Introduction to animal reinforcement learning

Cognition and Neuroscience

Academic year 2023/2024

Francesca Starita

francesca.starita2@unibo.it

Reinforcement Learning

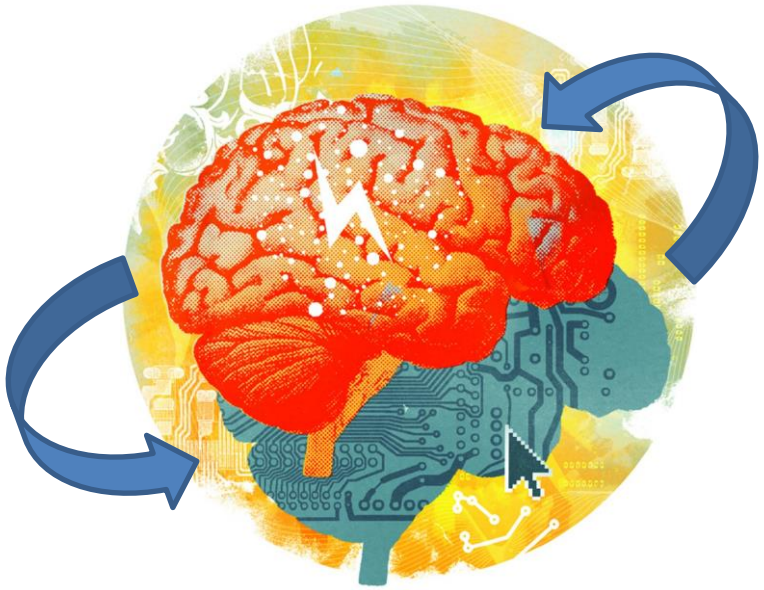
Alongside its important role in the development of deep learning, neuroscience was also instrumental in erecting a second pillar of contemporary AI, stimulating the emergence of the field of reinforcement learning (RL). RL methods address the problem of how to maximize future reward by mapping states in the environment to actions and are among the most widely used tools in AI research (Sutton and Barto, 1998). Although it is not widely appreciated among AI researchers, RL methods were originally inspired by research into animal learning. In particular, the development of temporal-difference (TD) methods, a critical component of many RL models, was inextricably intertwined with research into animal behavior in conditioning experiments. TD methods are real-time models that learn from differences between temporally successive predictions, rather than having to wait until the actual reward is delivered. Of particular relevance was an effect called second-order conditioning, where affective significance is conferred on a conditioned stimulus (CS) through association with another CS rather than directly via association with the unconditioned stimulus (Sutton and Barto, 1981). TD learning provides a natural explanation for second-order conditioning and indeed has gone on to explain a much wider range of findings from neuroscience, as we discuss below.

Here, as in the case of deep learning, investigations initially inspired by observations from neuroscience led to further developments that have strongly shaped the direction of AI research. From their neuroscience-informed origins, TD methods and related techniques have gone on to supply the core technology for recent advances in AI, ranging from robotic control (Hafner and Riedmiller, 2011) to expert play in backgammon (Tesauro, 1995) and Go (Silver et al., 2016).

Neuron
Review 2017

Neuroscience-Inspired Artificial Intelligence

Demis Hassabis,[1,2,*] Dharshan Kumaran,[1,3] Christopher Summerfield,[1,4] and Matthew Botvinick[1,2]



Reinforcement Learning

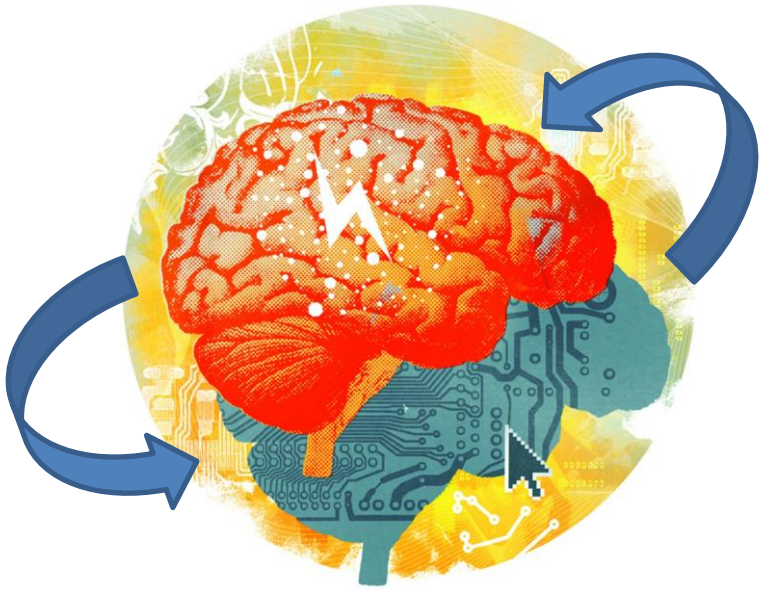
Alongside its important role in the development of deep learning, neuroscience was also instrumental in erecting a second pillar of contemporary AI, stimulating the emergence of the field of reinforcement learning (RL). RL methods address the problem of how to maximize future reward by mapping states in the environment to actions and are among the most widely used tools in AI research (Sutton and Barto, 1998). Although it is not widely appreciated among AI researchers, RL methods were originally inspired by research into animal learning. In particular, the development of temporal-difference (TD) methods, a critical component of many RL models, was inextricably intertwined with research into animal behavior in conditioning experiments. TD methods are real-time models that learn from differences between temporally successive predictions, rather than having to wait until the actual reward is delivered. Of particular relevance was an effect called second-order conditioning, where affective significance is conferred on a conditioned stimulus (CS) through association with another CS rather than directly via association with the unconditioned stimulus (Sutton and Barto, 1981). TD learning provides a natural explanation for second-order conditioning and indeed has gone on to explain a much wider range of findings from neuroscience, as we discuss below.

Here, as in the case of deep learning, investigations initially inspired by observations from neuroscience led to further developments that have strongly shaped the direction of AI research. From their neuroscience-informed origins, TD methods and related techniques have gone on to supply the core technology for recent advances in AI, ranging from robotic control (Hafner and Riedmiller, 2011) to expert play in backgammon (Tesauro, 1995) and Go (Silver et al., 2016).

Neuron
Review 2017

Neuroscience-Inspired Artificial Intelligence

Demis Hassabis,[1,2,*] Dharshan Kumaran,[1,3] Christopher Summerfield,[1,4] and Matthew Botvinick[1,2]
[1]DeepMind, 5 New Street Square, London, UK
[2]Gatsby Computational Neuroscience Unit, 25 Howland Street, London, UK
[3]Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London, UK
[4]Department of Experimental Psychology, University of Oxford, Oxford, UK
*Correspondence: dhcontact@google.com
http://dx.doi.org/10.1016/j.neuron.2017.06.011



Decisions, decisions, decisions!

Decisions, decisions, decisions!

Optimal decision making:

Decisions, decisions, decisions!

Optimal decision making:

- Maximize rewards
- Minimize punishments

Why is it hard?

Decisions, decisions, decisions!

Optimal decision making:

- Maximize rewards
- Minimize punishments

Why is it hard?

- Outcome (i.e. reward/punishment) may be delayed
- Outcomes may depend on a series of actions

Decisions, decisions, decisions!

Optimal decision making:

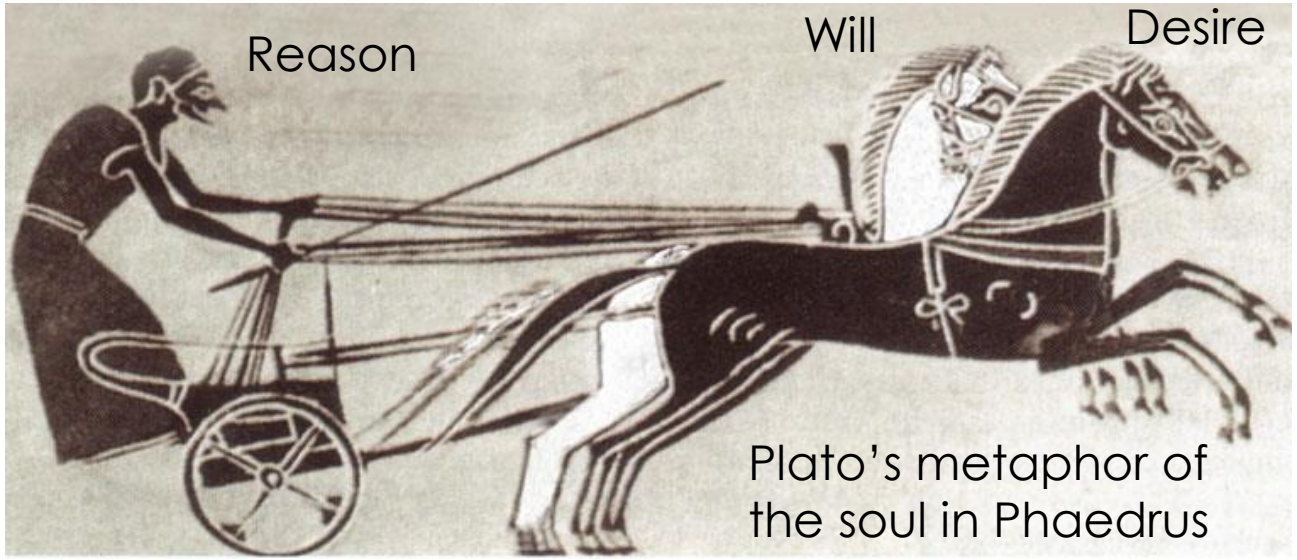
- Maximize rewards
- Minimize punishments

Credit assignment problem

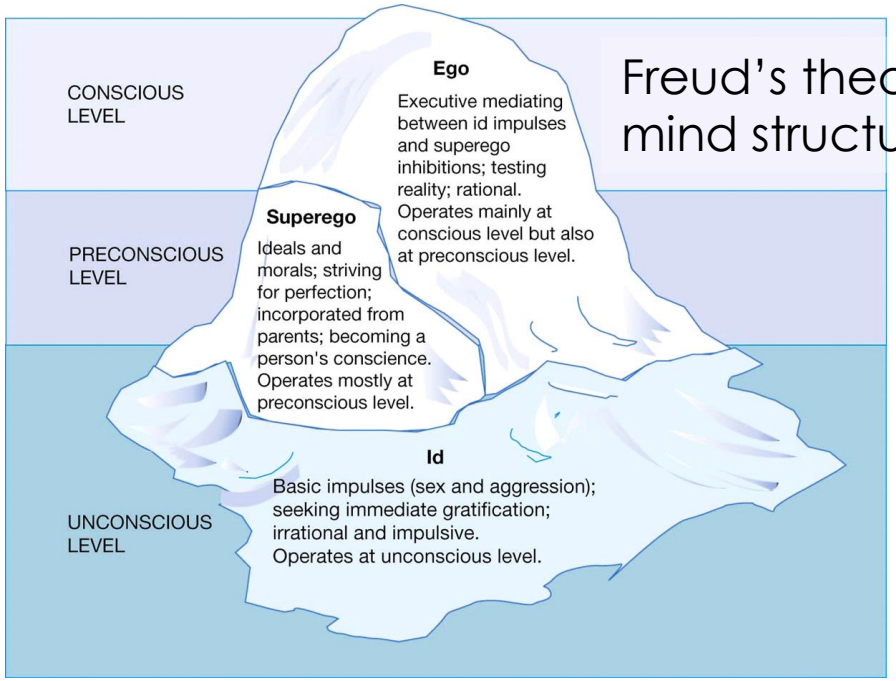
How do you distribute credit for success (or blame for failure) of a decision among the many component structures that could have been involved in producing it?

Multiple systems contribute to learning and controlling behavior in animals

- Human and animal decisions are governed not by a single unitary controller, but rather by multiple, competing sub-systems
- A given behavior can arise in multiple different ways, which are dissociable psychologically, neurally, and computationally
- Multiple roots that lead to a certain decision/behavior/action selection



Plato's metaphor of the soul in Phaedrus



Freud's theory of mind structure

Multiple systems contribute to learning and controlling behavior in animals

Some definitions

- Learning: Enduring change in response or behavior that occurs as a result of experience
- **Non-associative learning**: Change in response or behavior is caused by learning about the properties of a single stimulus: subject is exposed once or repeatedly to a single type of stimulus

Habituation: a **decrease** in an innate response to a stimulus that is presented repeatedly

Sensitization: an **increase** in an innate response to a stimulus that is presented repeatedly

Multiple systems contribute to learning and controlling behavior in animals

Some definitions

- Learning: Enduring change in response or behavior that occurs as a result of experience
- Non-associative learning: Change in response or behavior is caused by learning about the properties of a single stimulus: subject is exposed once or repeatedly to a single type of stimulus
- **Associative learning**: Change in response or behavior is caused by learning about the association of at least **two stimuli or events**
    - **Reinforcement learning**: learning about the association of at least **a neutral stimulus or event** and a **reinforcer**

Reinforcers are stimuli or events that cause a change in response

Primary reinforcer es. pain, food, sex

- A stimulus that is biologically prepared to elicit a response
- It is biologically relevant
- Can be positive or negative

Secondary reinforcer es. money

- A stimulus that comes to elicit a response following associative learning
- A stimulus that has become relevant following associative learning
- Can be positive or negative

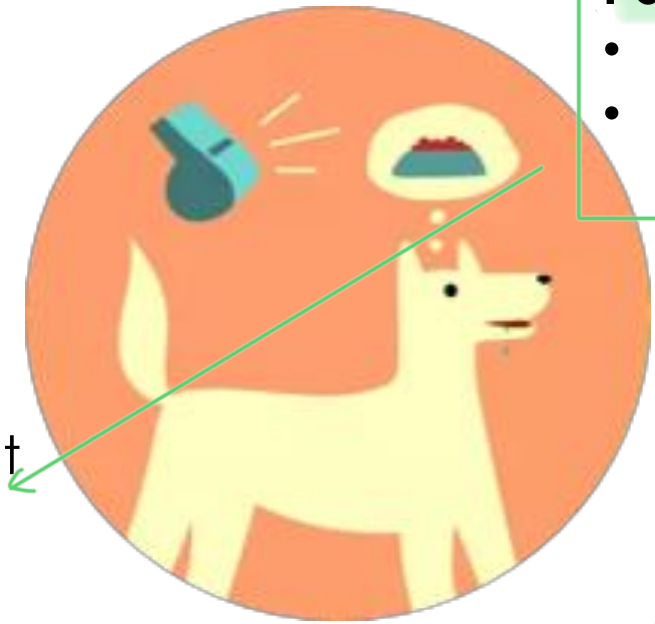
Multiple systems contribute to learning and controlling behavior in animals

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.



Pavlovian system
- Prediction learning
- Learns stimulus-outcome associations

Instrumental system
- Control learning
- Learns action-outcome associations

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

If we can predict what situations are associated with rewards we can try to bring those about through our actions

Multiple systems contribute to learning and controlling behavior in animals

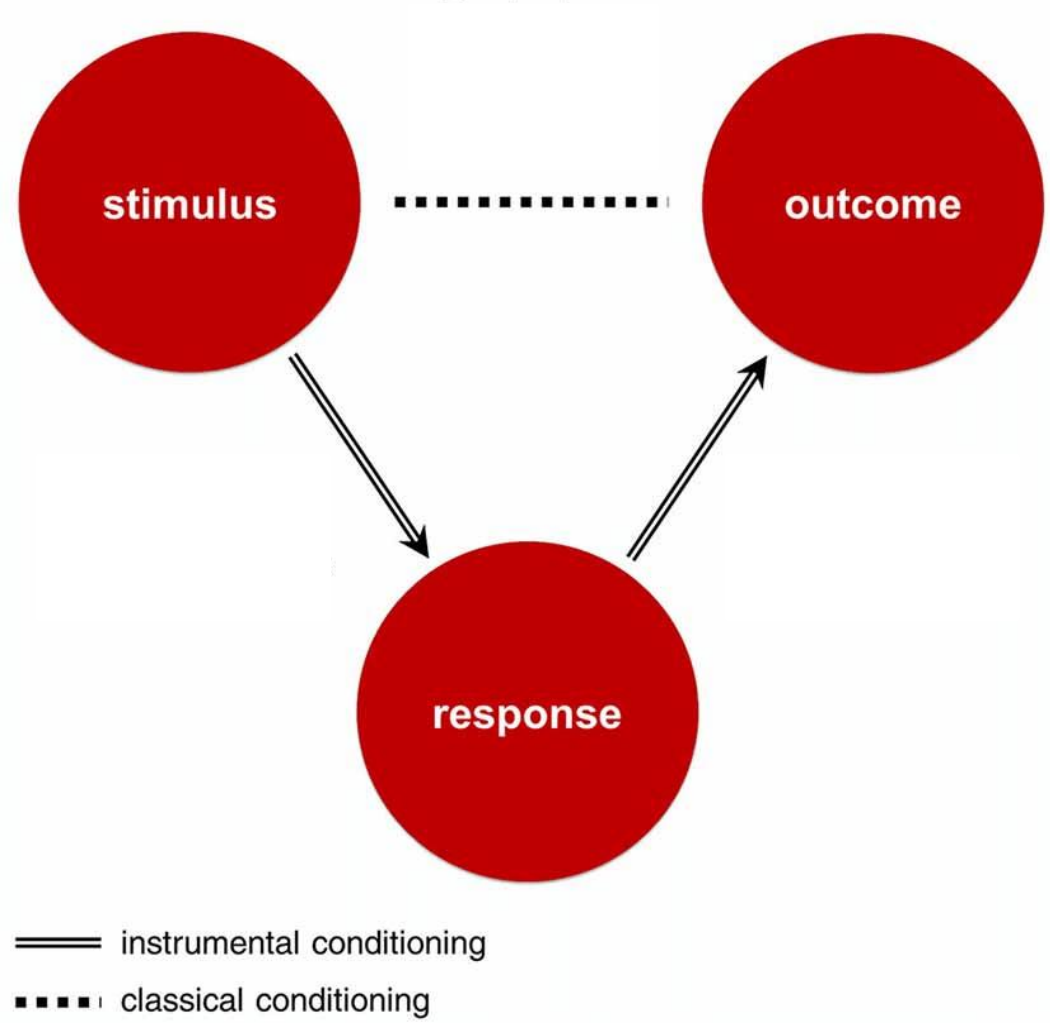
Three learning systems enable organisms to draw on previous experience to **make predictions** about the world and to **select behaviors** appropriate to those predictions:

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.

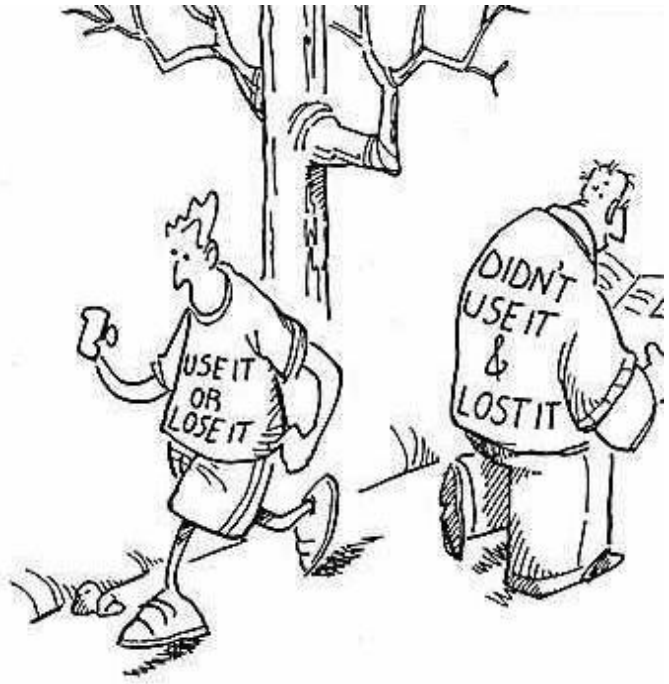
Predictions are for control



Learning at the neuronal level

Learning is the result of changes in the strength of synaptic interactions among neurons in neural networks

Plasticity: Neural connections can be modified by experience & learning



Changes in the strength of synaptic interactions can be:

- **Short-term changes:** functional physiological changes (lasting seconds to hours) that increase or decrease the effectiveness of existing synaptic connections. --> **Hebbian plasticity**

- **Long-term changes:** structural changes (lasting days) that can give rise to further physiological changes that lead to anatomical alterations, including pruning of preexisting synapses or growth of new ones.

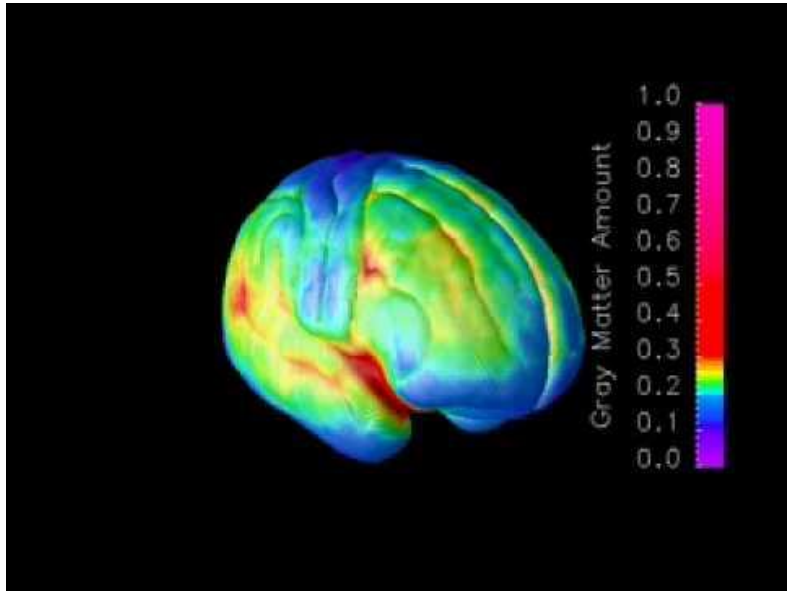
Learning is the result of changes in the strength of synaptic interactions among neurons in neural networks

Plasticity: Neural connections can be modified by experience & learning

Changes in the strength of synaptic interactions can be:

- **Short-term changes:** functional physiological changes (lasting seconds to hours) that increase or decrease the effectiveness of existing synaptic connections. --> **Hebbian plasticity**

- **Long-term changes:** structural changes (lasting days) that can give rise to further physiological changes that lead to anatomical alterations, including pruning of preexisting synapses or growth of new ones.

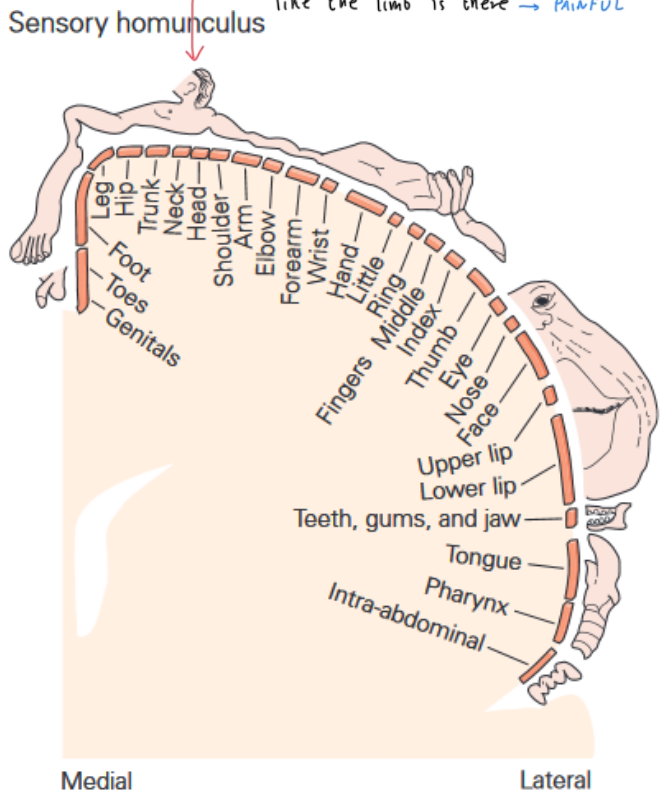


Right oblique view of gray matter maturation over the cortical surface between ages 4 and 21. The side bar shows a color representation in units of GM volume. Gogtay et al., 2004, PNAS https://doi.org/10.1073/pnas.0402680101

A peculiar example of plasticity: phantom limb pain

USES STIMULI FROM NEAR AREAS ← THE PART OF THE BRAIN WHICH REFERS TO THE MISSING LIMB DOESN'T RECEIVE STIMULI ANYMORE AND "GETS HUNGRY" OF THEM

STRUCTURAL CHANGES ARE FOLLOWED BY NEURONAL CHANGES IN ORDER TO CHANGE THE LIMB REPRESENTATION IN THE CORTEX (BODY IMAGE) → so that brain doesn't behave anymore like the limb is there → PAINFUL



https://www.youtube.com/watch?v=1mHIv5ToMTM

Intro to Pavlovian learning

Aka

Pavlovian conditioning

Classical conditioning

https://youtu.be/xnf8i_IRCcw

Pavlovian learning system

- Prediction learning
- Learns stimulus-outcome associations
- Learn to predict
    - when reinforces are likely to occur
    - which stimuli tend to precede those reinforcers
- These predictions enable the animal to emit reflexive responses in **anticipation** of reinforcers, instead of responding exclusively in a reactive manner once reinforcers have occurred

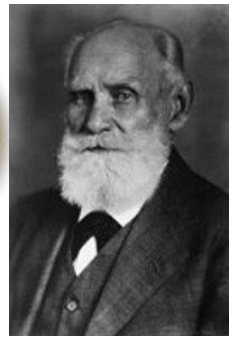
Pavlovian learning involves associating a stimulus with an outcome

a) A stimulus is presented that has no meaning to an animal, such as the sound of a bell, there is no response (NR)

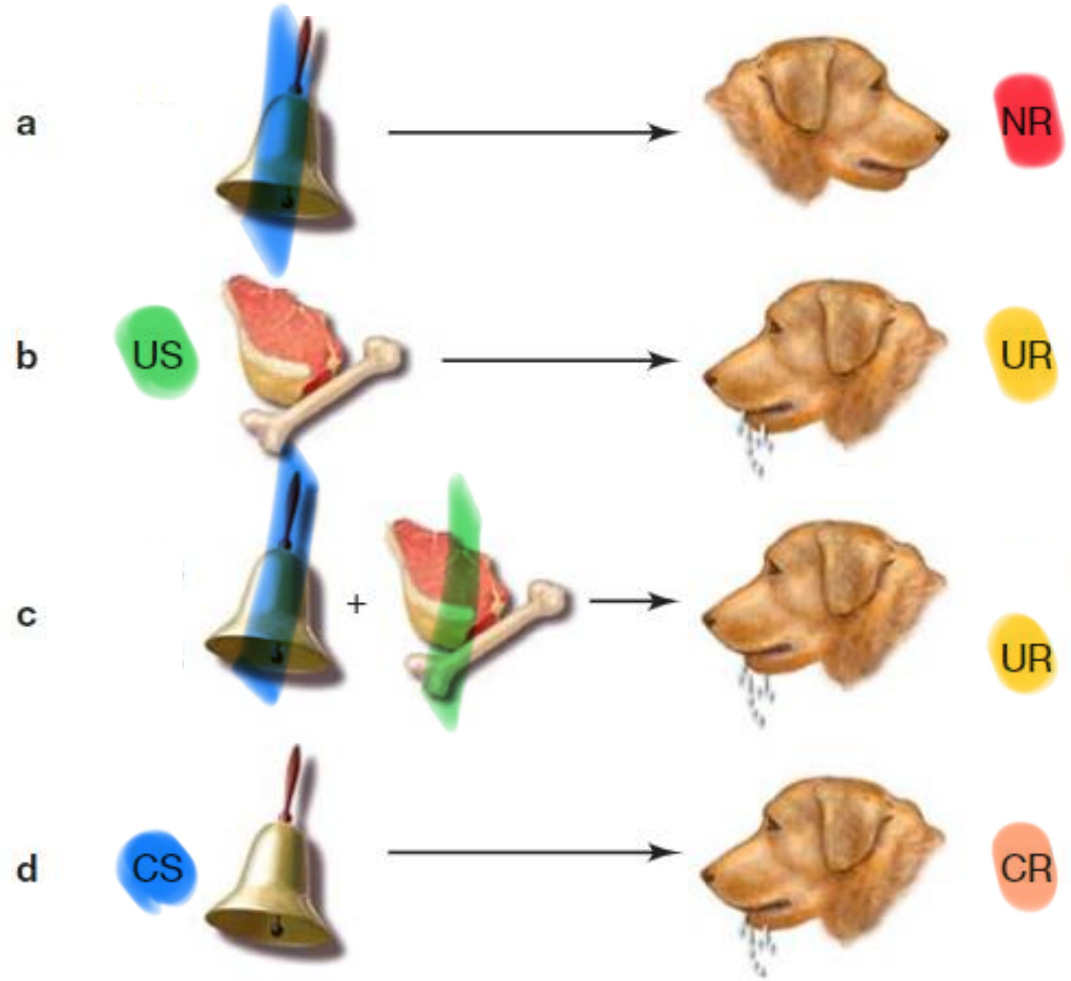
b) Presentation of a reinforcer like food (i.e. unconditioned stimulus, US) generates an unconditioned response (UR)

c) When the sound is paired with the food, the animal learns the association

d) the newly conditioned stimulus (CS) alone can elicit the response, which is now called a conditioned response (CR)

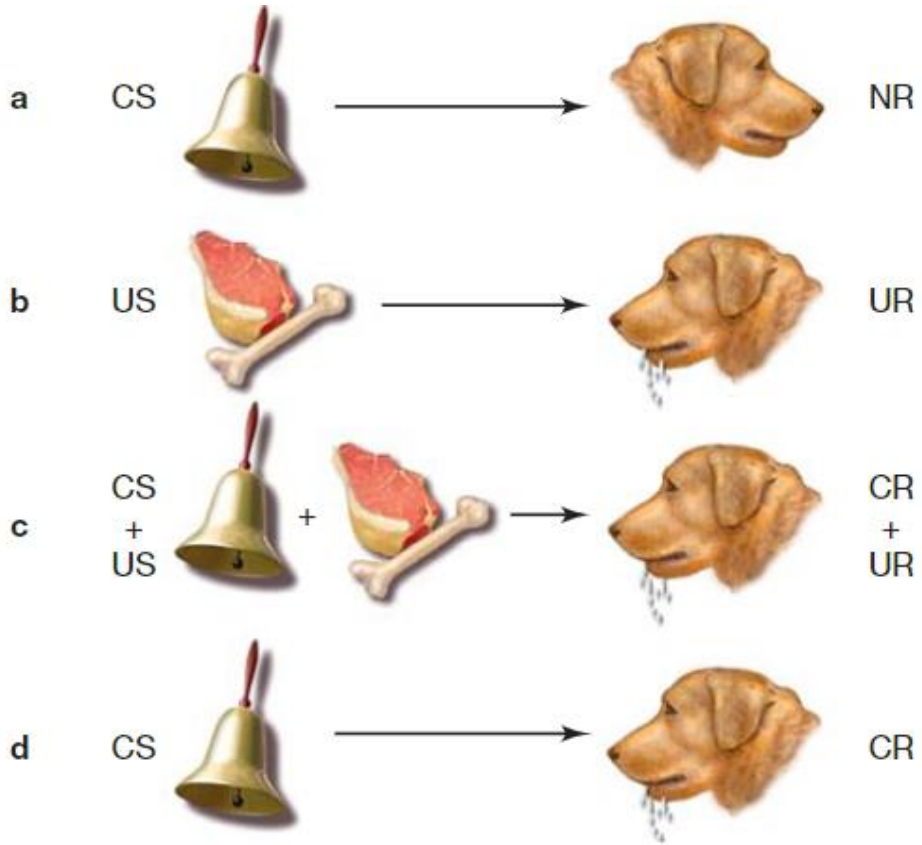


Ivan Pavlov (1849–1936) received a Nobel Prize after first demonstrating this type of learning with his dogs

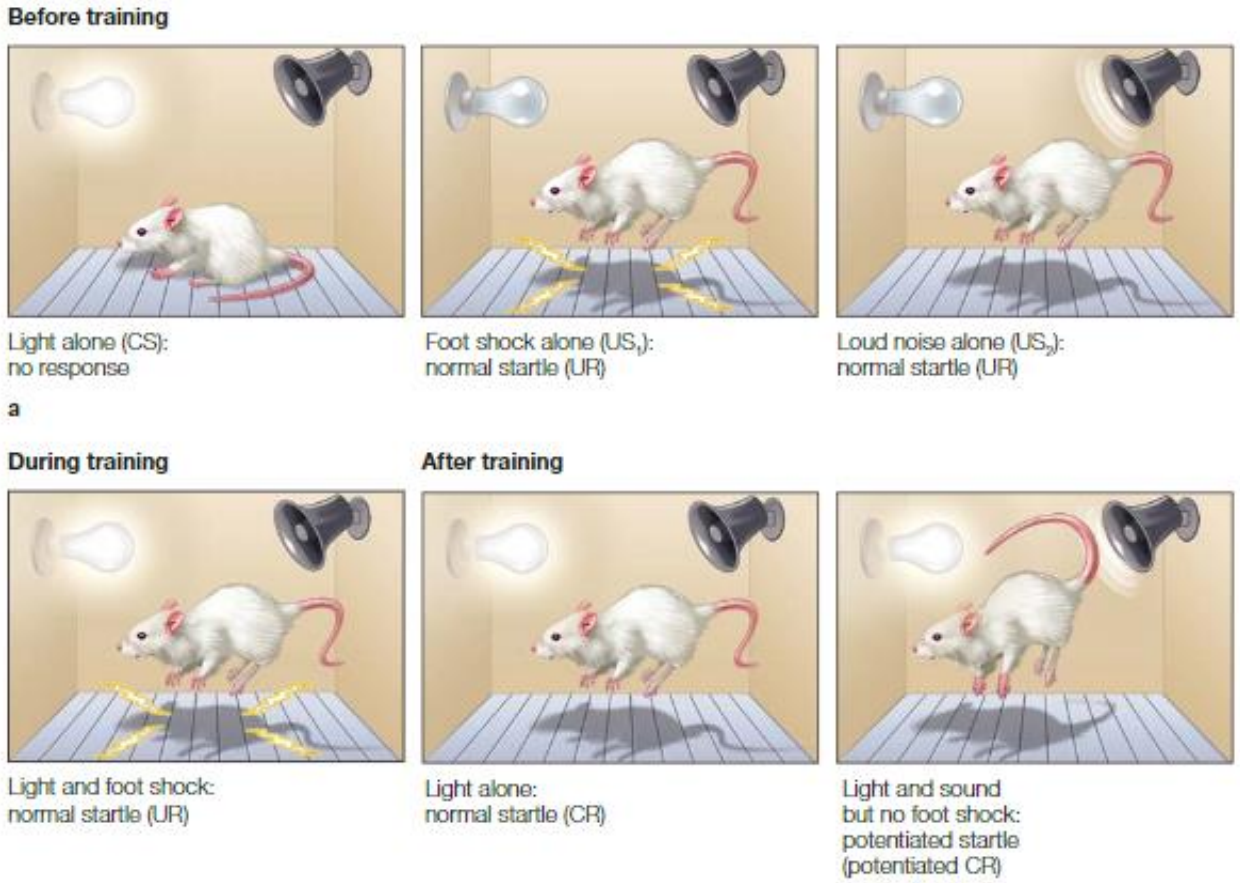


The nature of the outcome triggers different responses

Appetitive



Aversive



https://youtu.be/FMnhyGozLyE

Behaviorism: only observable behavior can be studied

John B. Watson (1878–1958)

Watson proposed that psychology could be objective only if it were based on **observable behavior**. All talk of mental processes, which cannot be publicly observed, should be avoided

Learning was the key, everybody had the same neural equipment on which learning could build.

The brain as a **blank slate** upon which to build through learning and experience.

Frequency of outcome delivery also shapes responses

Continuous Reinforcement

- the CS is reinforced with the US every single time it occurs
- most effective when trying to teach a new association

Partial reinforcement

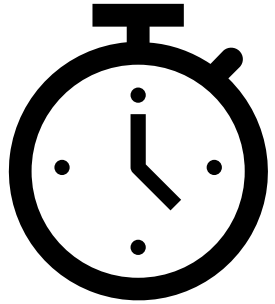
- the CS is reinforced only part of the time
- Associations are acquired more slowly with partial reinforcement, but the response is more resistant to extinction

Pavlovian learning in everyday life

Can you come up with some examples?

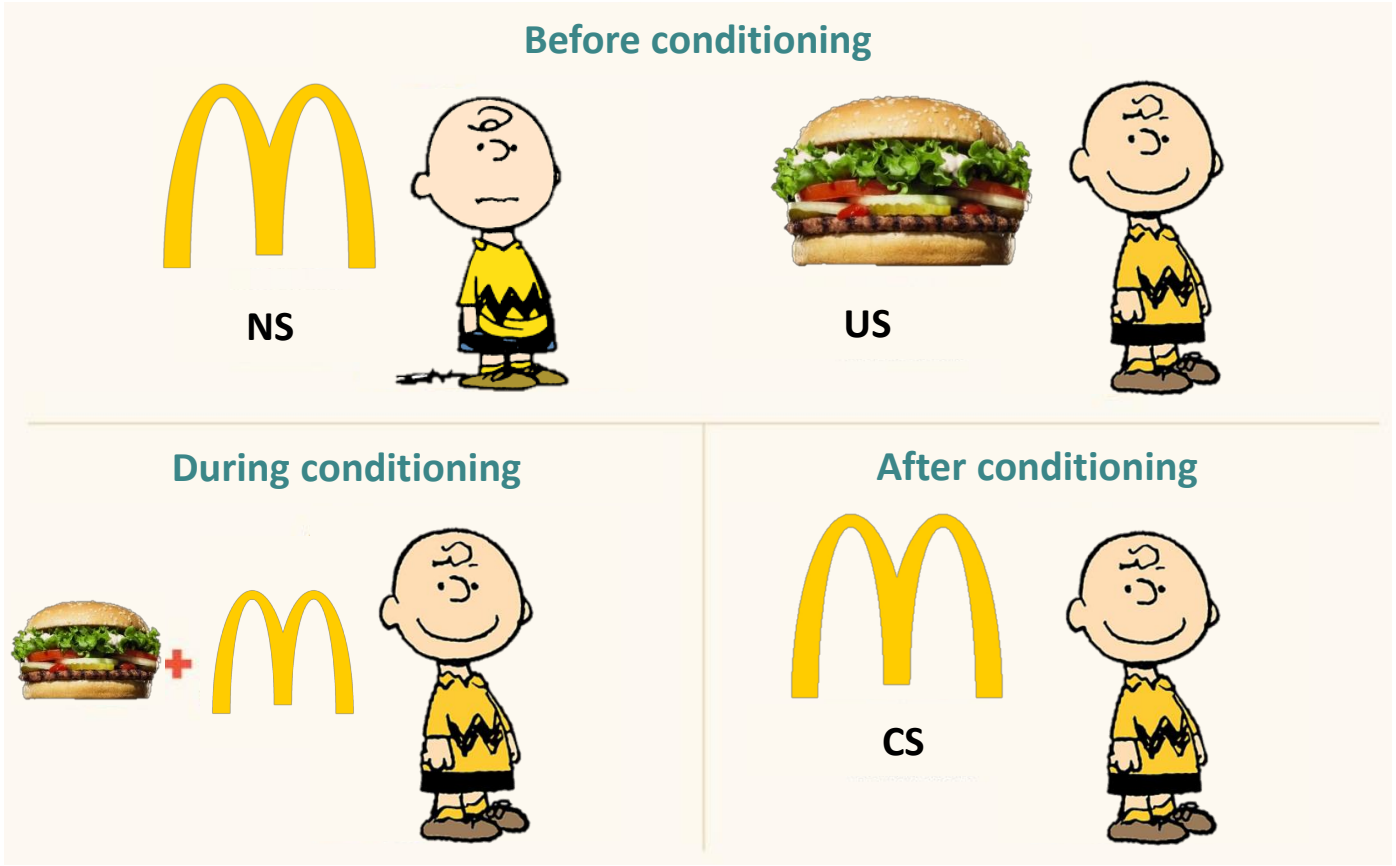
Discuss in pairs:

- Examples of Pavlovian learning
- Identify the
    - conditioned stimulus
    - unconditioned stimulus
    - conditioned response
    - unconditioned response



5 minutes

Pavlovian learning in everyday life



Intro to instrumental learning

Aka

Instrumental conditioning

Operant conditioning

https://youtu.be/5XUvm_smWHY

Instrumental learning system

- Control learning
- Learns action-outcome associations
- Learn to predict
    - when reinforces are likely to occur
    - which action bring about those reinforcers
- These predictions enable the animal to produce specific actions in **anticipation** of reinforcers, instead of responding exclusively in a reactive manner once reinforcers have occurred

Instrumental learning involves associating an action with an outcome

Described by Edgar Thorndike and systematically studied by B. F. Skinner and others.

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)

Skinner Box

The nature of the outcome shapes behavior

Positive reinforcement

- Delivery of rewarding outcome increases the probability of emitting the action

Positive punishment

- Delivery of aversive outcome decreases the probability of emitting the action

Negative reinforcement

- Omission of aversive outcome increases the probability of emitting the action

Negative punishment

- Omission of rewarding outcome decreases the probability of emitting the action

	Delivery	Omission
Appetitive	Positive reinforcement	Negative punishment
Aversive	Positive punishment	Negative reinforcement

	Delivery	Omission
Appetitive	Increases behavior	Decreases behavior
Aversive	Decreases behavior	Increases behavior

Instrumental conditioning in everyday life

Can you come up with some examples?

Discuss in pairs

5 mins

Then report some examples

Instrumental conditioning in everyday life

Can you come up with some examples?

Discuss in pairs

5 mins

Then report some examples

The frequency of the outcome also shapes behavior

Continuous schedule

- the desired behavior is followed by the outcome every single time it occurs
- most effective when trying to teach a new behavior

Partial schedule

- the desired behavior is followed by the outcome only part of the time it occurs
- Behaviors are acquired more slowly, but the response is more resistant to extinction

Four different partial schedules

1. **Fixed-ratio**

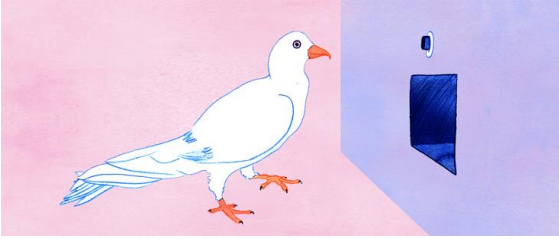
The outcome becomes available only after a **specified number of responses**.

This schedule produces a high, steady rate of responding with only a brief pause after the delivery of the outcome.

First learning ratio training, 2 lever presses for 1 food

Pressing the lever 10 times to get 1 reward

Four different partial schedules

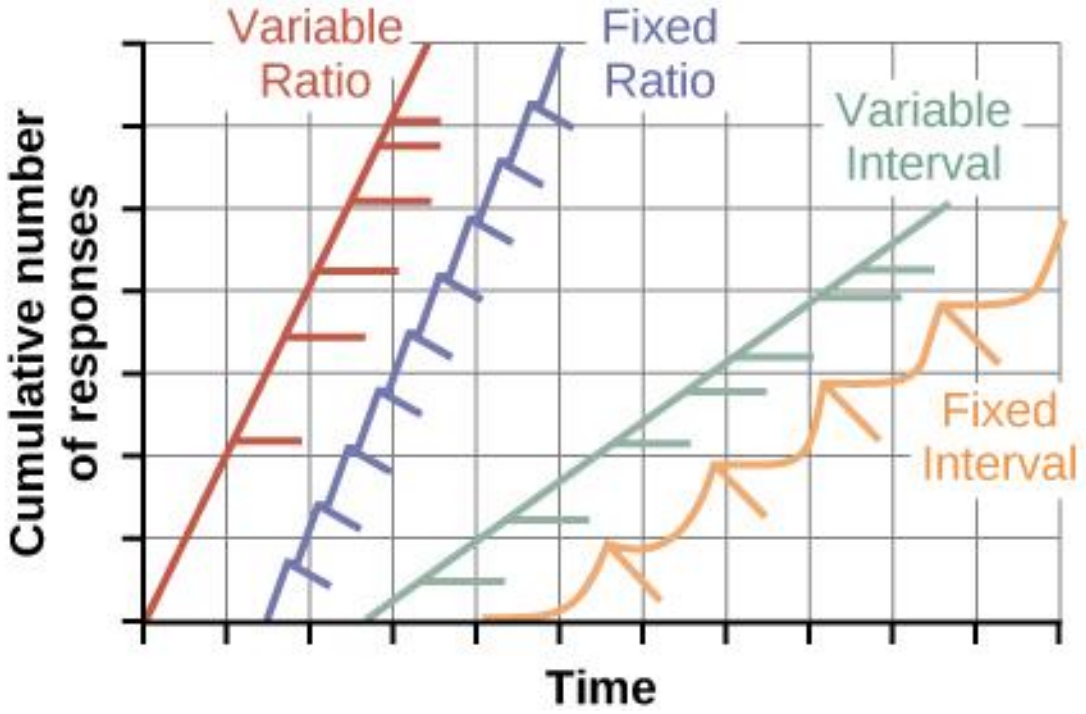


1. **Fixed-ratio**
- The outcome becomes available only after a **specified number of responses**.
- This schedule produces a high, steady rate of responding with only a brief pause after the delivery of the outcome.

2. **Variable-ratio**
- The outcome becomes available after an **unpredictable number of responses**.
- This schedule creates a high steady rate of responding.

3. **Fixed-interval**
- The outcome becomes available after a **specified interval of time**.
- This schedule causes high amounts of responding near the end of the interval but slower responding immediately after the delivery of the outcome.

4. **Variable-interval**
- The outcome becomes available after an **unpredictable interval of time.**
- This schedule produces a slow, steady rate of response.



Partial schedules of instrumental conditioning in everyday life

Can you come up with some examples?

Discuss in pairs

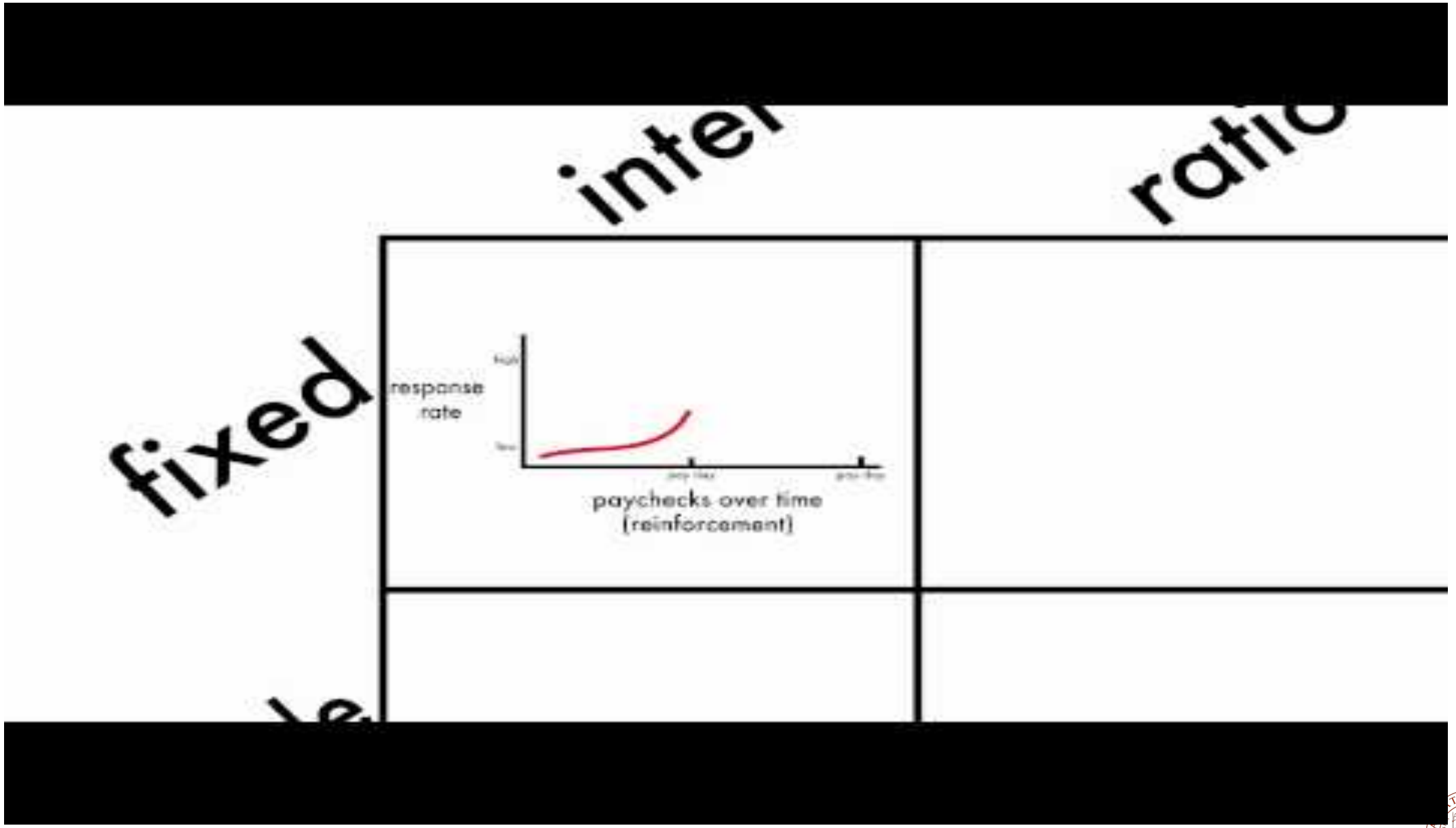
5 mins

Then report some examples

Partial reinforcement everyday life

1. Fixed-ratio
- Supermarket points
- Videogames

2. Variable ratio
- Gambling
- Lottery games



https://www.youtube.com/watch?v=GLx5yl0sxeM

https://youtu.be/qG2SwE_6uVM

wooclap

Recommended readings

- Daw, N. D., & O'Doherty, J. P. (2014). Multiple systems for value learning. In *Neuroeconomics* (Chapter 21, pp. 393-410). Academic Press.
- Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2014). Cognitive Neuroscience, The biology of the mind.
  - Page 393
- Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., & Mack, S. (Eds.). (2000). Principles of neural science. New York: McGraw-hill.
  - chapter 65 sections:
    - Implicit Memory Can Be Associative or Non-associative
    - Classical Conditioning Involves Associating Two Stimuli
    - Operant Conditioning Involves Associating a Specific Behavior with a Reinforcing Event