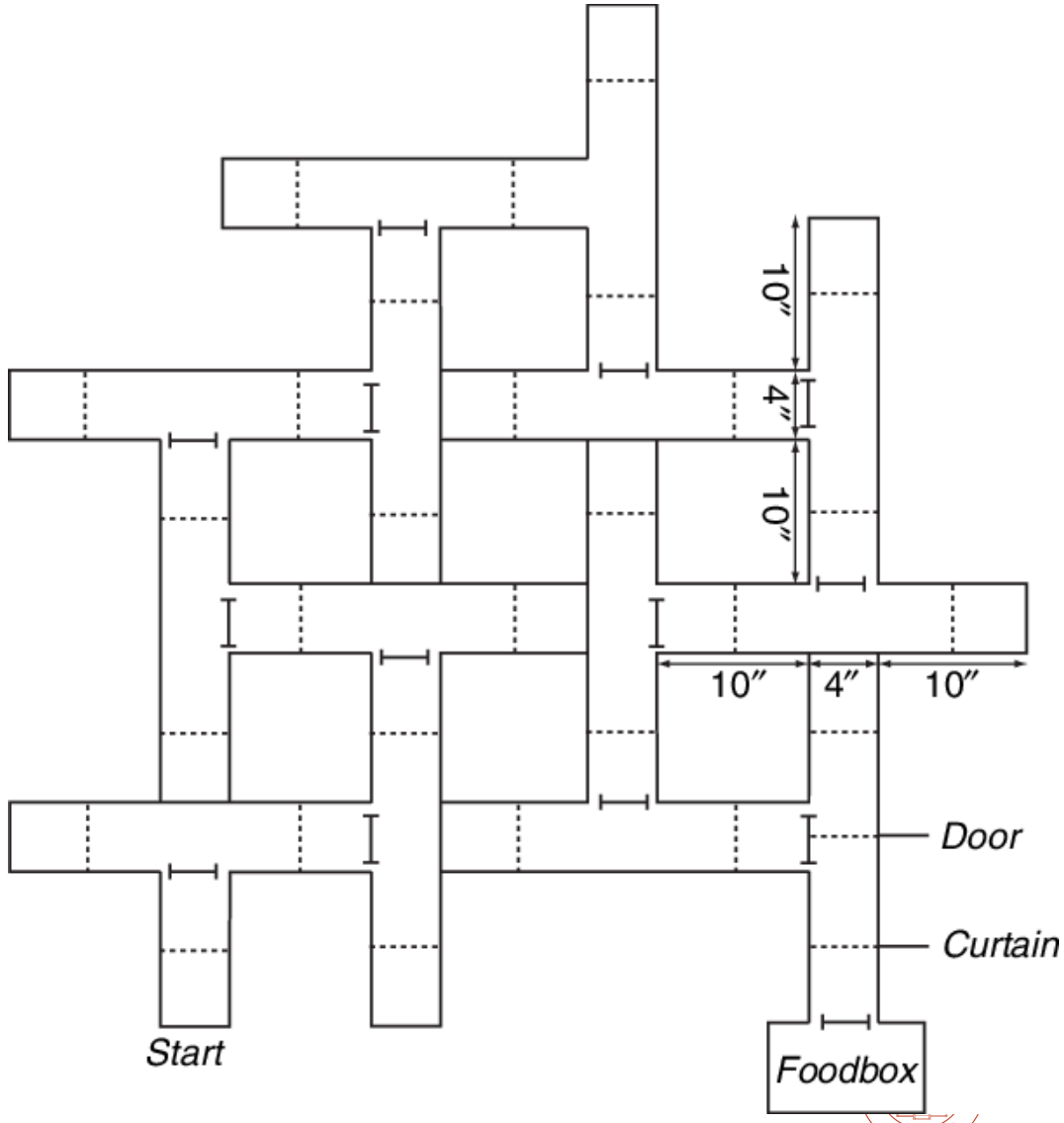
Group 1: no reward for solving the maze

Group 2: food reward for solving the maze

Group3: food reward for solving the maze provided only at day 11

What do you think happens to performance?

Note: the S-R theories argues that no learning occurs when there is no reward



Maze used by Tolman and Honzik (1930) to study latent learning in rats. From Tolman EC (1948) Cognitive maps in rats and men. Psychol. Rev. 55: 189-208

Generation 0: cognitive maps vs stimulus-response

Experiment

Hungry rats have to find their way out through a maze

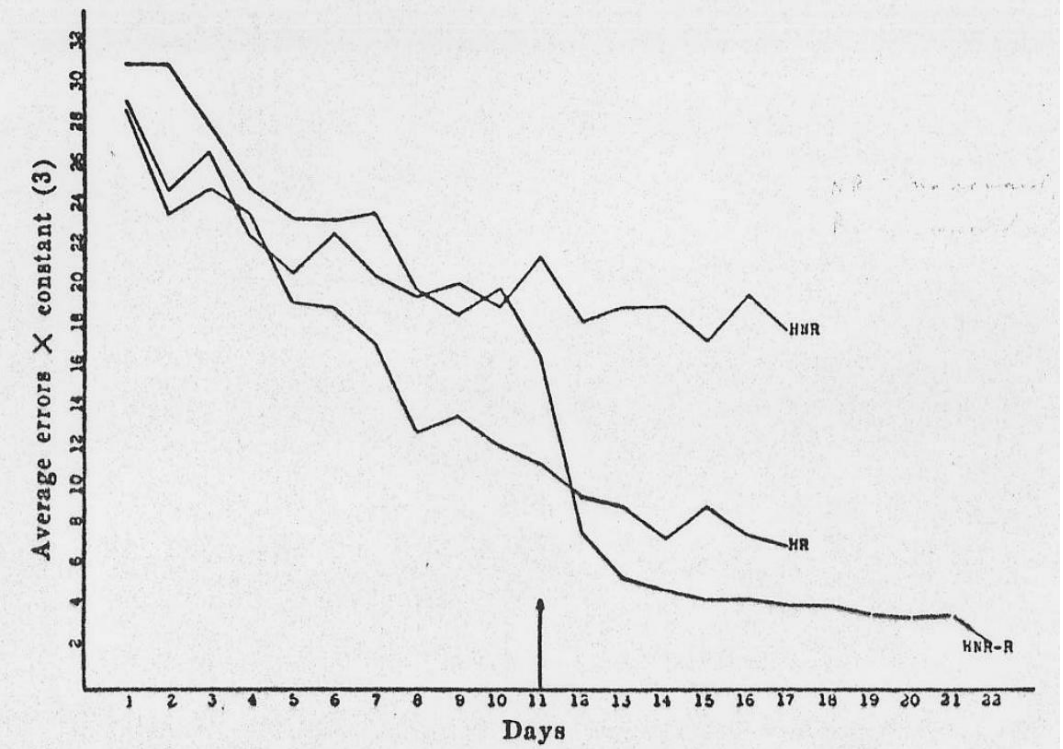
Group 1: no reward for solving the maze

Group 2: food reward for solving the maze

Group3: food reward for solving the maze provided only at day 11

Results:

- As soon as the rats in group 3 was given the food, they were able to find their way through the maze quickly, just as quickly as the comparison group, which had been rewarded with food all along



Error curves for HR, HNR, and HNR–R

FIG. 6

(From E. C. Tolman and C. H. Honzik, Introduction and removal of reward, and maze performance in rats. *Univ. Calif. Publ. Psychol.*, 1930, 4, No. 19, p. 267.)

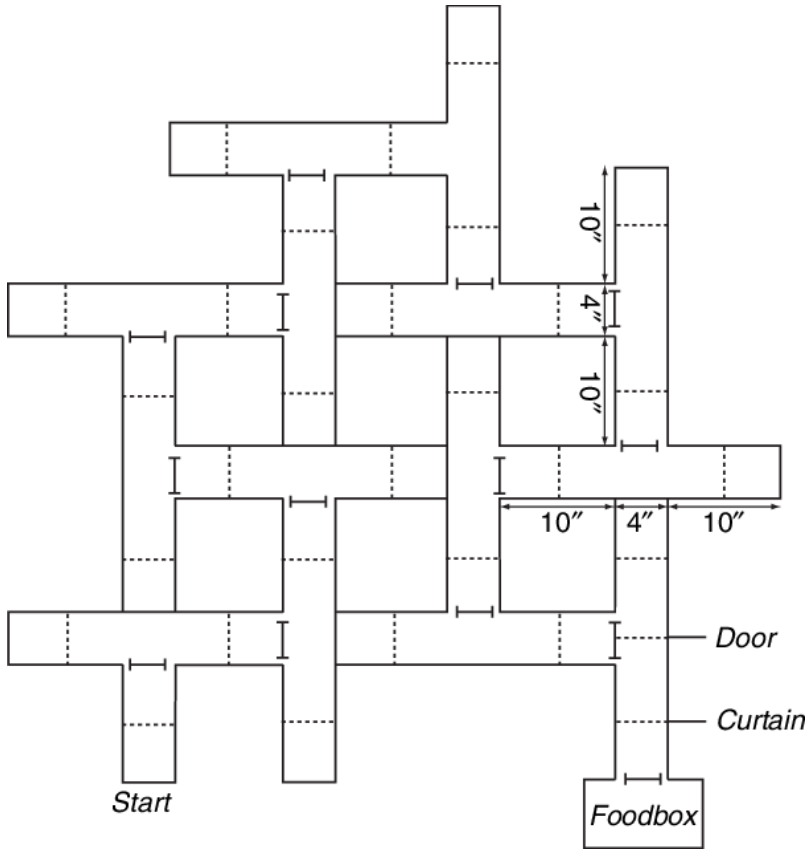
Latent learning & cognitive maps

Latent learning

- Learning that is not shown behaviorally until there is sufficient motivation
- It occurs without any obvious reinforcement of the behavior or associations that are learned

Cognitive map

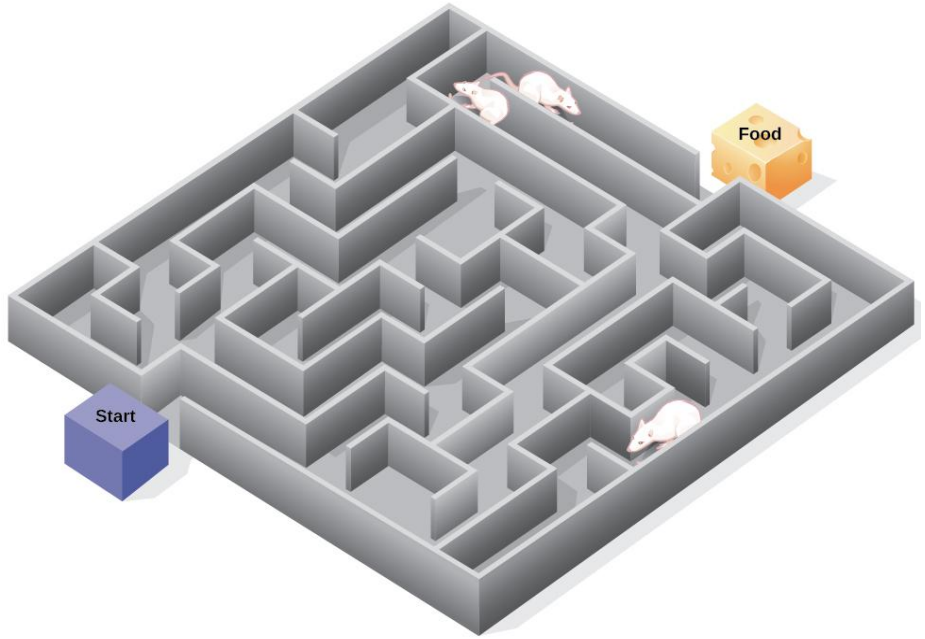
- Rats behaved as if they were responding to a mental representation of the overall layout of the maze rather than blindly exploring different parts of the maze through trial and error
- Mental representation of the space field that can guide what actions should be performed at any stage to achieve a particular goal



Maze used by Tolman and Honzik (1930) to study latent learning in rats. From Tolman EC (1948) Cognitive maps in rats and men. Psychol. Rev. 55: 189-208

Generation 0: Implications for the field

- Challenged the constraints of behaviorism, which stated that processes must be directly observable and that learning was the direct consequence of conditioning to stimuli
- Challenged the prevailing stimulus-response (S–R) view of learning and behavior, which corresponds to the simplest model-free way of learning policies
- Conditioning in volves more than the simple formation of associations between sets of stimuli or between responses and reinforcers. It includes learning and representing other facets of the total behavioral context



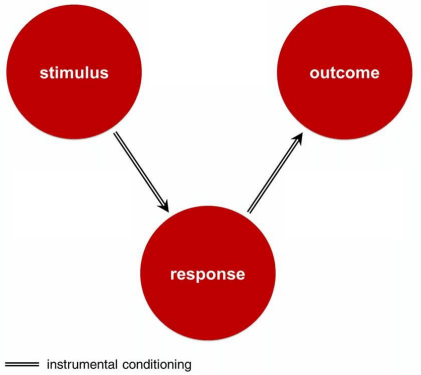
Generation 0 studies established a dichotomy between decision behavior controlled by a cognitive map and by S-R associations

Generation 1

goal-directed vs habitual actions

[experimental psychology]

Generation 1: Goal-Directed vs Habitual actions

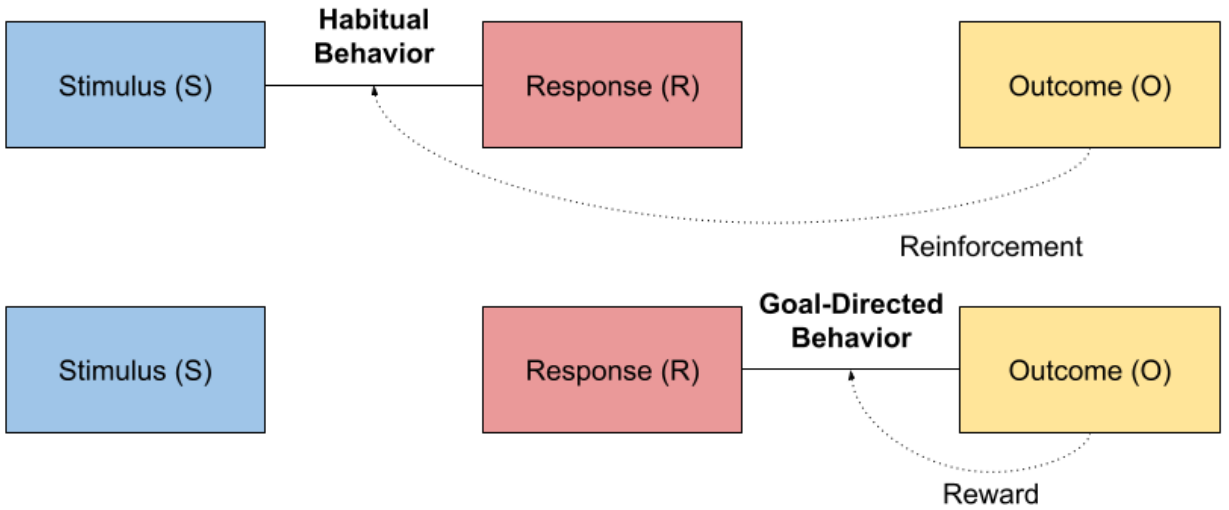


- Operationalized the use of cognitive maps for learning to choose appropriate actions (I.e. that maximize rewards/minimize punishments) in nonspatial domains

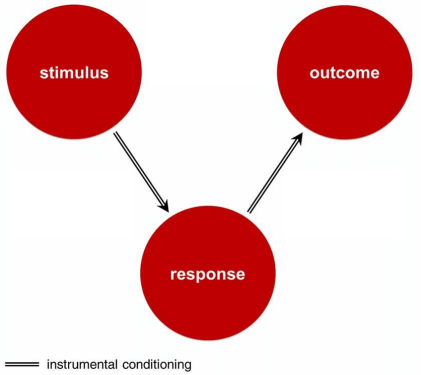
- Termed this as **goal-directed behavior/actions**

- Contrasted it with **habitual behavior/actions**

- Focused on animal studies to identify the neural bases of the two types of behaviors



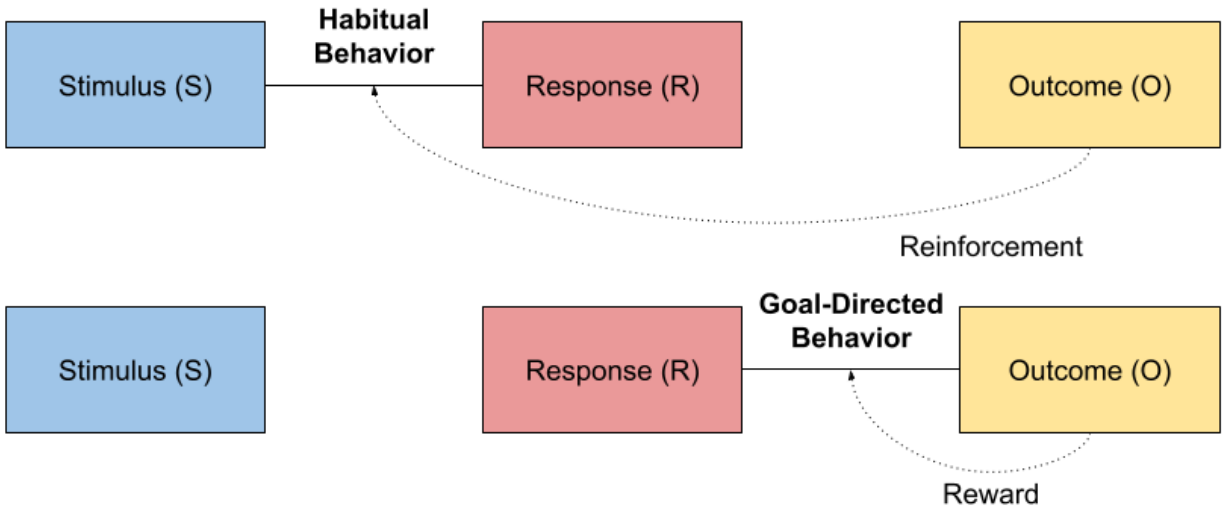
Goal-directed behavior/actions



-The action is made because we think that they will lead to outcomes that we desire

-Two criteria make an action goal-directed

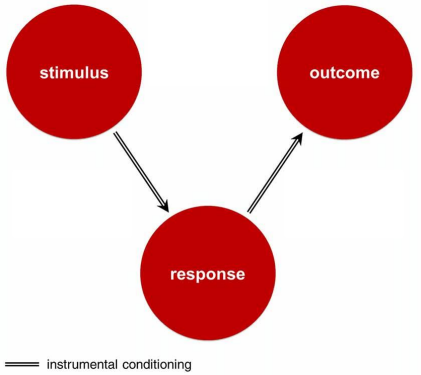
1. There must be **knowledge of the relationship between an action** (or sequence of actions) and its **consequences** --> response-outcome or R-O control
2. The **outcome should be motivationally relevant** or desirable at the moment of choice/action



Goal-directed behavior:

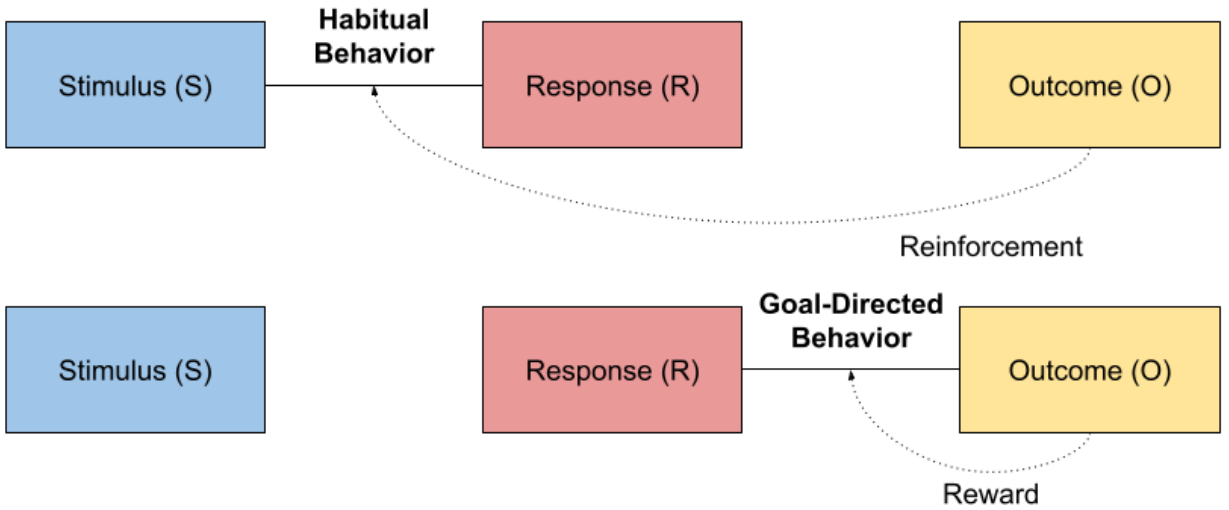
- Involves active deliberation
- Has high computational cost
- Shows adaptive flexibility to changing of environmental contingencies (e.g. the behavior stops if no reward follows the action)

Habitual behavior/actions



The action is

- made automatically, just because it has been rewarded in the past
- not influenced by the current value of the outcome it leads to
- continues to be enacted even when the outcome is undesired



Habitual behavior:

- Automatic (no active deliberation)
- Has low computational cost
- Is inflexible to changing of environmental contingencies (e.g. the behavior does not stop even if no reward follows the action)

Habitual behavior/actions

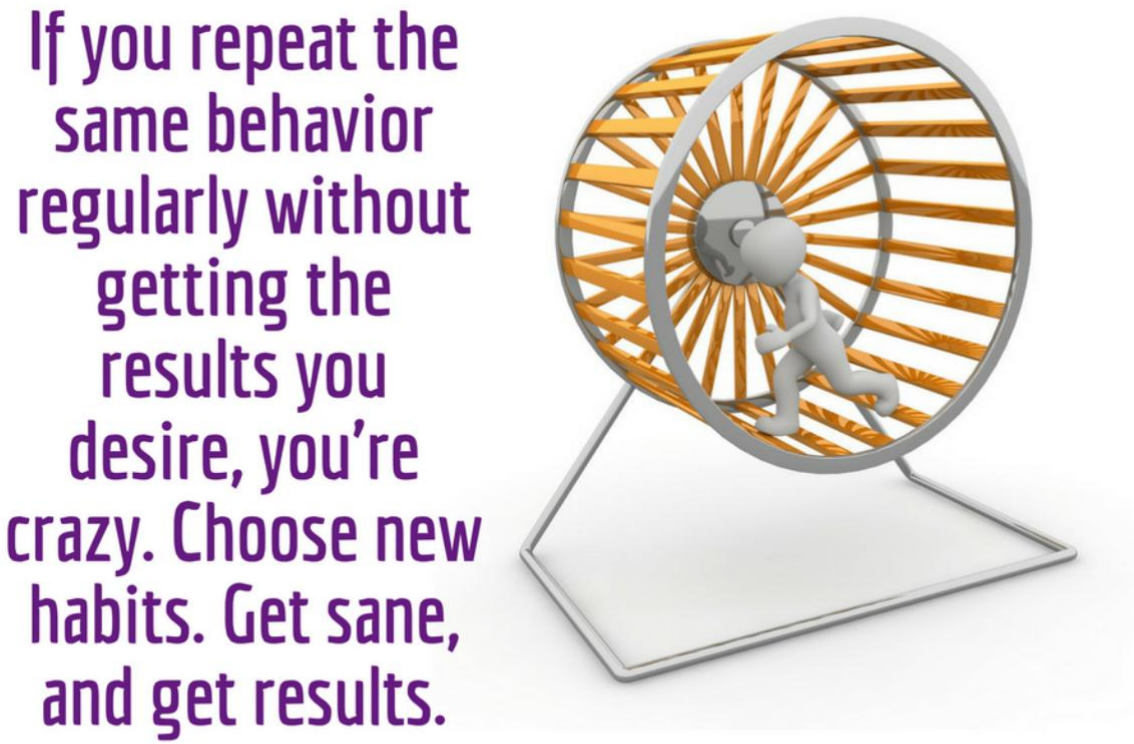
The action is

- made automatically, just because it has been rewarded in the past
- not influenced by the current value of the outcome it leads to
- continues to be enacted even when the outcome is undesired

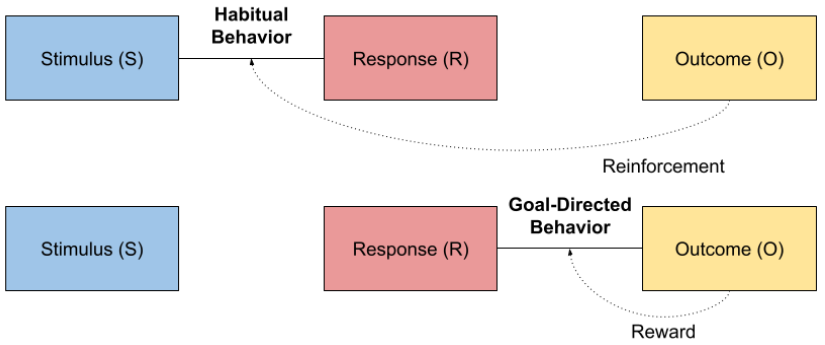
Habitual behavior:

- Automatic (no active deliberation)
- Has low computational cost
- Is inflexible to changing of environmental contingencies (e.g. the behavior does not stop even if no reward follows the action)

IT'S NOT THAT SIMPLE



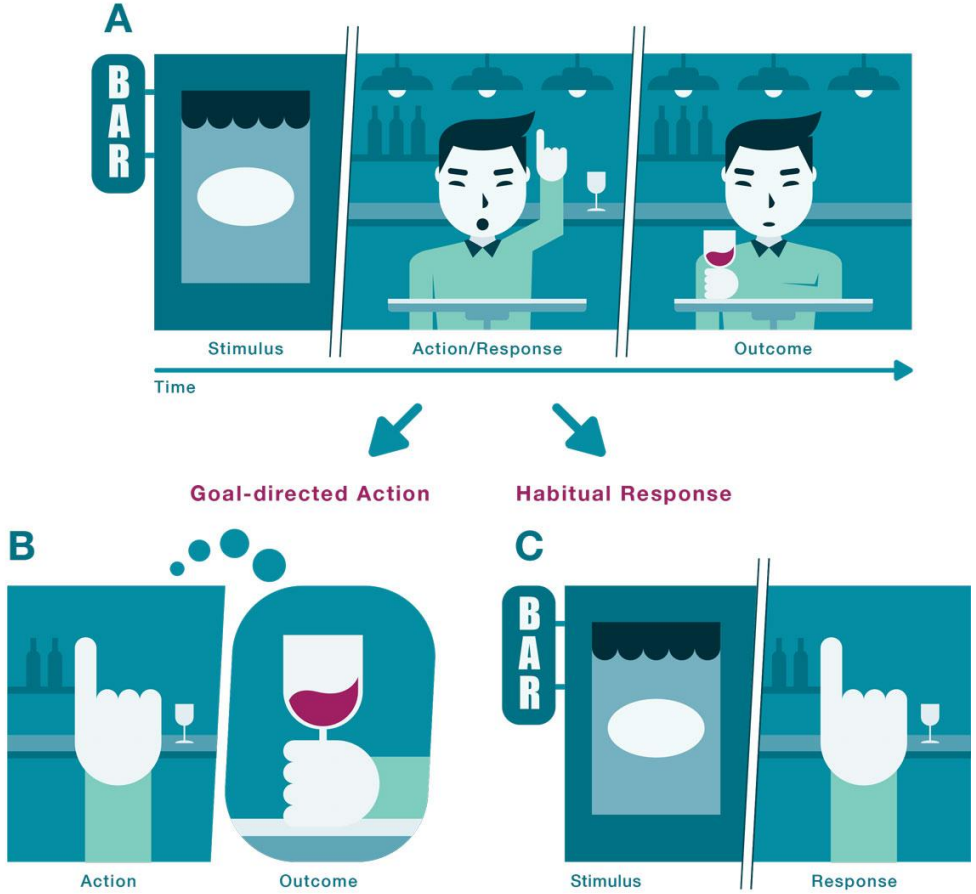
Generation 1: Goal-Directed vs Habitual actions



(A) Sequence of behavior: A person sees a bar (stimulus), enters and orders wine (action/response), and drinks the wine (outcome/reward).

(B) Goal-directed actions: Driven by the expected value of the outcome (e.g., ordering wine to drink). If the outcome loses value (e.g., after intoxication), the action is less likely to occur.

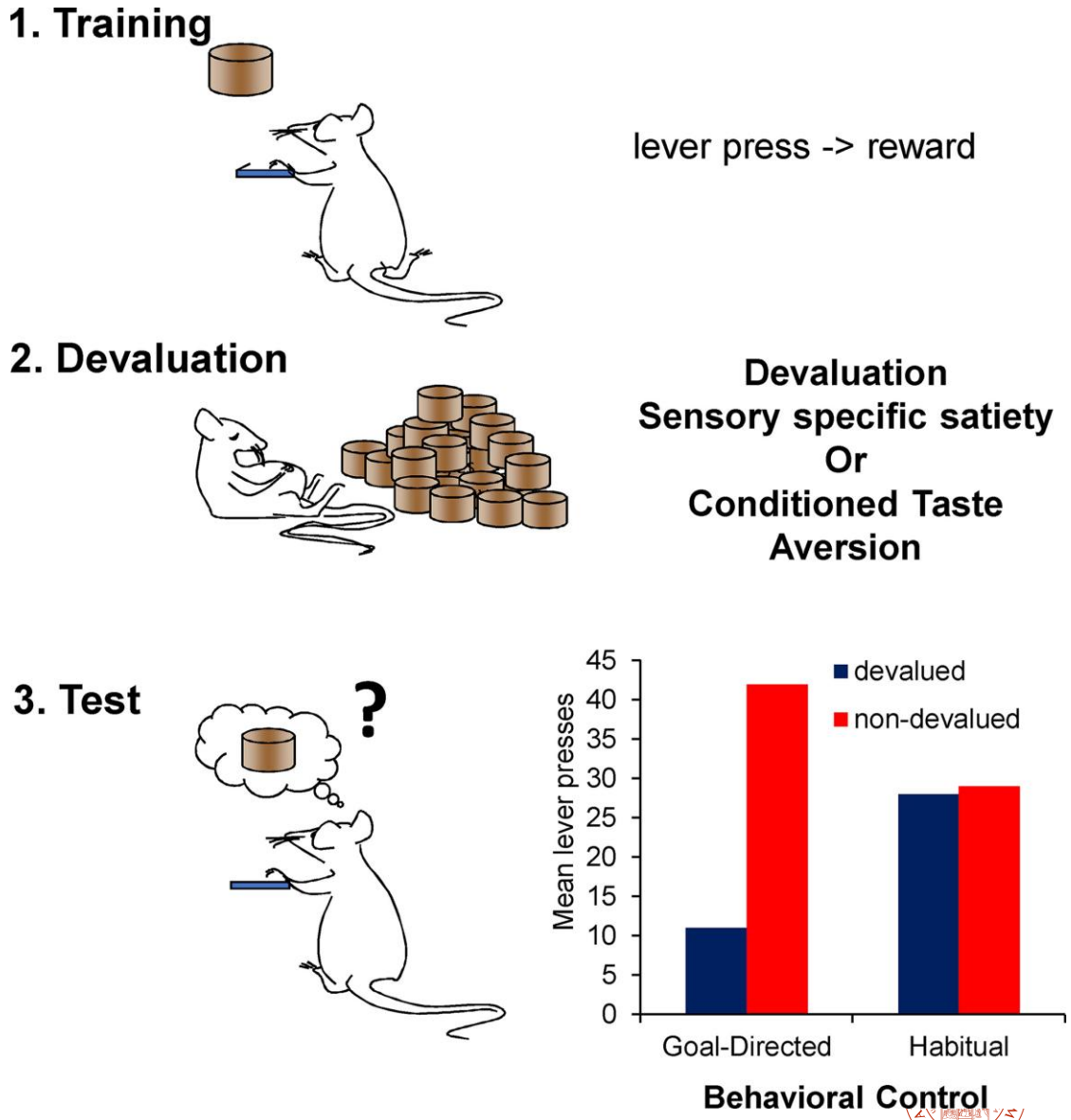
(C) Habitual actions: Triggered by environmental cues (e.g., sight of the bar) without considering the outcome. Behaviors persist even if the reward is devalued (e.g., after intoxication).



Pool, E. R., & Sander, D. (2019). Vulnerability to relapse under stress: Insights from affective neuroscience. *Swiss Medical Weekly, 149*(4748). https://doi.org/10.4414/smw.2019.20151

Testing if a behavior is goal-directed (vs habitual)

1. training session: the animal undergoes instrumental learning (learns that some actions will lead to rewards)
2. Post-training manipulation
    1. reinforcer **devaluation**
    2. contingency degradation
3. Testing session: the animal repeats the actions learned during instrumental training under extinction
    - If the action associated to the devalued reinforcer is performed less, then the behavior is goal-directed
    - if not, it's habitual

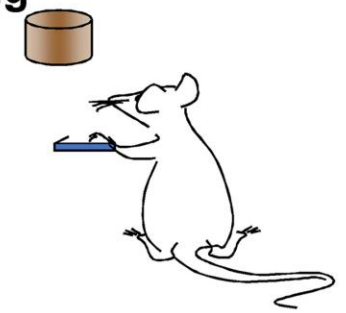


Corbit, Laura H. "Understanding the balance between goal-directed and habitual behavioral control." *Current opinion in behavioral sciences* 20 (2018): 161-168.

Testing if a behavior is habitual (vs goal-directed)

1. Extensive training session: **overtraining**
   - the animal undergoes instrumental learning (learns that some actions will lead to rewards)
   - This time the training is extensive
2. Post-training manipulation
   1. reinforcer **devaluation**
   2. contingency degradation
3. Testing session: the animal repeats the actions learned during instrumental training under extinction
   - If the action associated to the devalued reinforcer is performed less, then the behavior is goal-directed
   - if not, it's habitual

1. Training

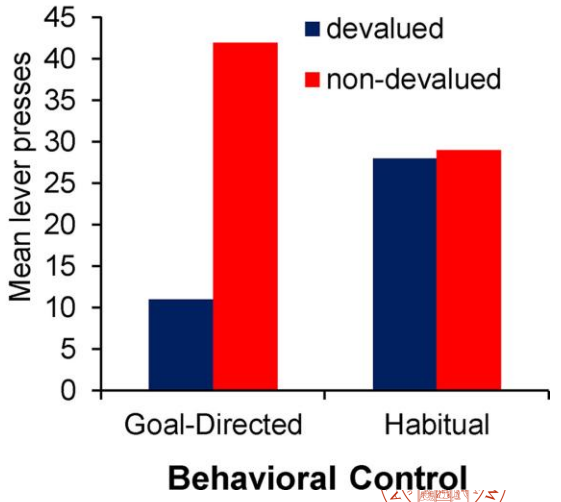


lever press -> reward

2. Devaluation

Devaluation
Sensory specific satiety
Or
Conditioned Taste Aversion

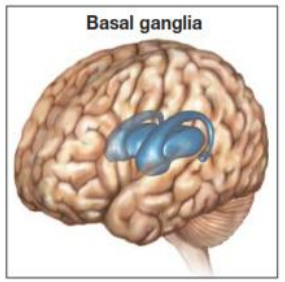
3. Test



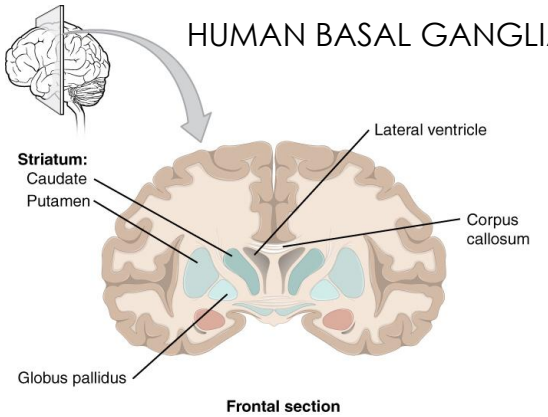
Dissociation of goal-directed vs habitual behavior in the striatum

Behavioral dissociation between goal-directed and habitual behavior corresponds to a neural dissociation:

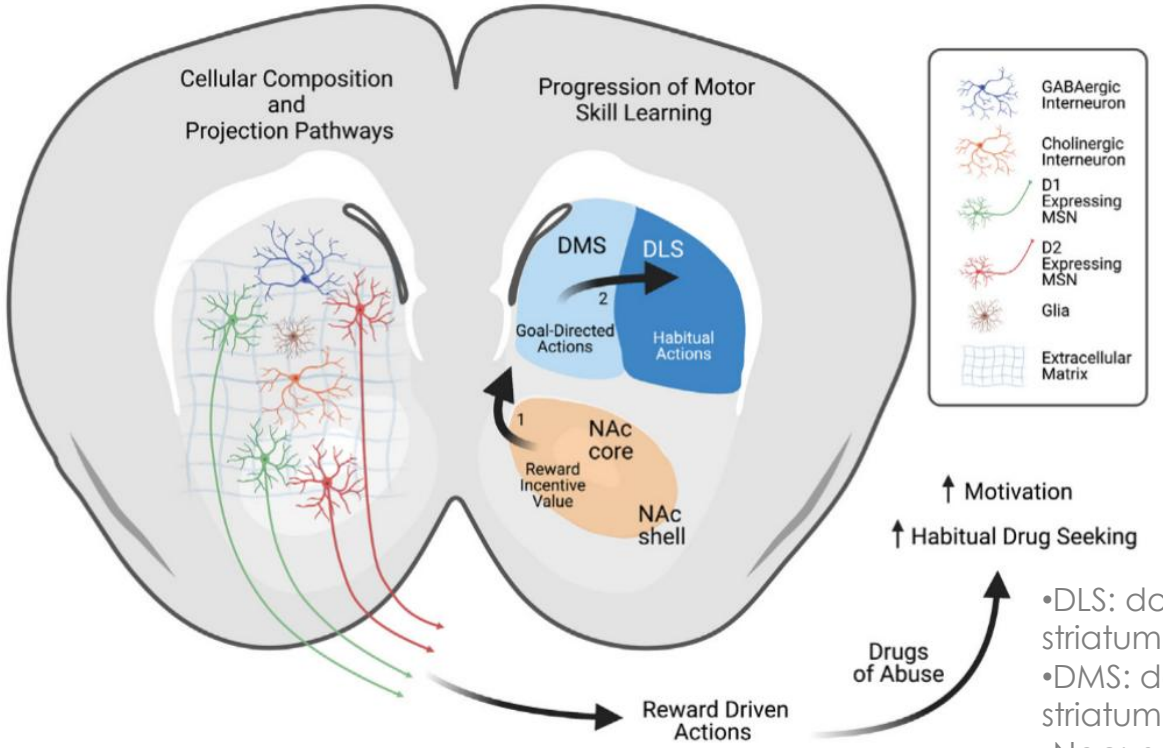
- **Dorsomedial striatum (DMS) →** supports goal-directed behavior
- **Dorsolateral striatum (DLS) →** supports habitual behavior



HUMAN BASAL GANGLIA



RODENT BASAL GANGLIA



- DLS: dorsolateral striatum
- DMS: dorsomedial striatum
- Nac: nucleus accumbens

Patton, M. S., & Mathur, B. N. (2023). Striatal circuits. In N. W. Gilpin (Ed.), *Neurocircuitry of addiction* (pp. 73–124). Academic Press. https://doi.org/10.1016/B978-0-12-823453-2.00010-2

The dopaminergic pathways

1. Nigrostriatal pathway

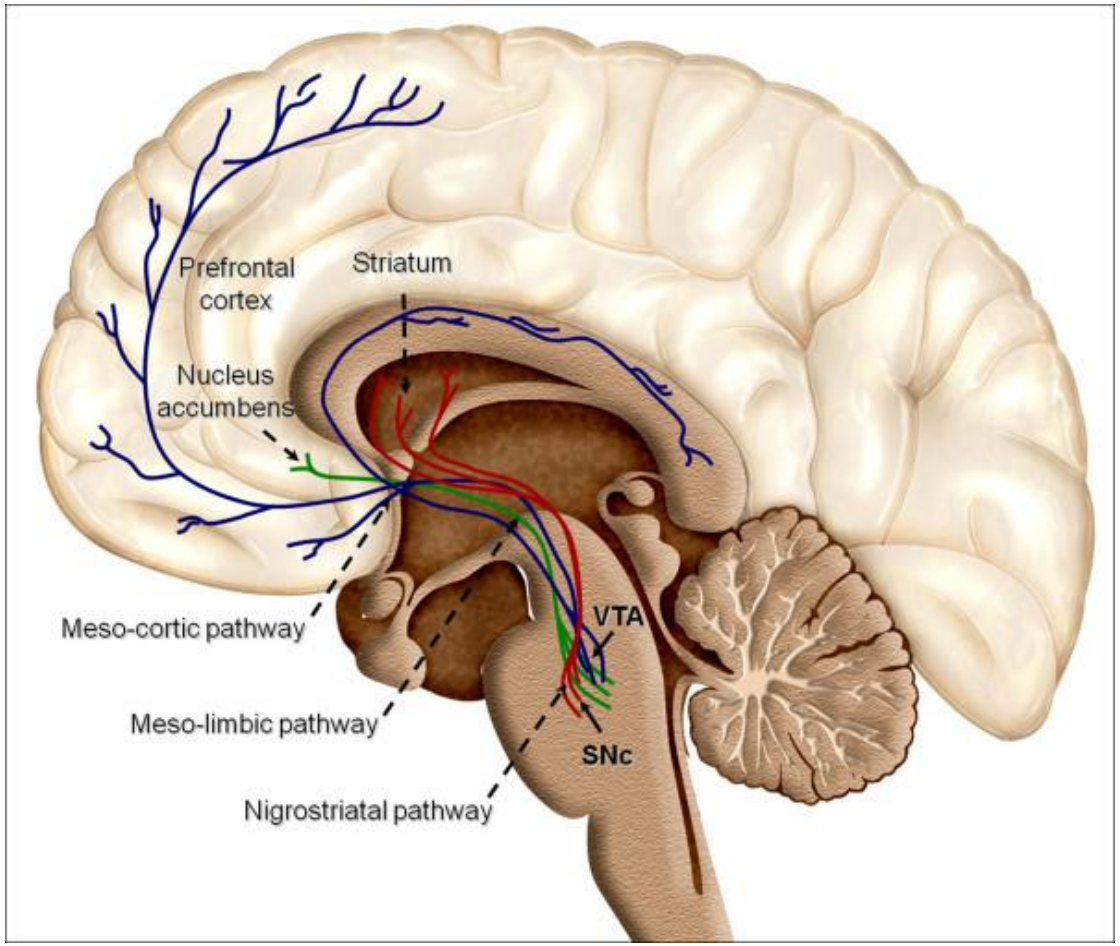
- originates in the substantia nigra pars compacta (SNc)
- projects primarily to the **caudate–putamen** (dorsal striatum in rodents)
- It is critical in the production of **movement** as part of the basal ganglia motor loop

2. Mesolimbic pathway

- originates in the VTA
- projects to the **nucleus accumbens**, septum, amygdala and hippocampus

3. Mesocortical pathway

- Originates in the VTA
- projects to the medial prefrontal, cingulate, **orbitofrontal** and perirhinal cortex

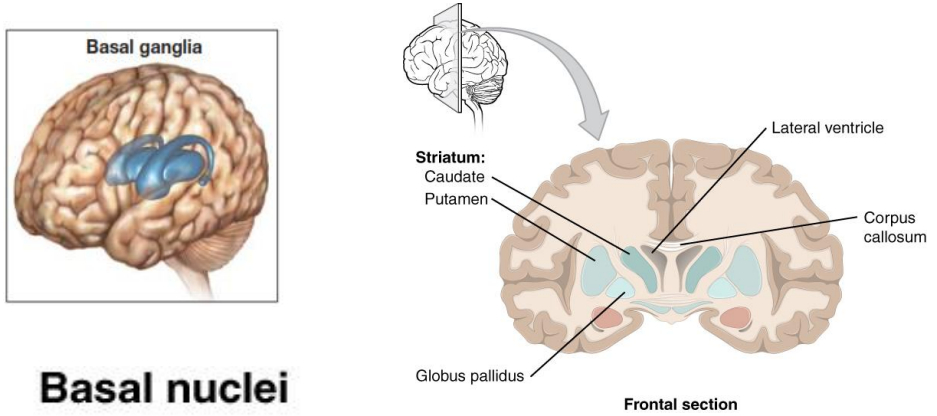


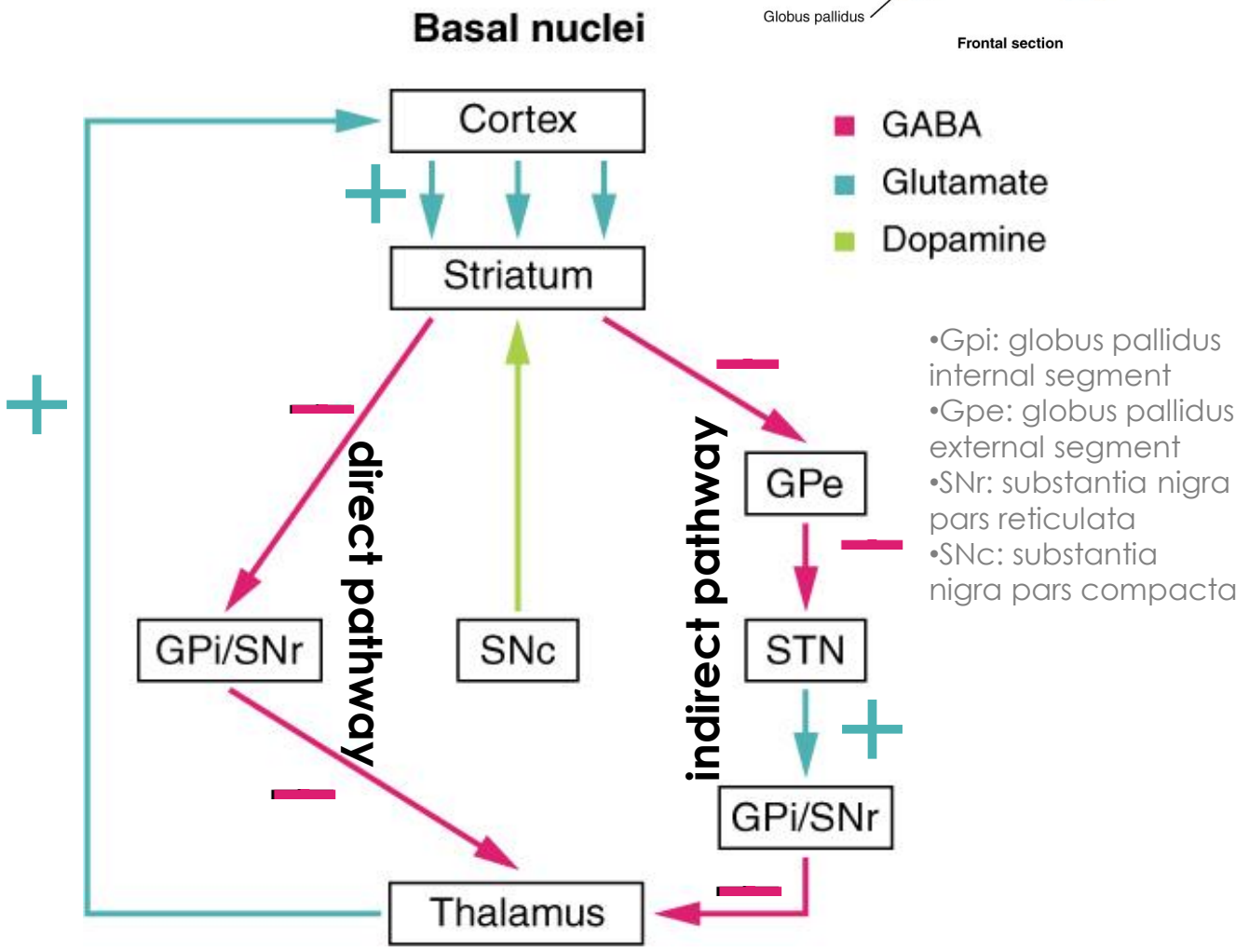
https://www.brainfacts.org/3d-brain#intro=false&focus=Brain

Striatum: linking motivation-action

The **striatum** may be the **interface where reward influences action**

- The basal ganglia are involved in the selection of actions
- Rewards may influence which actions are selected
    - by affecting plasticity in the striatum, so as to reinforce rewarded actions and make them more likely to recur

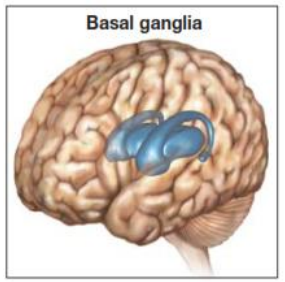




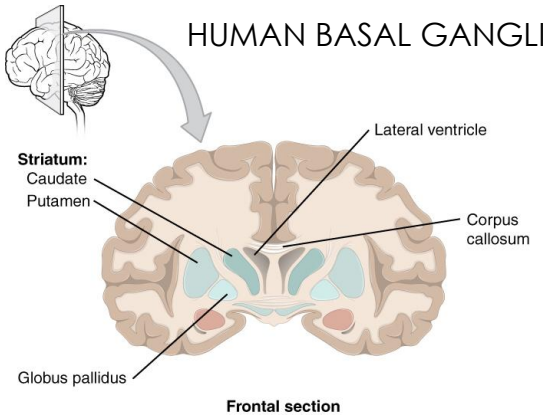
Dissociation of goal-directed vs habitual behavior in the striatum

Behavioral dissociation between goal-directed and habitual behavior corresponds to a neural dissociation:

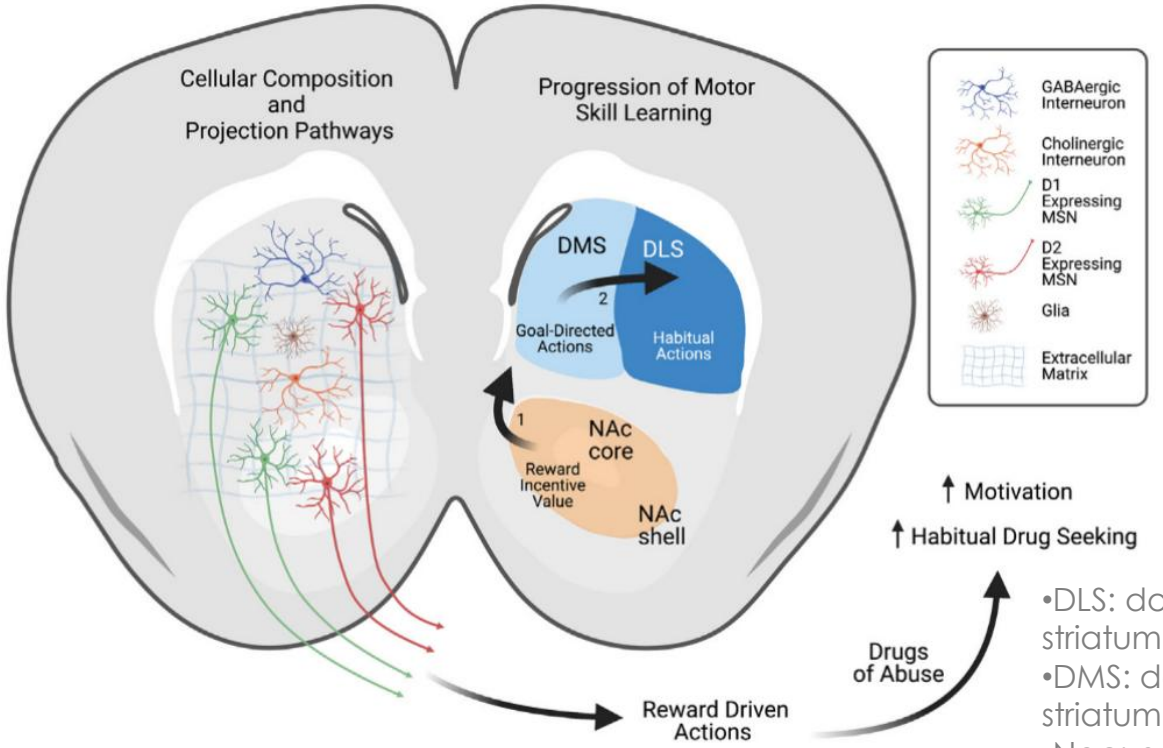
- **Dorsomedial striatum (DMS) →** supports goal-directed behavior
- **Dorsolateral striatum (DLS) →** supports habitual behavior



HUMAN BASAL GANGLIA



RODENT BASAL GANGLIA



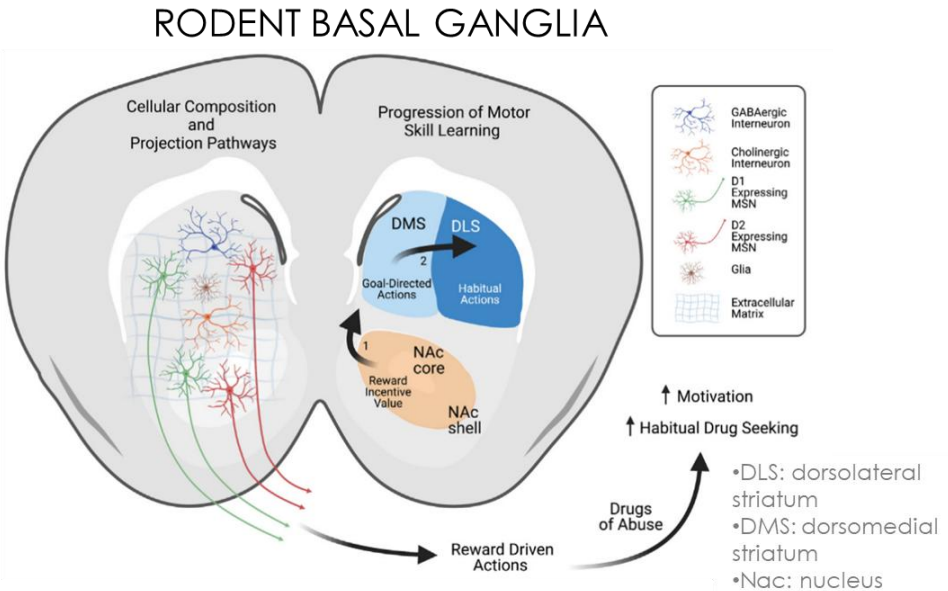
- DLS: dorsolateral striatum
- DMS: dorsomedial striatum
- Nac: nucleus accumbens

Patton, M. S., & Mathur, B. N. (2023). Striatal circuits. In N. W. Gilpin (Ed.), *Neurocircuitry of addiction* (pp. 73–124). Academic Press. https://doi.org/10.1016/B978-0-12-823453-2.00010-2

Generation 1: Implications for the field

- The need for overtraining to make a behavior habitual suggests that **behavior is initially goal-directed but then becomes habitual** over the course of experience
    - Adaptive in stable environment but can be maladaptive when environment changes

- Behavioral dissociation between goal-directed and habitual behavior corresponds to a neural dissociation:
    - **Dorsomedial striatum** → supports goal-directed behavior
    - **Dorsolateral striatum** → supports habitual behavior



Patton, M. S., & Mathur, B. N. (2023). Striatal circuits. In N. W. Gilpin (Ed.), *Neurocircuitry of addiction* (pp. 73–124). Academic Press. https://doi.org/10.1016/B978-0-12-823453-2.00010-2

Generation 2

Goal-directed vs habitual actions in the human brain
[cognitive neuroscience]

Generation 2: Actions and Habits in the Human Brain

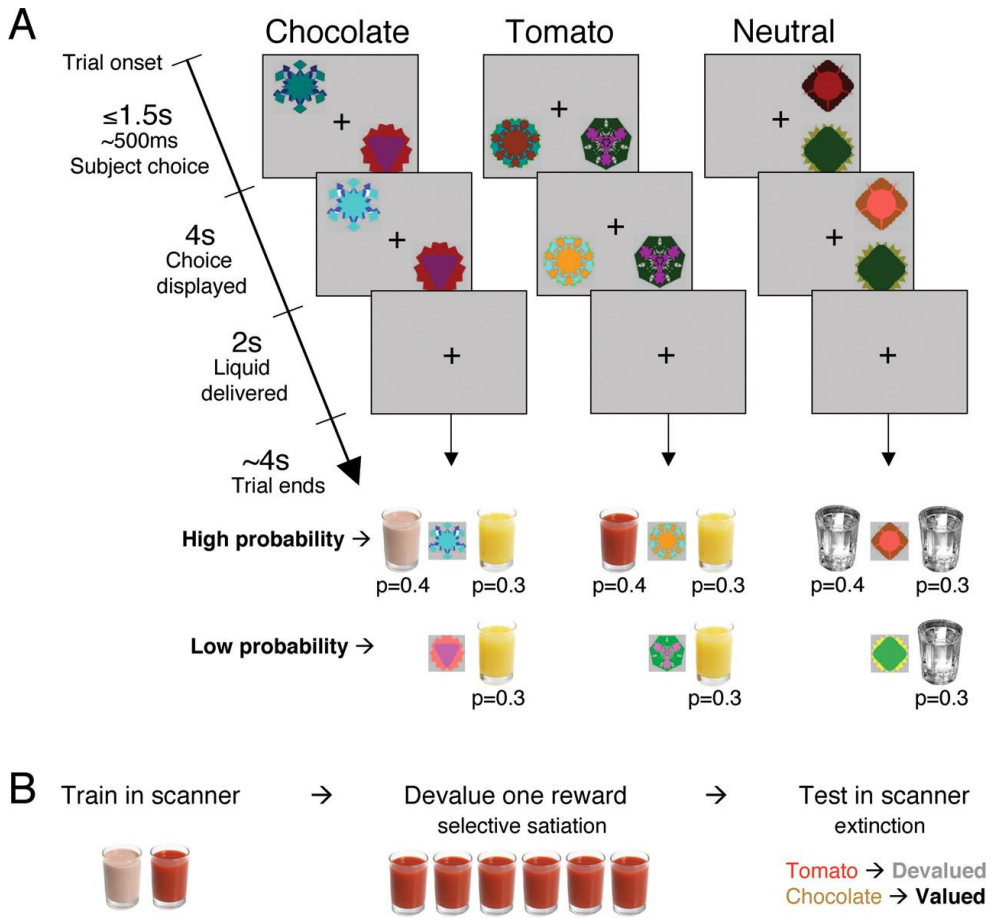
- Successful animal paradigms were adapted for human experiments

- Use of fMRI in order to investigate the neural bases of
    - Goal-directed actions
    - Habitual actions

Neural substrates of goal-directed behavior in humans

Method:

- human subjects were trained on a task in which two different actions resulted in two distinct food reward outcomes (I.e. instrumental conditioning)
- One of the outcomes was then devalued(by feeding subjects that food to satiety, i.e., until they would consume no more of it)
- the values of other foods not eaten remained high
- After devaluation participants performed the instrumental actions (choice of stimuli) under extinction

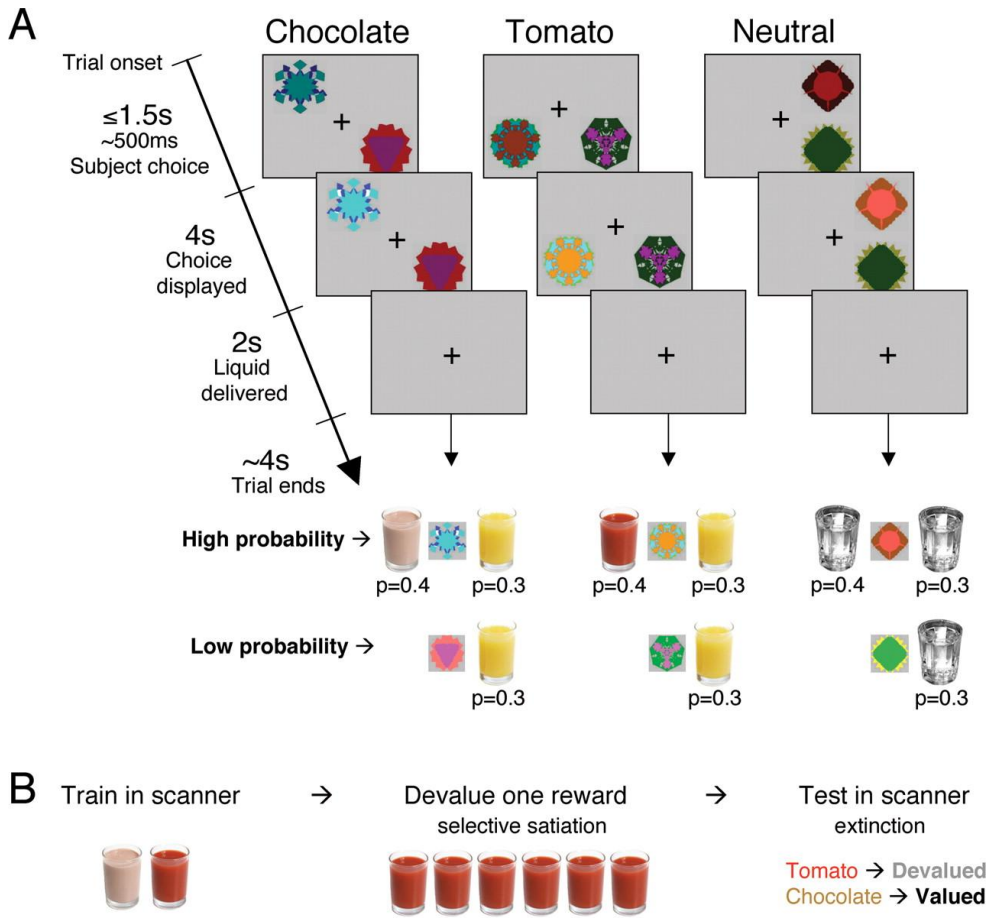


Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the neu-ral substrates of goal-directed learning in the human brain. J. Neurosci.27,4019–4026.

Neural substrates of goal-directed behavior in humans

Method:

- fMRI was recorded at train and test to examine brain areas responding during action selection
  - looking for areas that showed sensitivity to the change in value of the associated outcomes
  - such area(s) would be candidate regions for implementing goal-directed behavior in humans

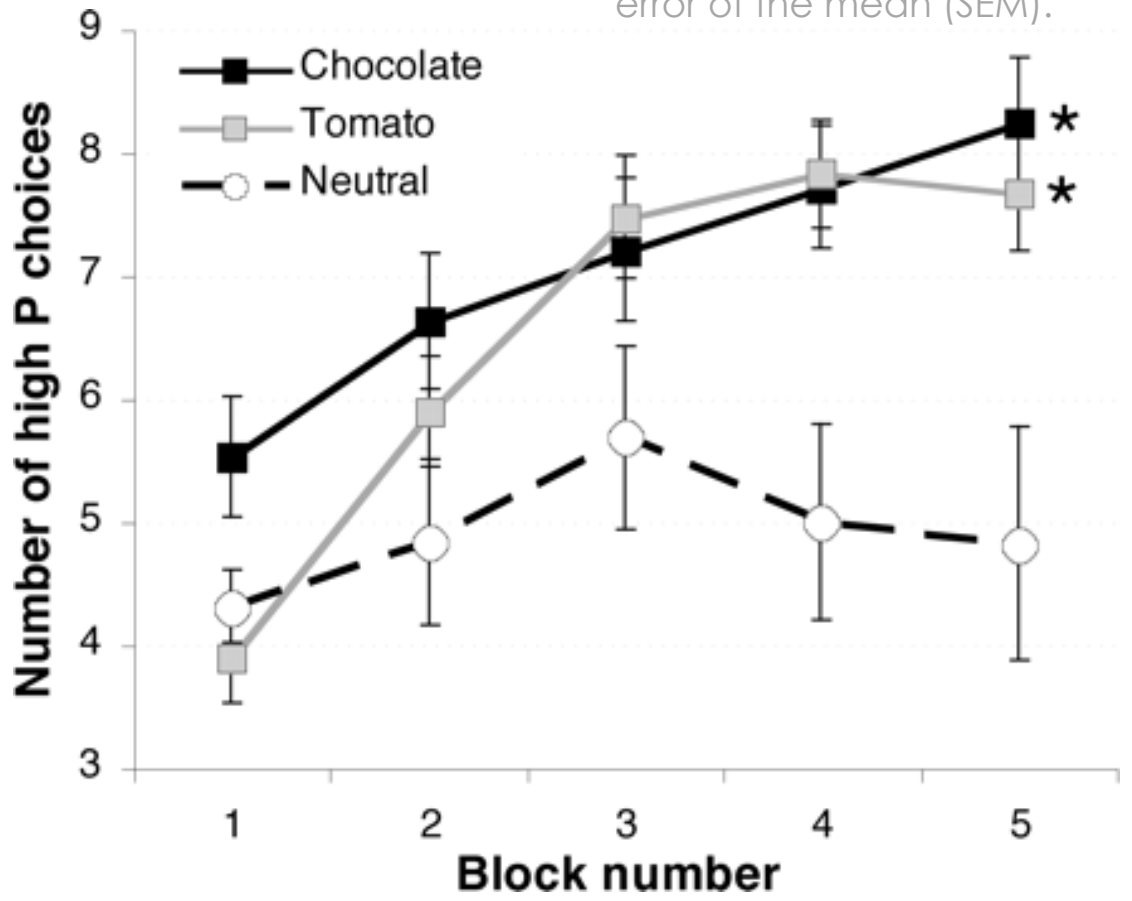


Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the neu-ral substrates of goal-directed learning in the human brain. J. Neurosci.27,4019–4026.

Neural substrates of goal-directed behavior in humans

Results: behavioral learning curves

- The total number of high-probability action choices across five 10-trial blocks was averaged across 19 subjects during training.
- **Over time, subjects increasingly favored the high-probability actions associated with tomato juice or chocolate milk over low-probability counterparts.**
- In the neutral condition, however, subjects were indifferent between high- and low-probability actions.

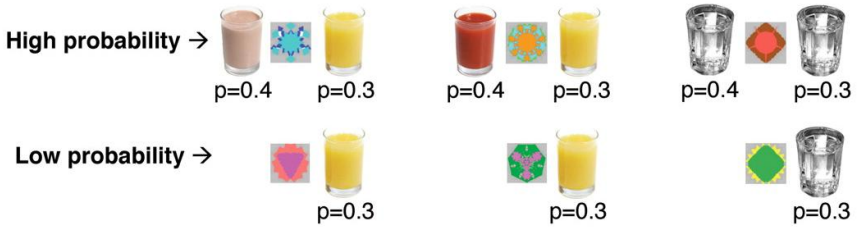


Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the neu-ral substrates of goal-directed learning in the human brain. J. Neurosci.27,4019–4026.

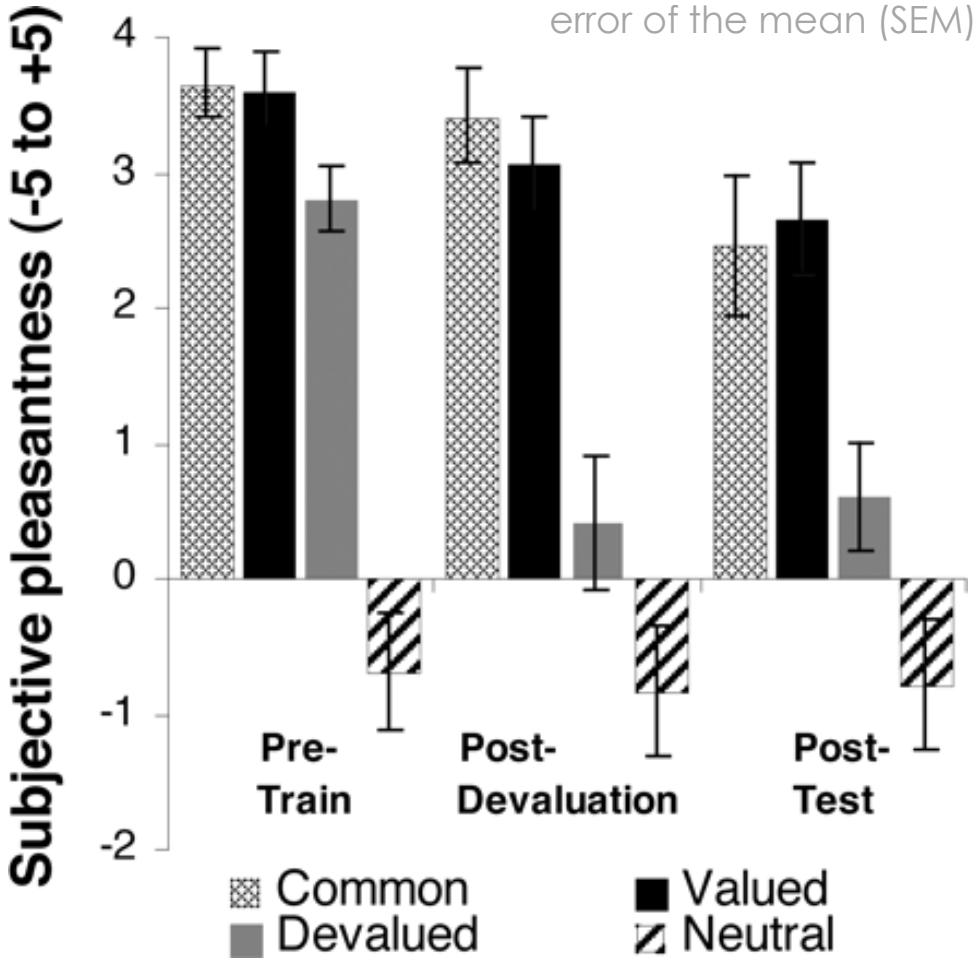
Neural substrates of goal-directed behavior in humans

Results: behavioral

- Ratings were made on a scale of −5 (very unpleasant) to +5 (very pleasant) before training, after devaluation, and after the test.
- **Post-devaluation, the rating for the food that was eaten (devalued) significantly decreased compared to the food not eaten (valued).**



Error bars represent standard error of the mean (SEM).

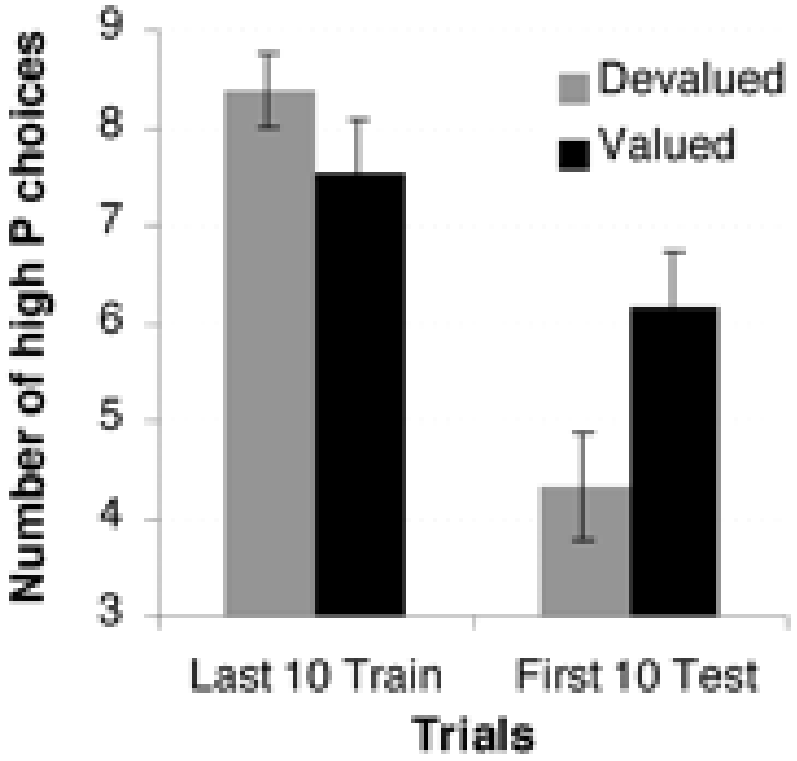


Neural substrates of goal-directed behavior in humans

Error bars represent standard error of the mean (SEM).

Results: behavioral

At testing **post-devaluation, subjects reduced their choices of the high-probability action associated with the devalued food** significantly more than that of the valued food.



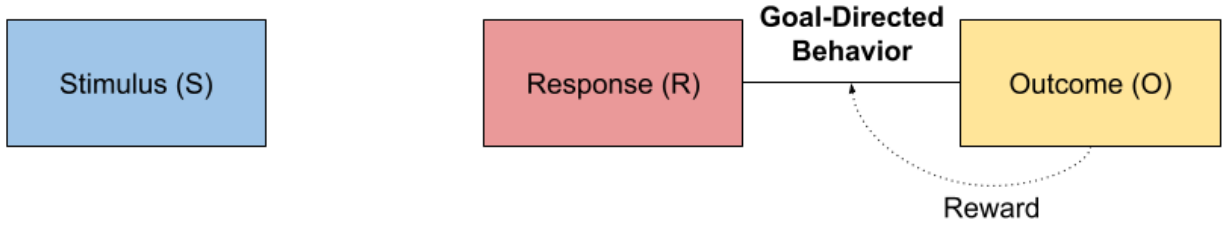
Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the neu-ral substrates of goal-directed learning in the human brain. J. Neurosci.27,4019–4026.

Goal-directed behavior/actions

The action is made because we think that they will lead to outcomes that we desire

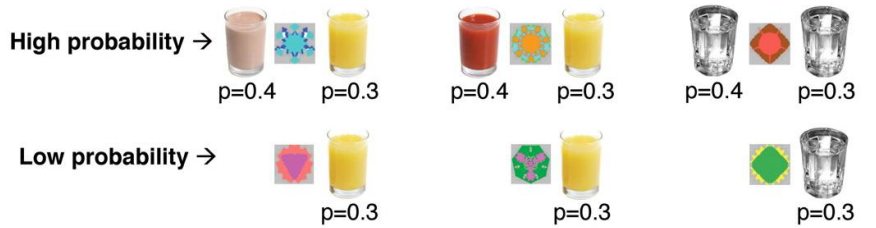
Two criteria make an action goal-directed

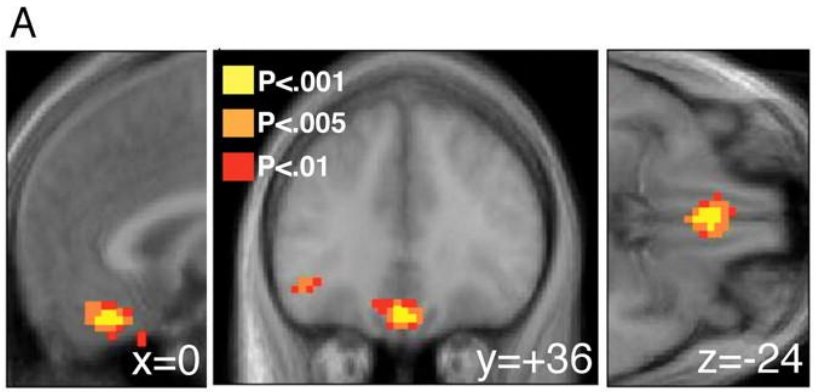
1. There must be **knowledge of the relationship between an action** (or sequence of actions) and its **consequences** --> response-outcome or R-O control
2. The **outcome should be motivationally relevant** or desirable at the moment of choice/action

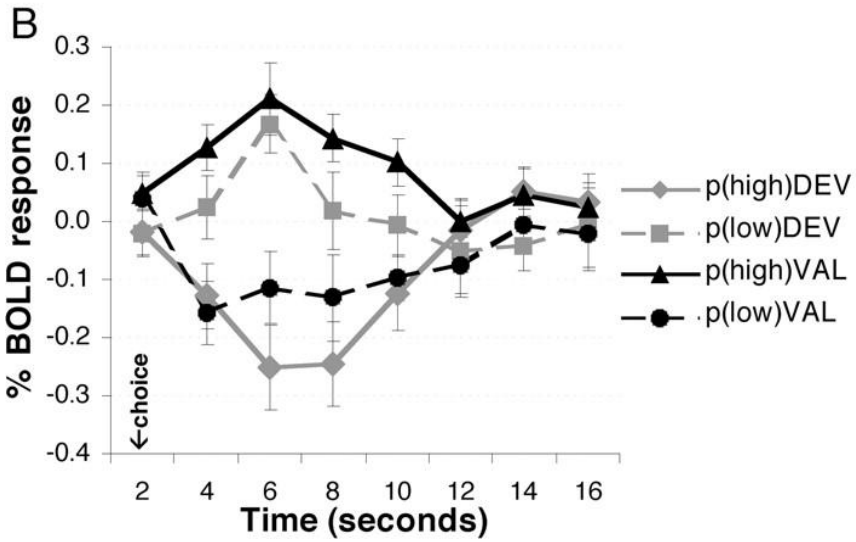


Neural substrates of goal-directed behavior in humans

Are there brain areas that respond differently between the still motivationally relevant outcome (I.e. valued) and the devalued one?

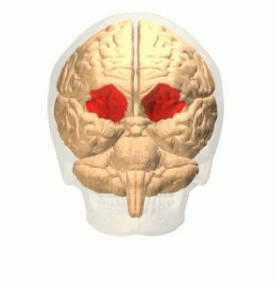






Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the neu-ral substrates of goal-directed learning in the human brain. J. Neurosci.27,4019–4026.

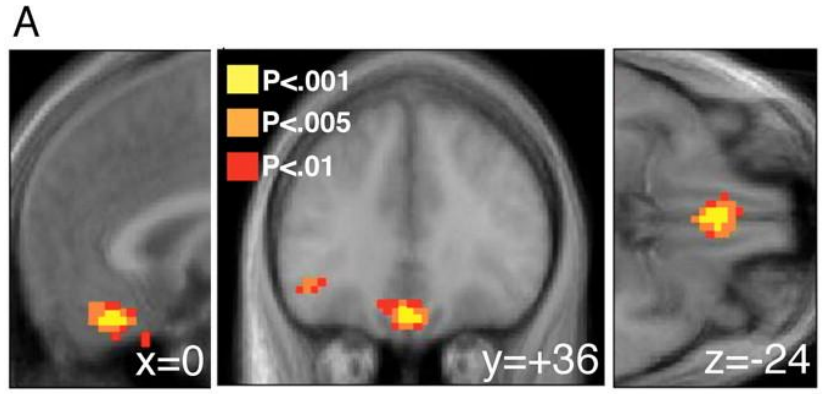
Neural substrates of goal-directed behavior in humans

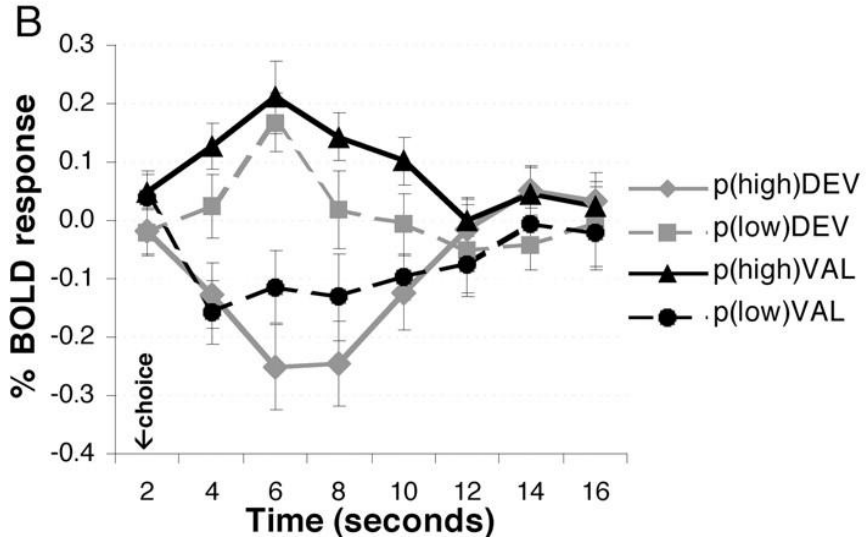


Results: neural

(A) A region of the **medial orbitofrontal cortex (OFC)** showed significant modulation in its activity during instrumental action selection.

(B) The modulation was dependent on the value of the associated outcome.





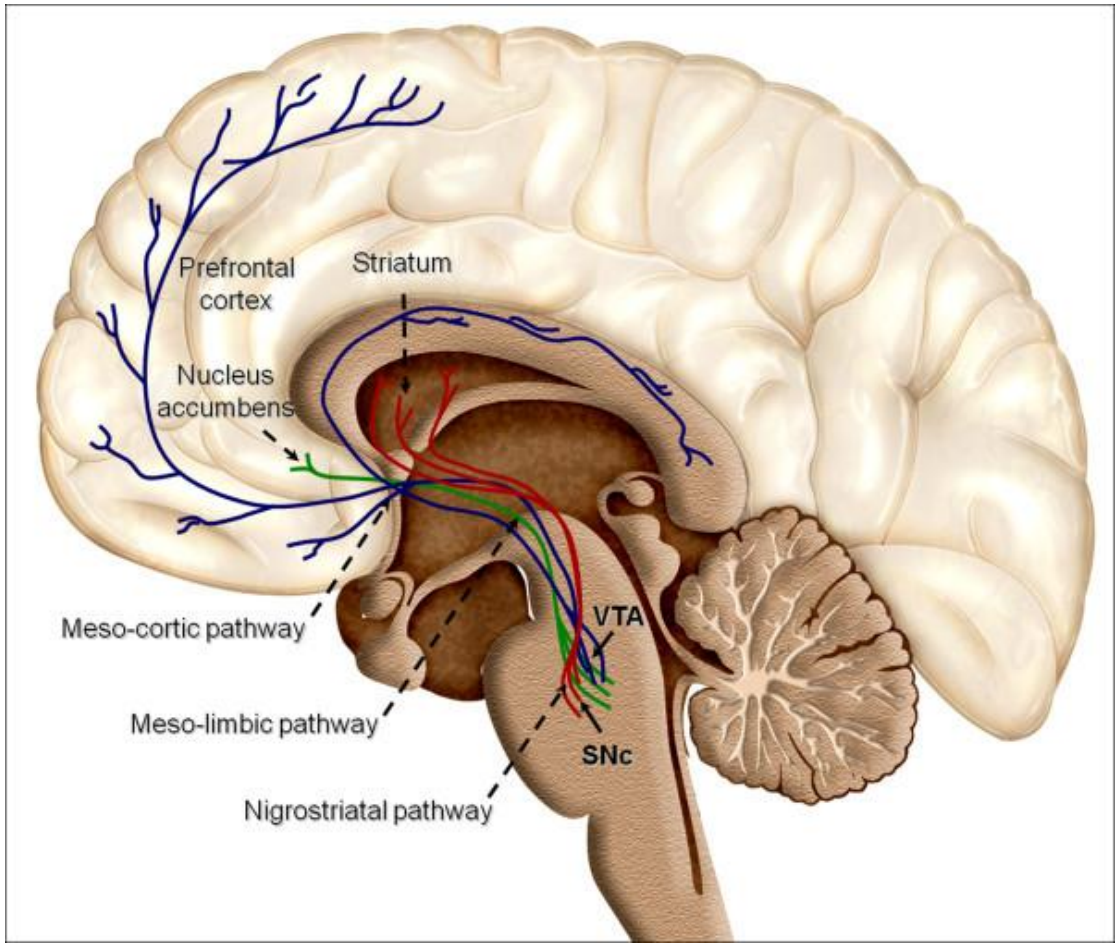
Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the neu-ral substrates of goal-directed learning in the human brain. J. Neurosci.27,4019–4026.

The dopaminergic pathways

1. **Nigrostriatal pathway**
- originates in the substantia nigra pars compacta (SNc)
- projects primarily to the **caudate–putamen** (dorsal striatum in rodents)
- It is critical in the production of **movement** as part of the basal ganglia motor loop

2. **Mesolimbic pathway**
- originates in the VTA
- projects to the **nucleus accumbens**, septum, amygdala and hippocampus

3. **Mesocortical pathway**
- Originates in the VTA
- projects to the medial prefrontal, cingulate, **orbitofrontal** and perirhinal cortex



https://www.brainfacts.org/3d-brain#intro=false&focus=Brain

Neural substrates of habitual behavior in humans

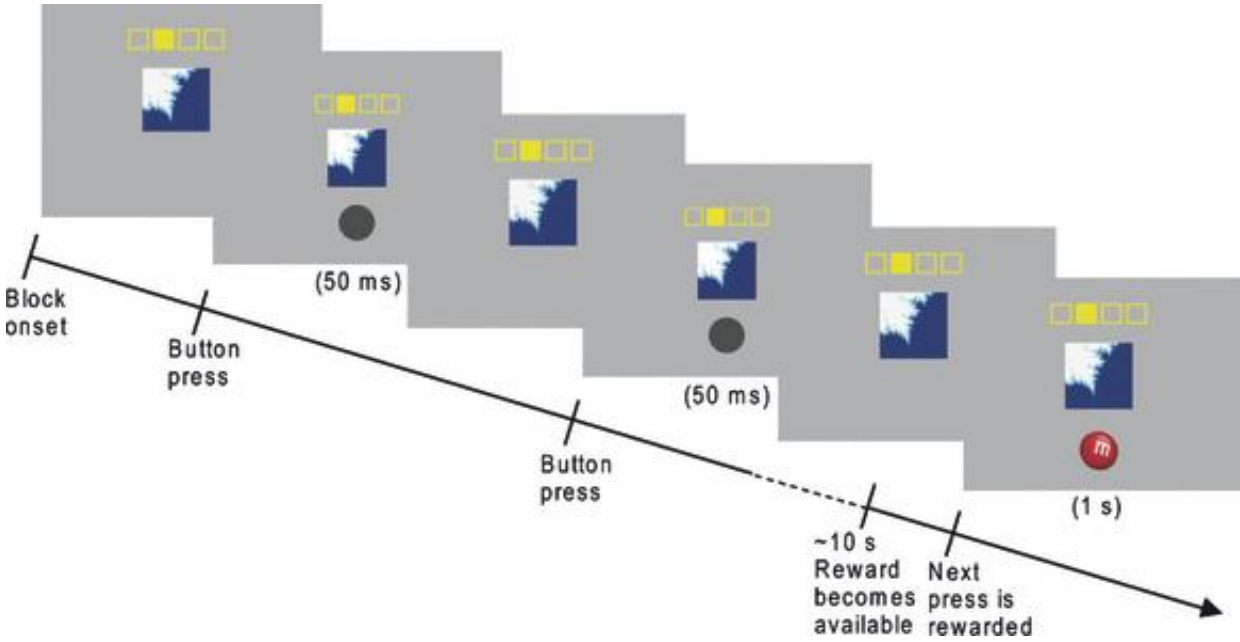
- Group 1: extensive training (6 times > group2)
- Group 2: little training

A fractal image was displayed on the screen with a schematic indicating which button to press.

- Participants could press the indicated button freely.

After each press:

- A gray circle (50ms) appeared → No reward.
- A picture of an M&M or Frito (1000ms) appeared → Food reward.
- Only presses of the correct button triggered the display of the outcome (reward or no reward)
- Rewards followed a variable interval 10-second schedule.



Tricomi, E.M., Balleine, B.W., and O'Doherty, J.P. (2009). A specific role forposterior dorsolateral striatum in human habit learning. Eur. J. Neurosci.29,2225–2232

Neural substrates of habitual behavior in humans

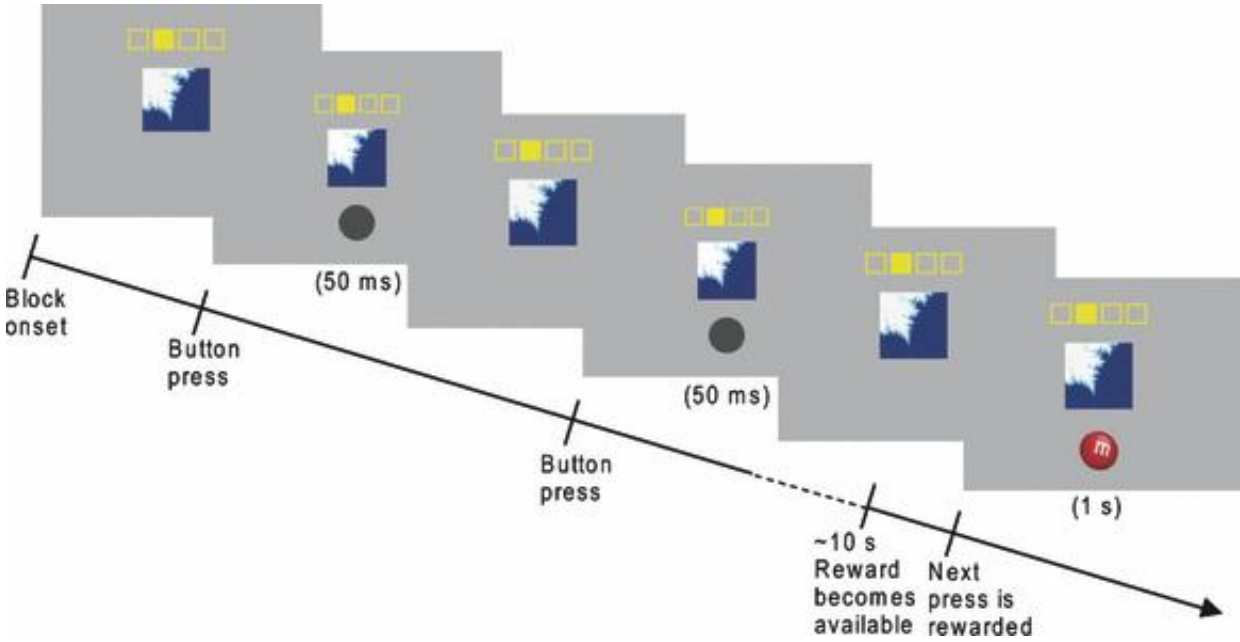
- Group 1: extensive training (6 times > group2)
- Group 2: little training

After the final session of training, **one of the two food outcomes was devalued through selective satiation:**

- participants were asked to eat that food until it was no longer pleasant to them.

To test the **effects of the devaluation** procedure on behavior, participants were placed back into the scanner for an **extinction test:**

- the fractal cues were presented again and participants' responses were collected
- No reward was provided



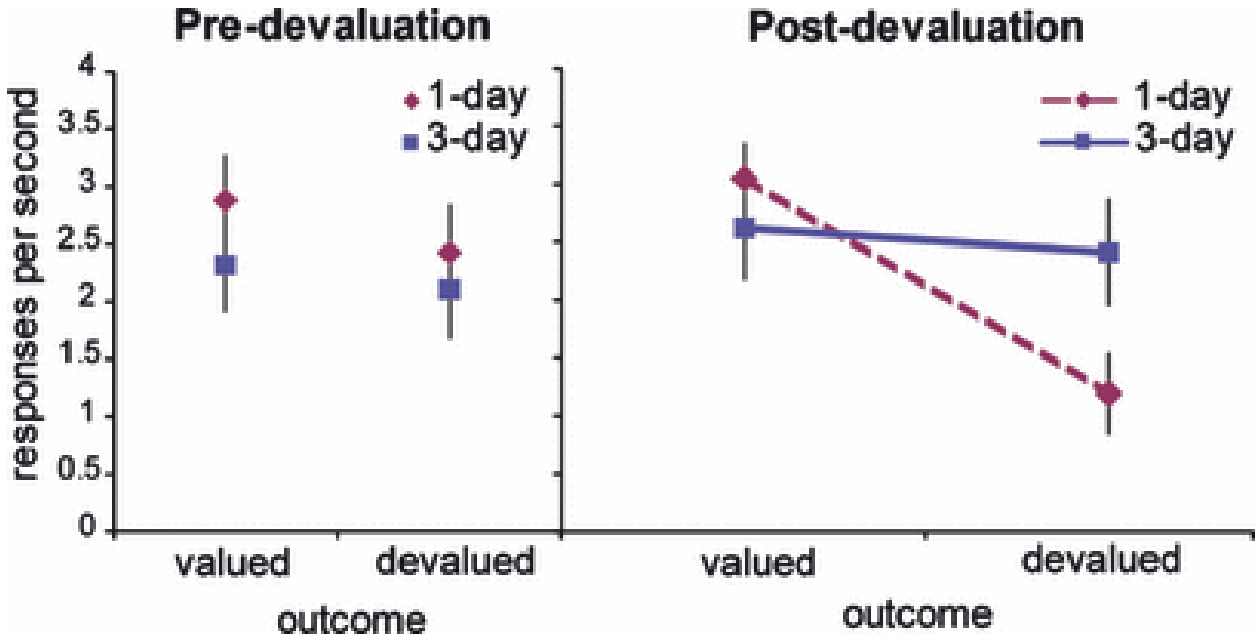
Tricomi, E.M., Balleine, B.W., and O'Doherty, J.P. (2009). A specific role forposterior dorsolateral striatum in human habit learning. Eur. J. Neurosci.29,2225–2232

Neural substrates of habitual behavior in humans

Results: behavioral

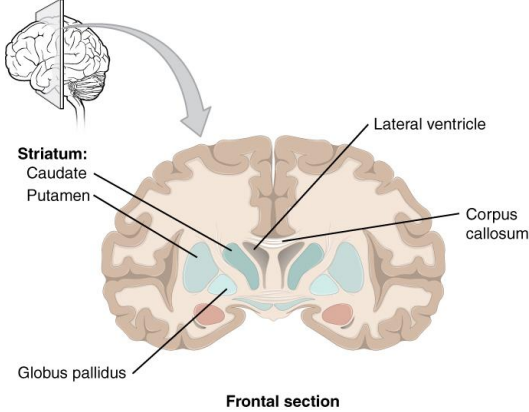
Response Rates Before and After Devaluation

- Last Training Session (Before Devaluation):
  - No significant differences in response rates between groups.
  - No differences when responding for the two food rewards (one later devalued, one not).
- **Test Session** (After Devaluation):
  - Responses for the still-valued outcome remained high, regardless of group.
  - **3-day group: Continued responding for the devalued outcome. → habitual behavior**
  - **1-day group: Reduced response rates for the devalued outcome.**



Tricomi, E.M., Balleine, B.W., and O'Doherty, J.P. (2009). A specific role forposterior dorsolateral striatum in human habit learning. Eur. J. Neurosci.29,2225–2232

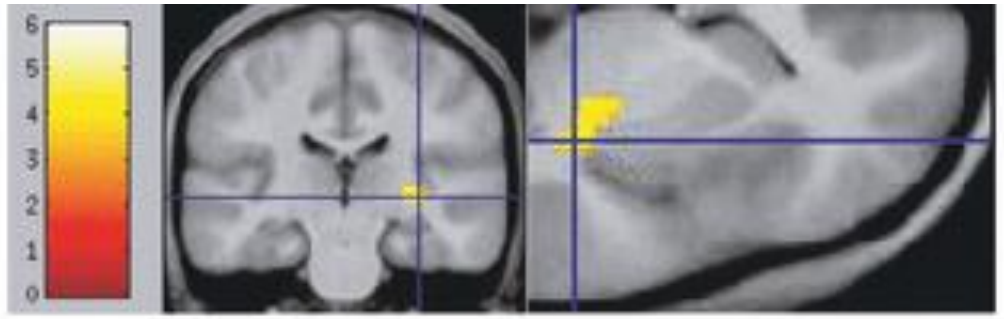
Neural substrates of habitual behavior in humans

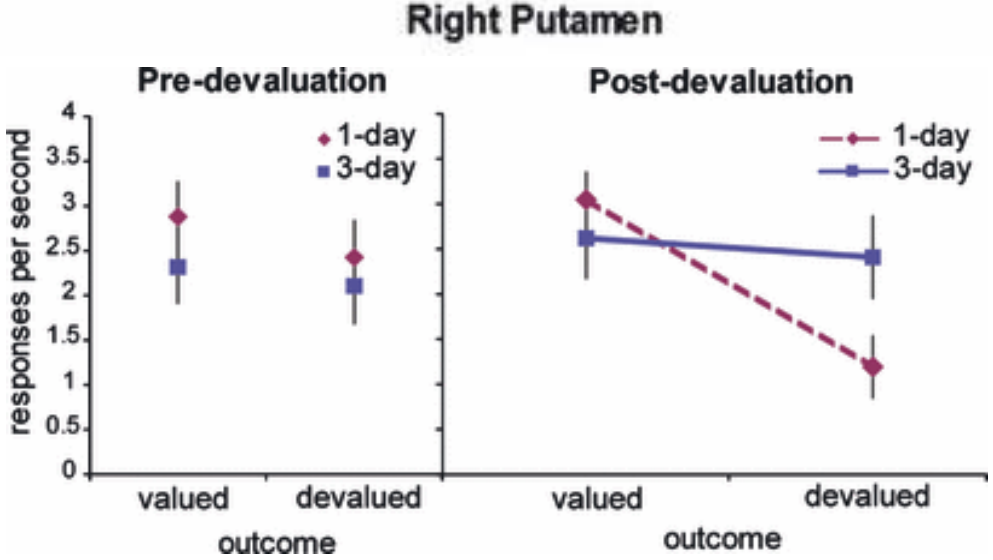


Results: neural

Within-subjects analysis in the 3-day group:

A comparison of the last two sessions of training versus the first two sessions revealed a significant cluster of activation in a region within the dorsolateral striatum (DLS), in the right posterior putamen/globus pallidus





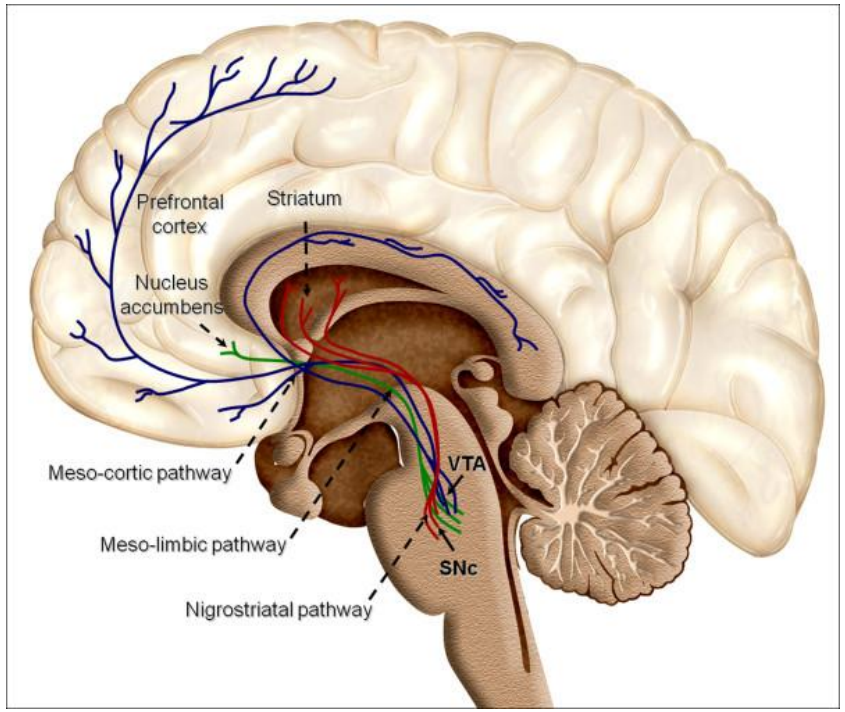
Tricomi, E.M., Balleine, B.W., and O'Doherty, J.P. (2009). A specific role forposterior dorsolateral striatum in human habit learning. Eur. J. Neurosci.29,2225–2232

Generation 2: Implications for the field

Similarly to rodents:

- humans show a transition from goal-directed to habitual behaviors following overtraining

- Behavioral dissociation between goal-directed and habitual behavior corresponds to a neural dissociation:
  - **Medial orbitofrontal cortex (OFC)** → supports goal-directed behavior
  - **Dorsolateral striatum** → supports habitual behavior



Generation 3

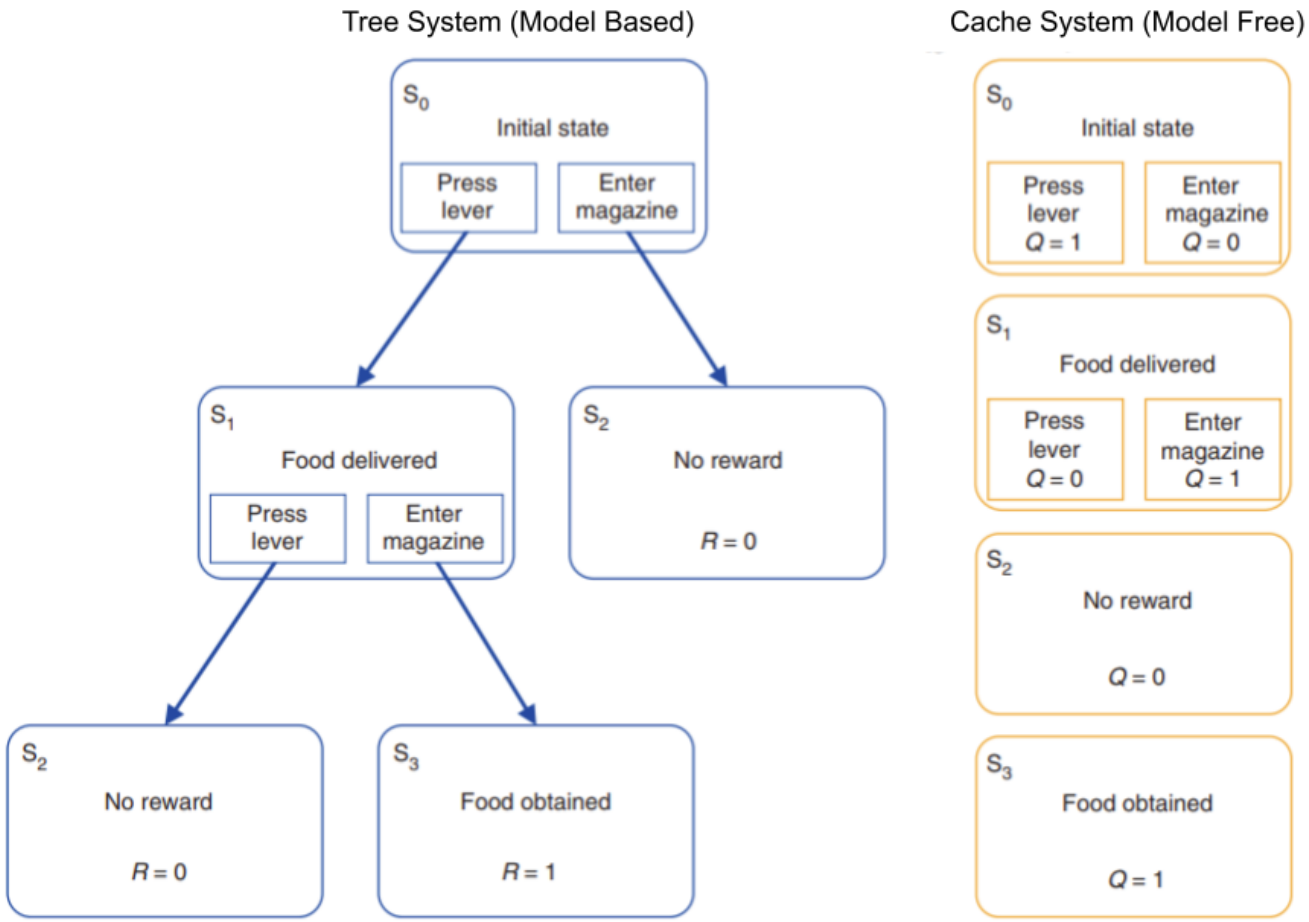
model-based vs model-free computational analyses
[computational neuroscience]

Generation 3: model-based vs model-free computational analyses

Computational formalization of

- Goal-directed actions --> model-based
- Habitual actions --> model-free
- Their interaction

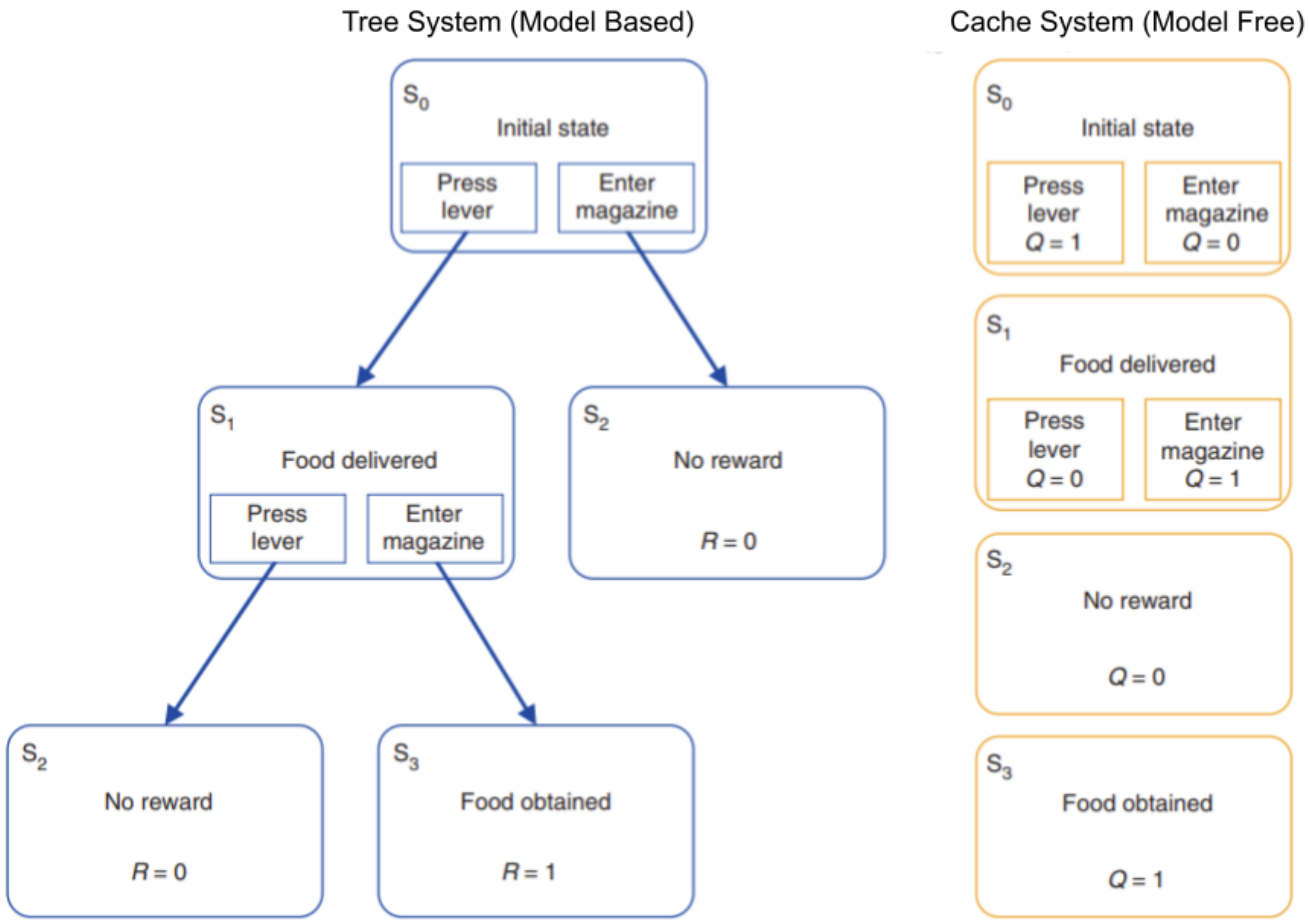
Model means anything an agent can use to predict how its environment will respond to its actions in terms of state transitions and rewards



Daw et al (2005)

Generation 3: model-based vs model-free computational analyses

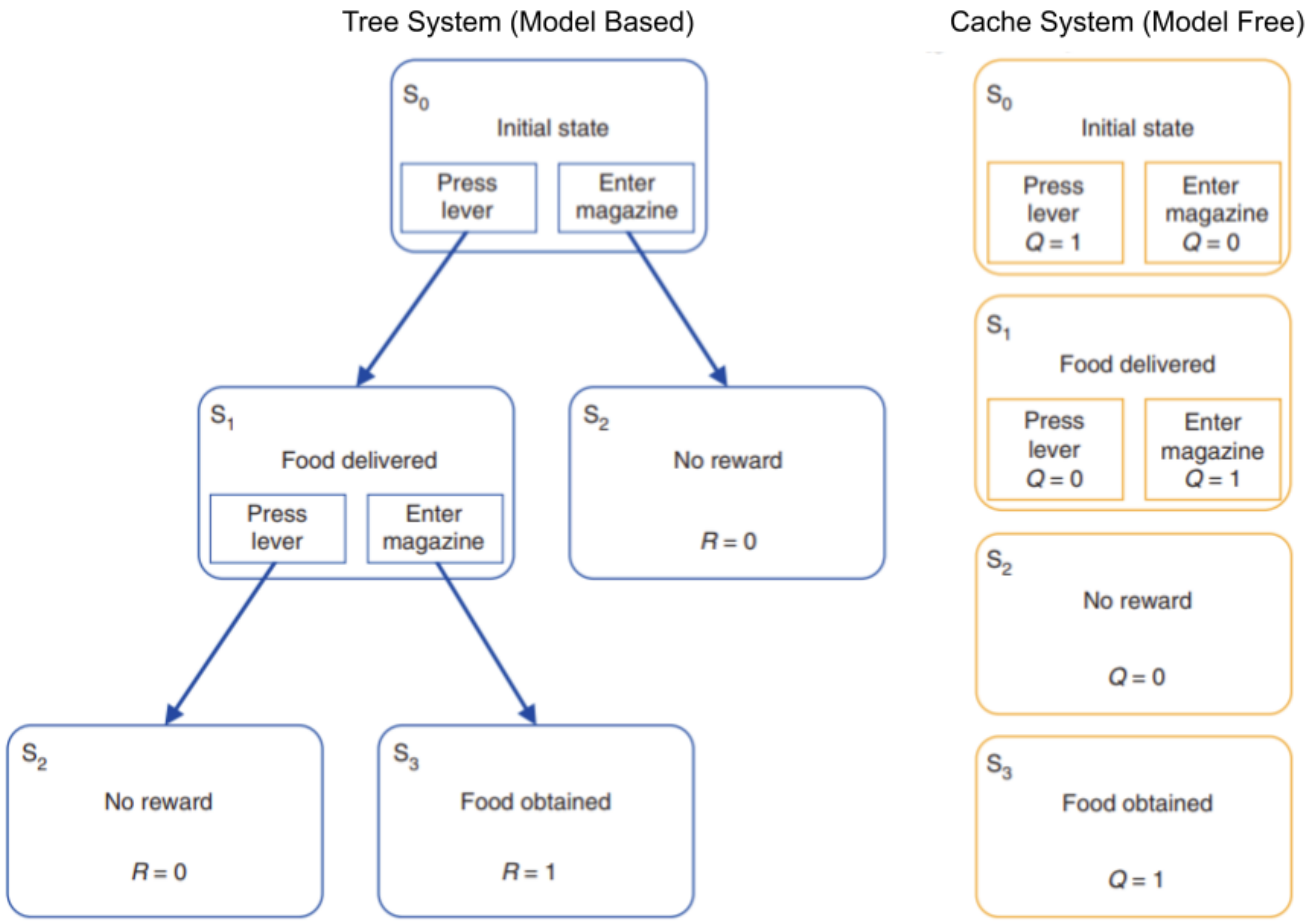
- Goal-directed actions --> model-based
  - A model-based algorithm selects actions by using a model to predict the consequences of possible courses of action in terms of future states and the reward signals expected to arise from those states
- Habitual actions --> model-free
  - A model-free algorithm selects actions relying on stored action values for all the state–action pairs obtained over many learning trials



Daw et al (2005)

Generation 3: model-based vs model-free computational analyses

- Goal-directed actions --> model-based
  - When the environment of a model-based agent changes the way it reacts to the agent's actions, the agent can update the value (policy) of future states without the need to move to them.
- Habitual actions --> model-free
  - When the environment of a model-free agent changes the way it reacts to the agent's actions, the agent has to move to that state, act from it, possibly many times, and experience the consequences of its actions.



Daw et al (2005)

Let's try this:

On a computer (sorry, not a phone) go to

https://nivlab.github.io/jspsych-demos/tasks/two-step/experiment.html

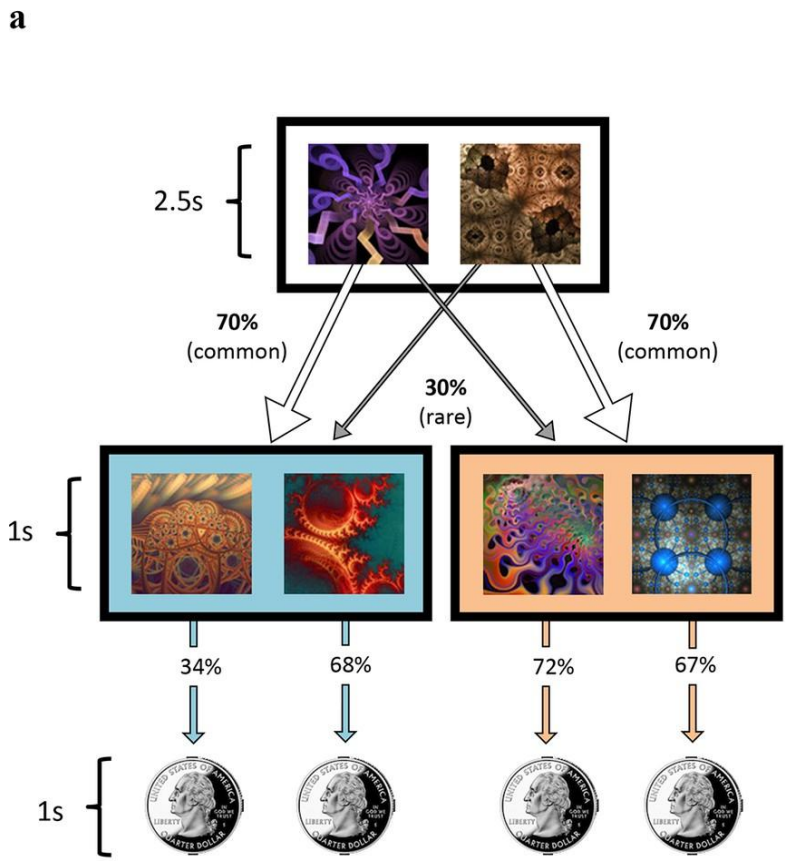
and play through instructions & game for a while

https://nivlab.github.io/jspsych-demos/

Sequential two-choice Markov decision tasks

Developed to

- discern the influence of model-free vs model-based controller on behavior
- to determine whether neural signals are correlated with predictions and prediction errors specific to each controller



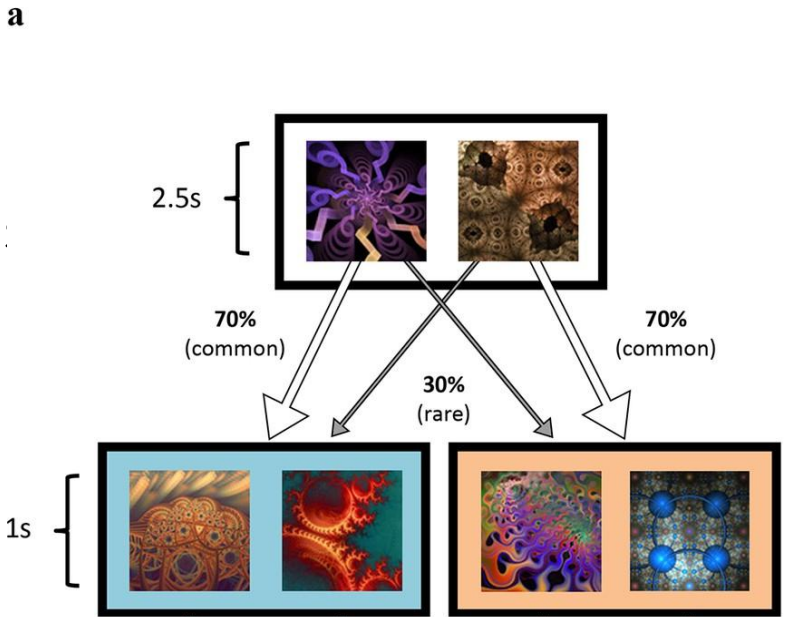
https://doi.org/10.7554/eLife.11305.003

Sequential two-choice Markov decision tasks

"Your task is to maximize the reward"

(**a**) Subjects chose between two fractal which probabilistically determined whether they would transition to the orange or blue second stage state.

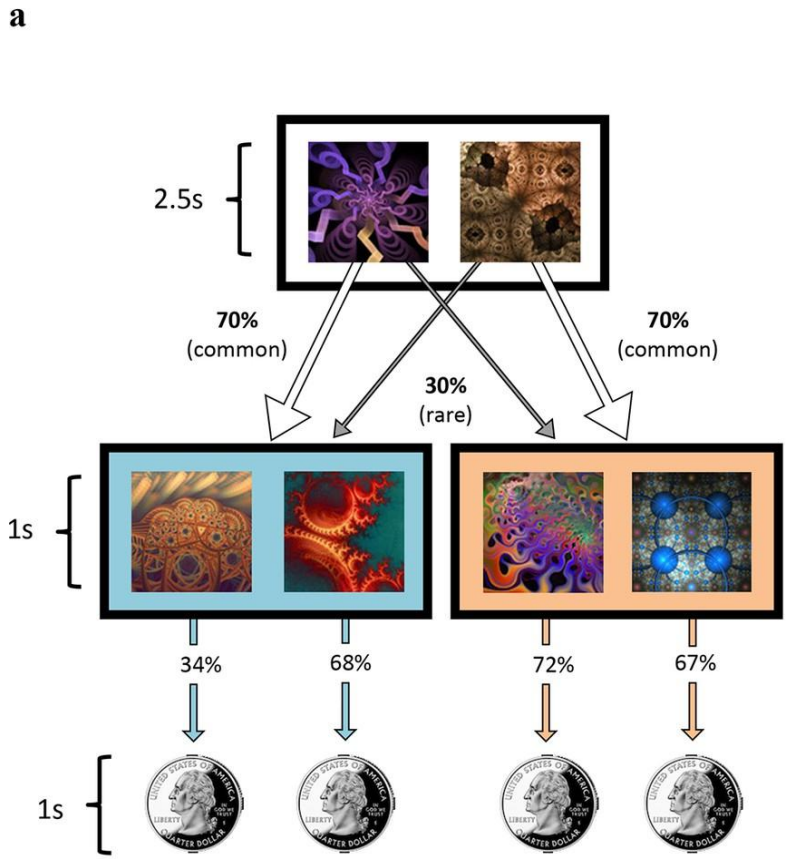
Action at the first state is associated with one likely and one unlikely transition. For example, the fractal on the left had a 70% chance of leading to the blue second stage state ('common' transition) and a 30% chance of leading to the orange state ('rare' transition). **These transition probabilities were fixed and could be learned over time.**



https://doi.org/10.7554/eLife.11305.003

Sequential two-choice Markov decision tasks

(**a**) In the second stage state, subjects chose between two fractals, each of which was associated with a distinct probability of being rewarded with a 25 cents coin.



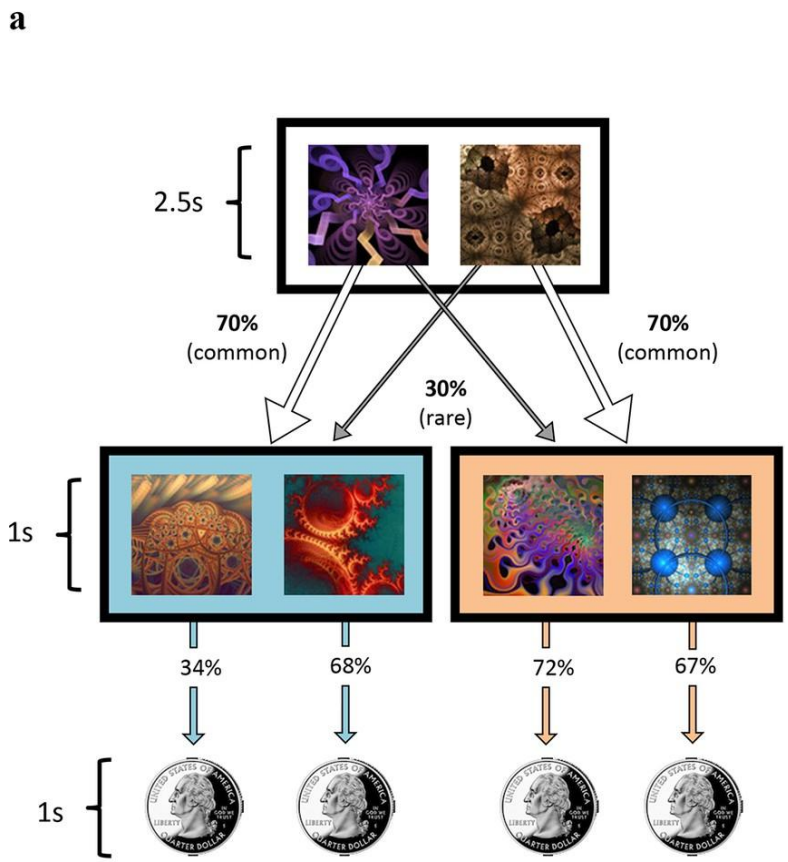
https://doi.org/10.7554/eLife.11305.003

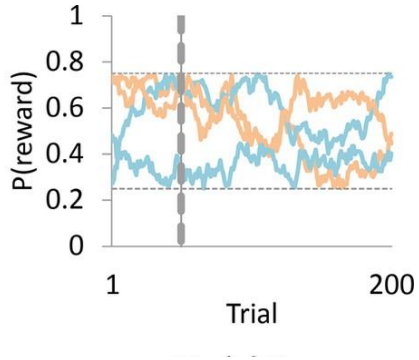
Sequential two-choice Markov decision tasks

(**b**) Drifting reward probabilities determined by Gaussian Random Walks for 200 trials with grey horizontal lines indicating boundaries at 0.25 and 0.75.

To incentivize subjects to continue learning throughout the task, the chances of p**ay off associated with the four second-stage options were changed** slowly and independently.

The chance of winning is almost stochastic.

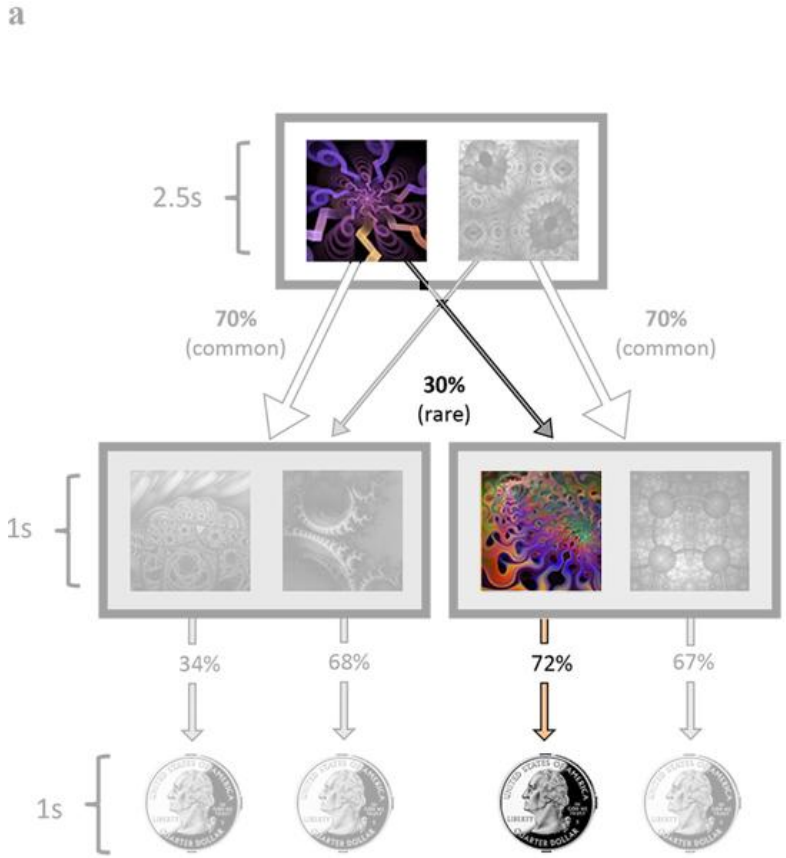




https://doi.org/10.7554/eLife.11305.003

Sequential two-choice Markov decision tasks

How does bottom-stage outcome affect top-stage choices?



https://doi.org/10.7554/eLife.11305.003

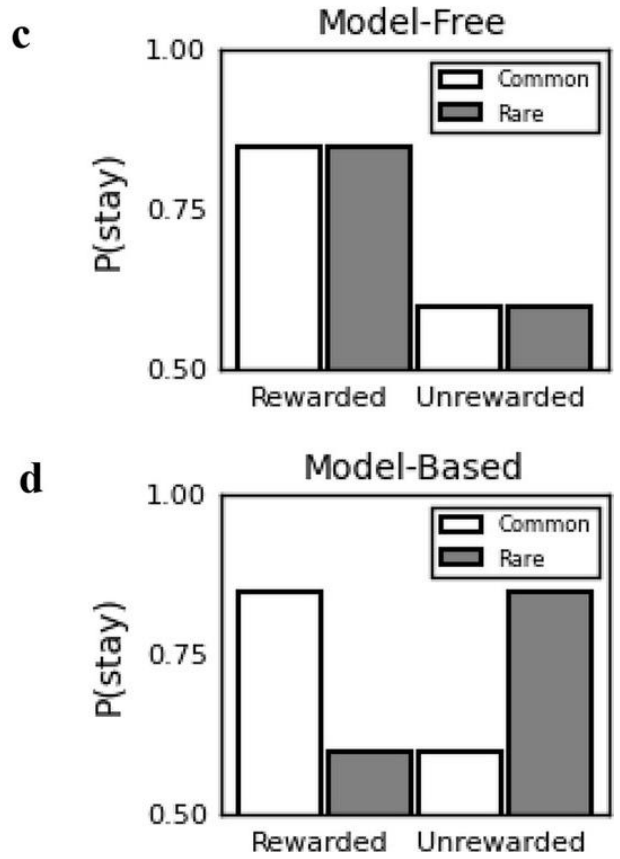
Sequential two-choice Markov decision tasks

https://doi.org/10.7554/eLife.11305.003

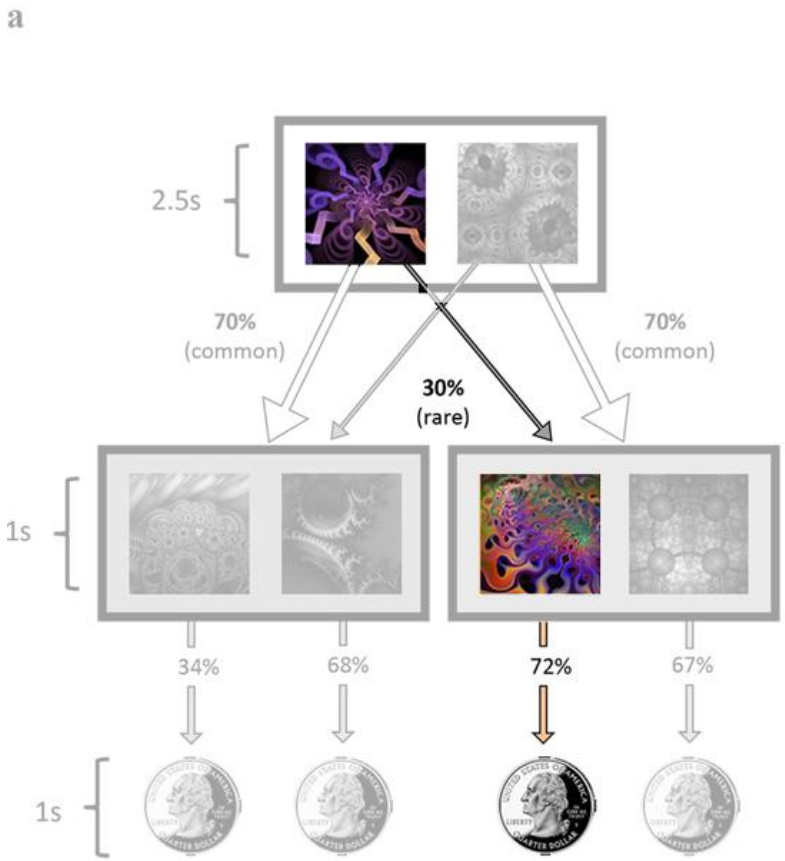
How does bottom-stage outcome affect top-stage choices?

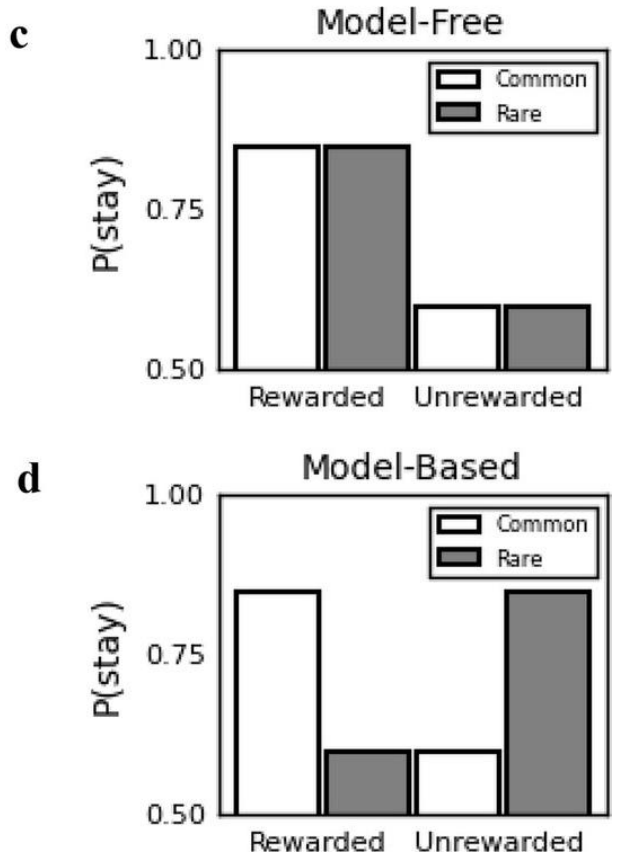
Model-free and model-based agents differ in the action selected after a **rare transition**.

- Model-free agent
  - ignores transition structure
  - prefers to repeat actions that lead to reward, irrespective of the likelihood of that first transition.
- Model-based agent
  - respects transition structure
  - can ascribe rewards following a rare transition to an alternative (non-selected) action—which, despite not predicting reward on the current trial, will be more likely to lead to reward on future trials.
  - model-based strategy predicts a crossover interaction between the two factors (reward and transition), because a rare transition inverts the effect of the subsequent reward.



Sequential two-choice Markov decision tasks



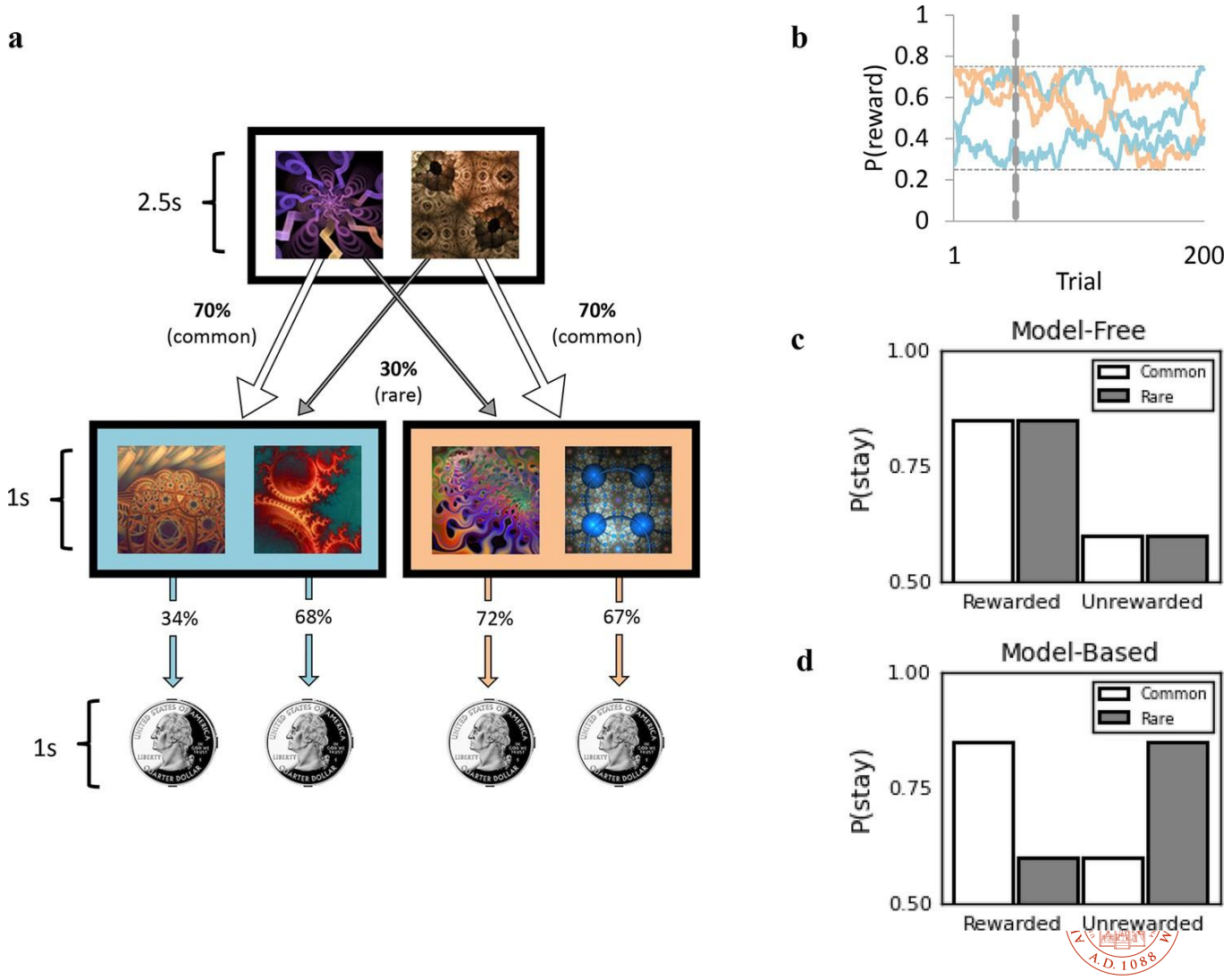


https://doi.org/10.7554/eLife.11305.003

Sequential two-choice Markov decision tasks

(**c**) Schematic representing the performance of a purely 'model-free' learner, who only exhibits sensitivity to whether or not the previous trial was rewarded vs. unrewarded, and does not modify their behavior in light of the transition that preceded reward.

(**d**) Schematic representing the performance of a purely 'model-based' learner, who is more likely to repeat an action (i.e. 'stay') following a rewarded trial, only if the transition was common. If the transition to that rewarded state was rare, they are more likely to switch on the next trial.

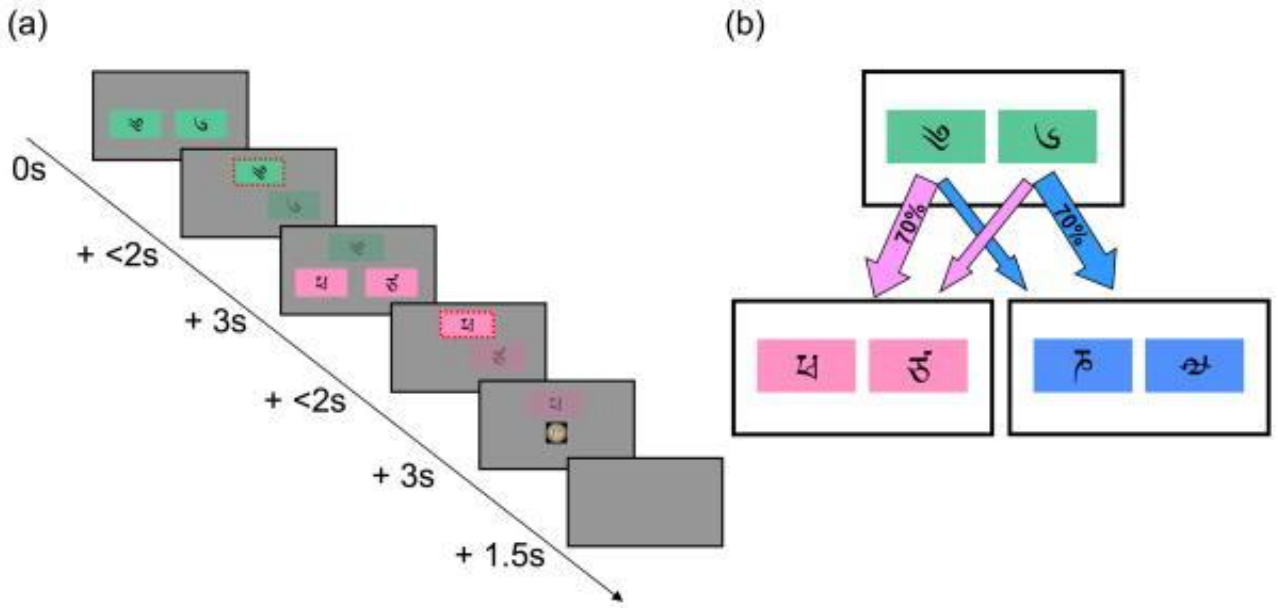


https://doi.org/10.7554/eLife.11305.003

Detecting simultaneous correlates of model-free and model-based systems

Method

(A) Timeline of events in trial. A first-stage choice between two options (green boxes) leads to a second-stage choice(here, between two pink options), which is reinforced with money.(B) State transition structure. Each first-stage choice is predominantly associated with one or the other of the second-stage states, and leads there 70% of the time.



To incentivize subjects to continue learning throughout the task, the chances of pay off associated with the four second-stage options are changed slowly and independently

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. *Neuron*. 2011;69(6):1204-1215. doi:10.1016/j.neuron.2011.02.027

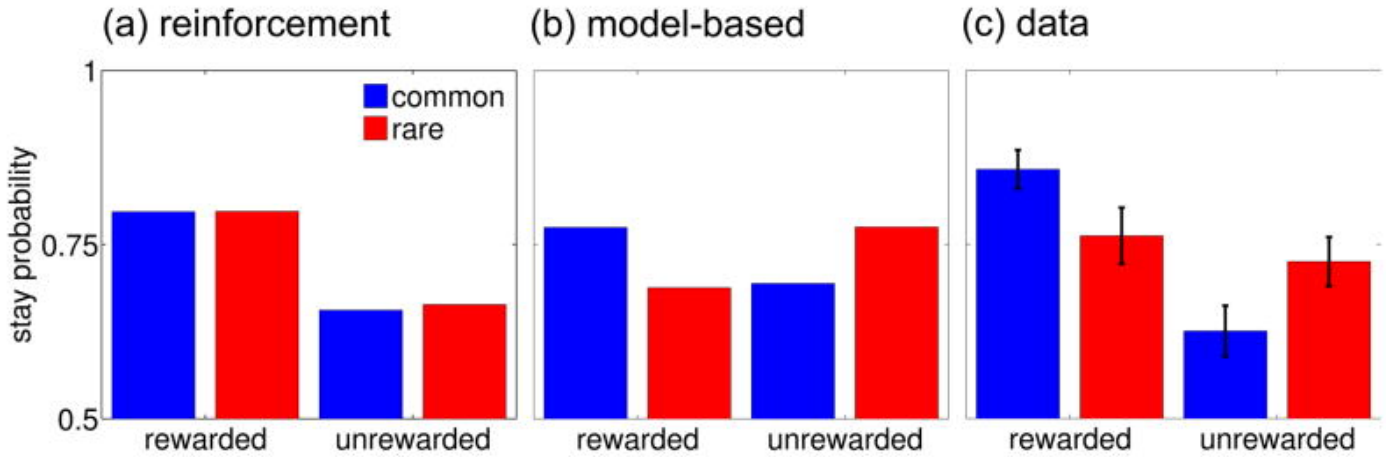
Detecting simultaneous correlates of model-free and model-based systems

Results: Analysis of choice behavior.

(a) Simple reinforcement predicts that a first-stage choice resulting in reward is more likely to be repeated on the subsequent trial, regardless of whether that reward occurred after a common or rare transition.

(b) Model-based prospective evaluation instead predicts that a rare transition should affect the value of the other first-stage option, leading to a predicted interaction between the factors of reward and transition probability.

(c) Actual stay proportions, averaged across subjects, display hallmarks of both strategies. Error bars: 1 SEM.



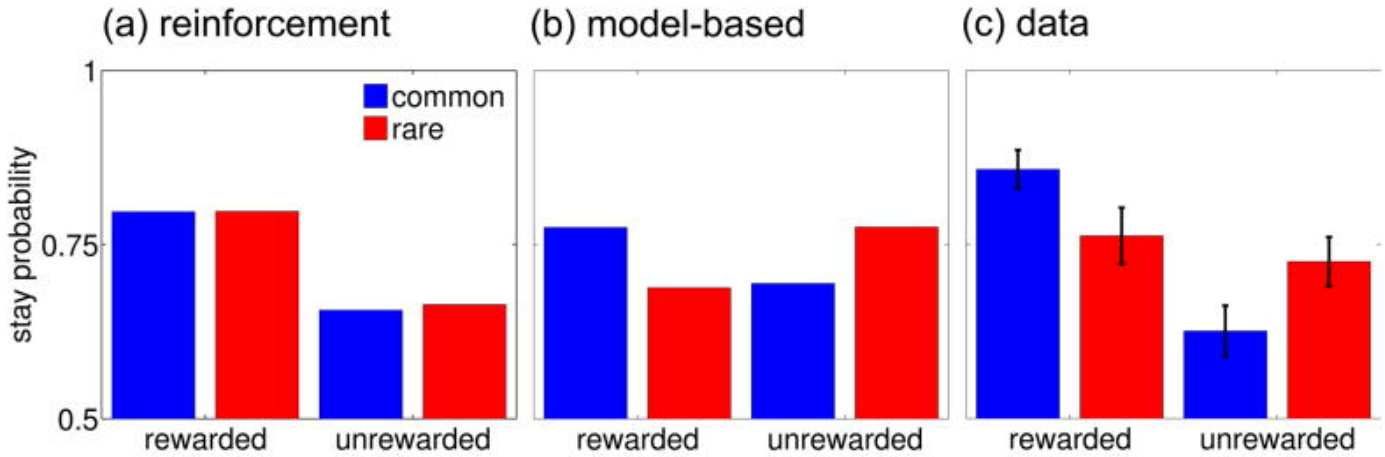
Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. *Neuron*. 2011;69(6):1204-1215. doi:10.1016/j.neuron.2011.02.027

Detecting simultaneous correlates of model-free and model-based systems

Results: Computational

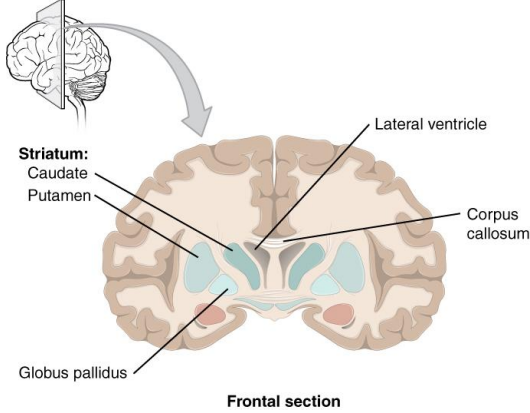
Choice behavior during was best explained by hybrid model that integrates both

- Reward PE: model-free (similar TD model)
- State PE: model-based



Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. *Neuron*. 2011;69(6):1204-1215. doi:10.1016/j.neuron.2011.02.027

Detecting simultaneous correlates of model-free and model-based systems

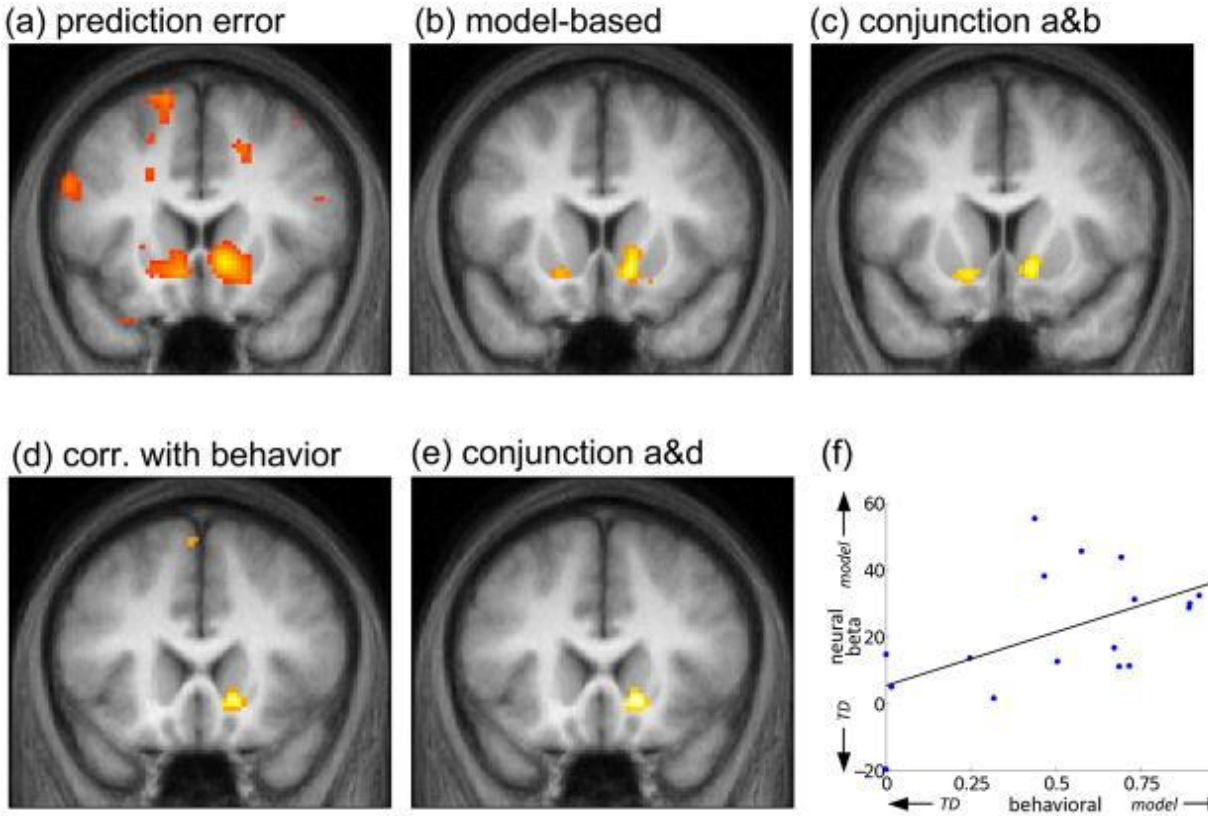


Results: Neural

Parameters estimated from computational models were used to find activations that correlated with SPE/model-based & RPE/model-free

Activity in striatum occurred both for model-free and model-based prediction error.

This activity correlated with the extent to which that subject's behavior was model based.

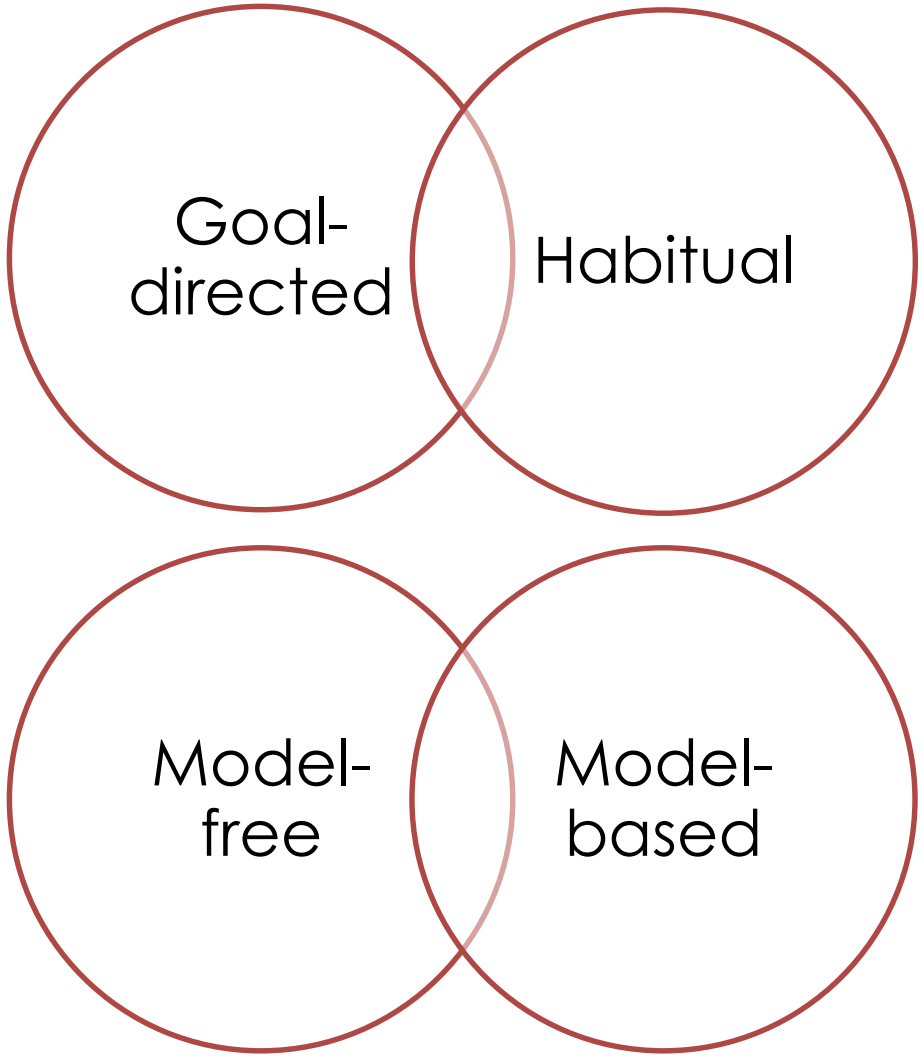


Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. *Neuron*. 2011;69(6):1204-1215. doi:10.1016/j.neuron.2011.02.027

Generation 3: Implications for the field

- Challenged the notion of a separate model-based vs model-free learner → suggesting a more **integrated computational and neural architecture** for high-level human decision-making

- In the brain, there is a dynamic **inter-dependency** between goal-directed/model-based and habitual/model-free systems



Recommended readings

- Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007

- Daw, N. D., & O'Doherty, J. P. (2014). Multiple systems for value learning. In Neuroeconomics (Chapter 21, pp. 393-410). Academic Press.

Revision questions

- Discuss the historical progression of thought regarding learning and behavioral control, tracing the evolution from stimulus-response theories to the understanding of goal-directed and habitual actions as distinct but interacting systems.
- Critically evaluate the experimental methodologies used in animal and human studies to dissociate goal-directed and habitual behaviors. What are the strengths and limitations of reinforcer devaluation and contingency degradation procedures?
- Explore the neural substrates underlying goal-directed and habitual behaviors in both rodent and human brains. How do the findings from animal studies translate to our understanding of human decision-making, and what role do different regions of the striatum and prefrontal cortex play?
- Explain the computational distinction between model-based and model-free reinforcement learning. How do these computational frameworks help us understand the cognitive processes underlying goal-directed and habitual control, and what evidence supports the idea that both systems contribute to behavior?
- Consider the implications of the dual-systems perspective of behavioral control (goal-directed vs. habitual, or model-based vs. model-free) for understanding various aspects of human behavior, such as addiction, learning new skills, and adapting to changing environments.

Glossary of key terms

- Stimulus-Response (S-R) Theory: A view of learning that emphasizes the formation of direct associations between environmental stimuli and behavioral responses, dependent on past reinforcement.
- Field Theory (Cognitive Map): A perspective suggesting that learning involves the creation of a mental representation of the environment, rather than just reinforcers, which guides behavior.
- Latent Learning: Learning that occurs without explicit reinforcement or behavioral manifestation until there is sufficient motivation.
- Reinforcer Devaluation: An experimental procedure where the value or desirability of a reward associated with an action is reduced, e.g. through satiation, used to assess whether the action is goal-directed (sensitive to the devaluation) or habitual (insensitive).
- Contingency Degradation: An experimental manipulation where the original relationship between an action and its outcome (reward) is weakened or removed, used to distinguish goal-directed (sensitive to degradation) from habitual control (insensitive).
- Striatum: A subcortical brain structure, part of the basal ganglia, implicated in both motor control and reward-based learning; divided into regions like the dorsomedial striatum (DMS) and dorsolateral striatum (DLS).

Glossary of key terms

- Dorsomedial Striatum (DMS): A region of the striatum primarily associated with goal-directed behavior in rodents.
- Dorsolateral Striatum (DLS): A region of the striatum primarily associated with habitual behavior in rodents.
- Medial Orbitofrontal Cortex (OFC): A region of the prefrontal cortex in humans implicated in processing the value of outcomes and supporting goal-directed behavior.
- Model-Based Learning: A computational approach to decision-making that involves building an internal model of the environment, including state transitions and rewards, to predict the consequences of actions.
- Model-Free Learning: A computational approach to decision-making that learns the value of actions directly through trial and error, without explicitly modeling the environment.
- Prediction Error (PE): The difference between an expected outcome and the actual outcome, used in reinforcement learning to update value estimates.
- Sequential Two-Choice Markov Decision Task: An experimental paradigm used in computational neuroscience to dissociate model-based and model-free contributions to decision-making by examining how choices are influenced by reward history and knowledge of environmental transitions.