SIT796 Reinforcement Learning

The Psychology behind Reinforcement Learning

Presented by:
Dr. Thommen George Karimpanal
School of Information Technology

Deakin University CRICOS Provider Code: 00113B

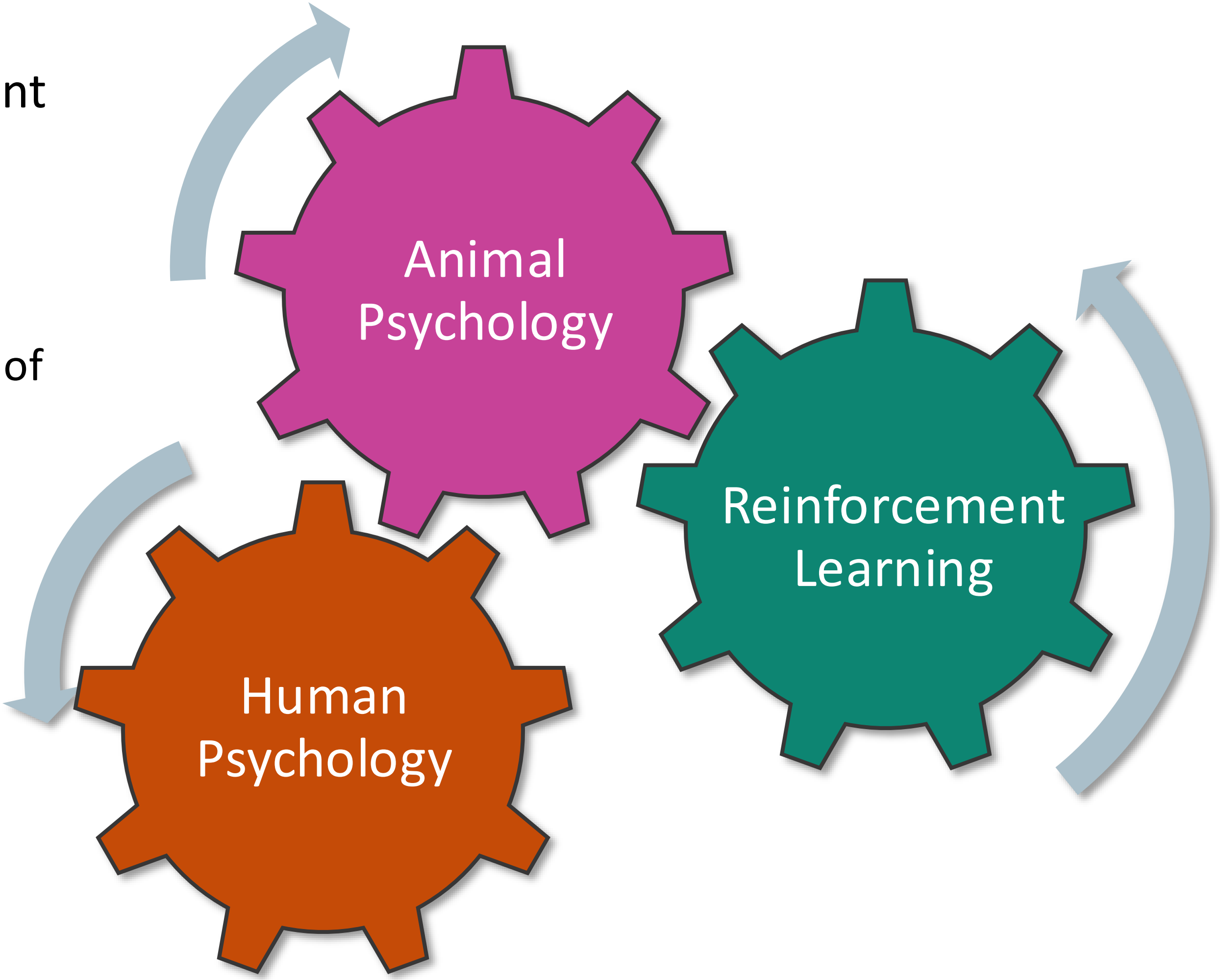
Motivation for Reinforcement Learning

Like many areas of Artificial Intelligence, Reinforcement Learning builds on our studies in other fields.

- Artificial Neural Networks based on Neurons.
- Evolutionary algorithms based on biological evolution
- Expert systems based on philosophical understanding of knowledge and logical inference
- …

Reinforcement Learning is founded on Animal Behaviour (Psychology)

- Classical Conditioning
- Instrumental Conditioning and the Law of effect
- Frequency of Reinforcement
- Delayed Reinforcement
- Habitual and Goal-directed behaviour



Pavlovian vs Operant Conditioning

Pavlovian (Classical)

- Occurs because of the subject’s instinctive responses
- A neutral stimulus gains the ability to elicit a response as a result of being paired with another stimulus

Operant (Instrumental)

- Contingent on the willful actions of the subject
- Occurs only after the organism executes a predesignated behavioral act.

Classical Conditioning

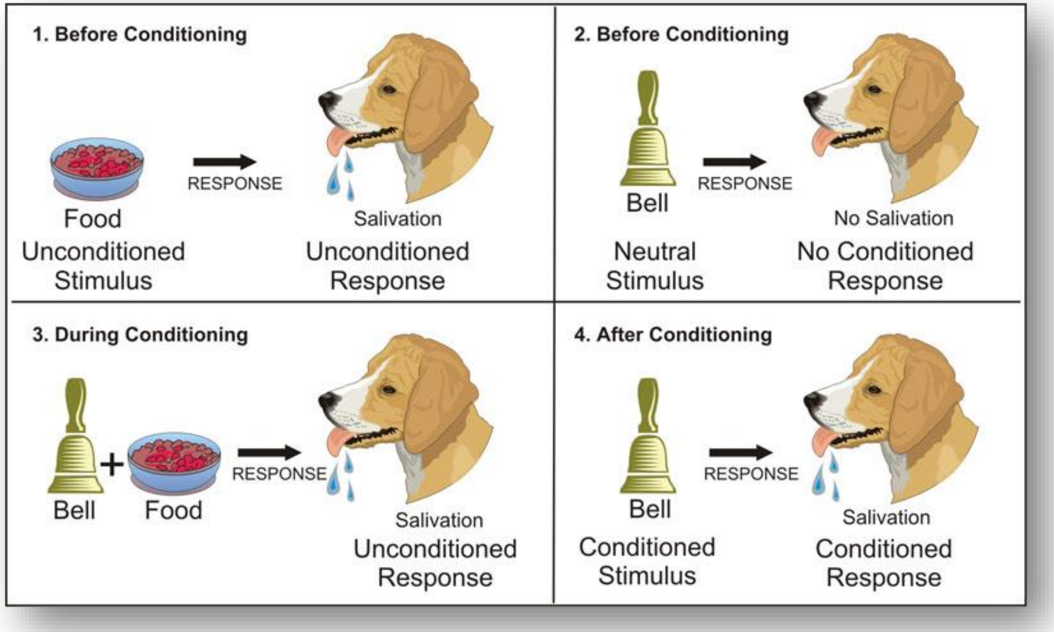
Based on work originally done by Ivan Pavlov (1927).

An ***Unconditioned Response*** (UR) is an inborn response and occurs automatically based on an ***unconditioned stimuli*** (US).

- A dog automatically salivating when food is placed in front of it.

A ***Conditioned Response*** (CR) is learnt. Takes a previously ***neutral stimuli*** and turns it into a ***conditioned stimuli*** (CS)

- Ringing a bell by itself is a neutral stimuli, producing no salivation in response.
- Each time food is place in front of the dog a bell is rung
- After a period of time ring the bell becomes a conditioned response and the dog will salivate when the bell is rung regardless of whether food is provided.



https://edu.glogster.com/glog/ivan-pavlov/224aae54ua8?=glogpedia-source

The US is called a ***reinforcer*** because it reinforces the production of a CR when stimuli.

Instrumental Conditioning

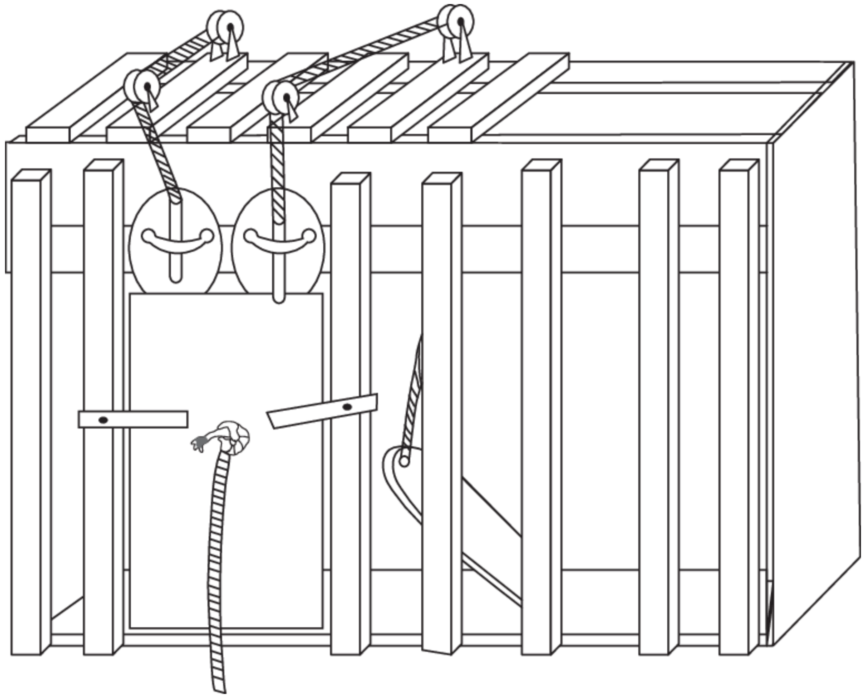
In ***Instrumental Conditioning***, or ***Operant conditioning*** as named by Skinner (1938), a stimulus is delivered contingent on an animal's behaviour.

- Whereas, a in Classical conditioning the US was provided regardless of behaviour.

Thorndike's experiments (1898)

- That a cat placed in a 'puzzle box' may take around 300 seconds to accidently activate three switches that open the door and allow access to visible food.
- After successive experiences it got this down to 6 or 7 seconds

Lead him to the development of the ***Law of effect***, describing the idea of ***trial-and-error***.



https://www.researchgate.net/figure/One-of-the-puzzle-boxes-used-by-Thorndike-to-study-the-acquisition-of-new-behaviors-in_fig9_285777770

Sutton and Barto

SIT796 Reinforcement Learning

Classical Conditioning

Presented by:
Dr. Thommen George Karimpanal
School of Information Technology

Classical Conditioning as Prediction

Common types of classical conditioning experiments

Delayed Conditioning

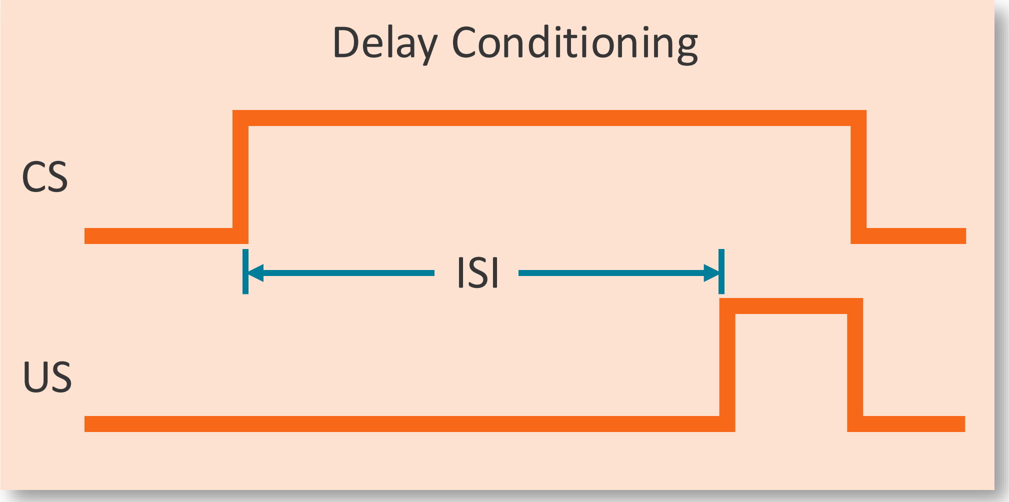
- Applies the ***conditioned stimuli*** (CS) throughout the ***Interstimulus interval*** (ISI).
- The ***unconditioned stimuli*** (US) is only applied at the end.

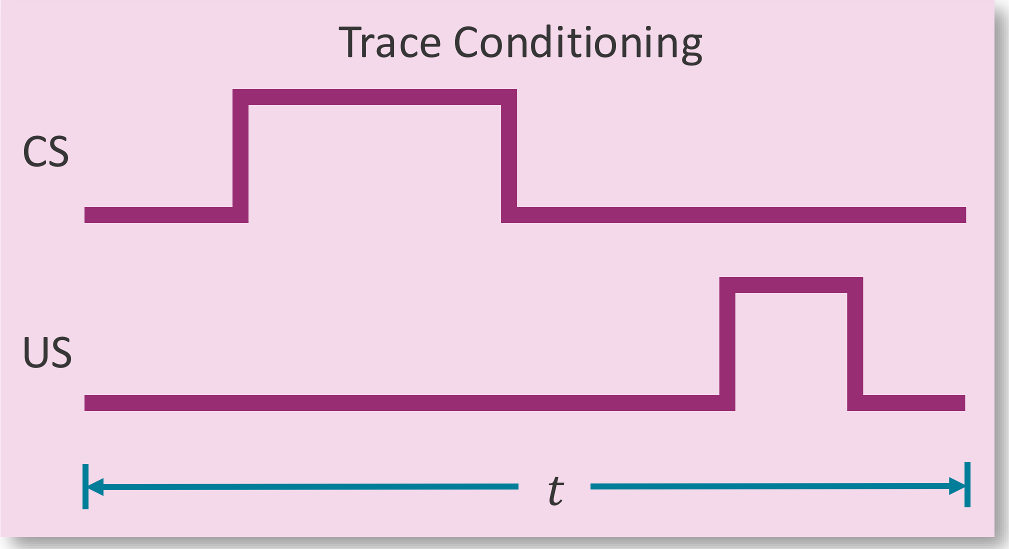
Trace Conditioning

- The US is only applied after the CS has completed.

Results have illustrated that applying these approaches over a series of trials causes the animal to learn to ***predict*** the US

Hence, ***Predictive Reinforcement Learning*** is often regarded as an algorithmic form of ***Classical Conditioning***.

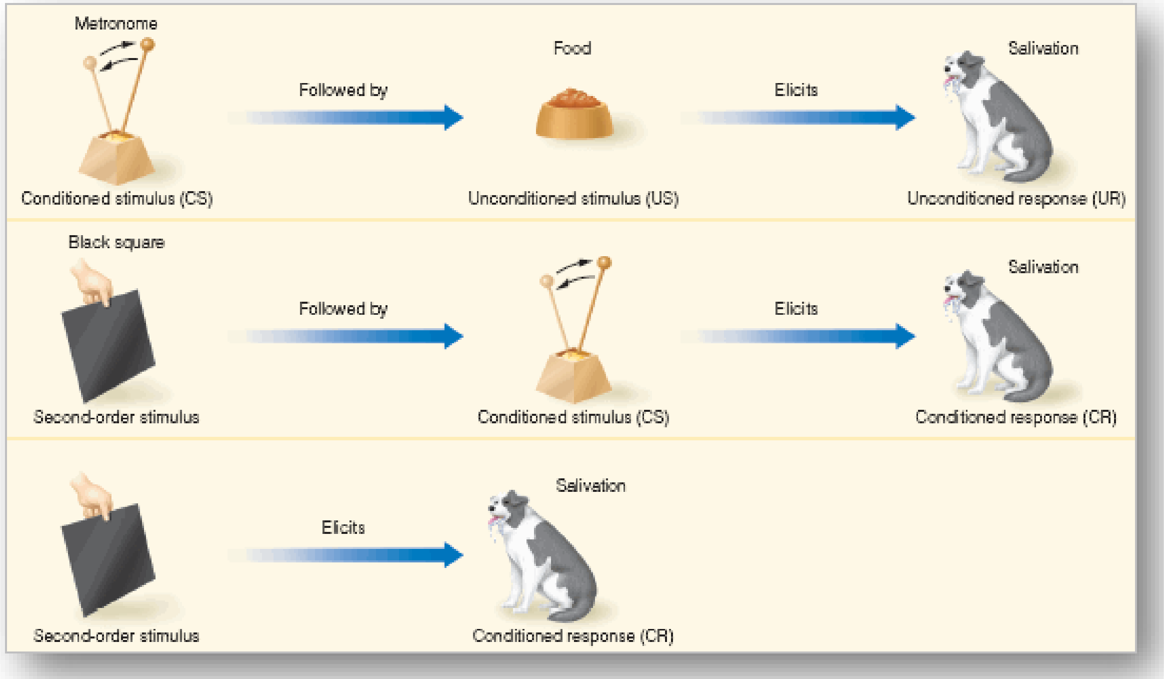




Classical Conditioning (*Higher-order conditioning*)

Higher-order conditioning

- Higher order conditioning occurs when an animal is presented with a CS followed by a previously learn CS.
- The animal learns to produce the same CR to the second CS.
  - This occurs even if the second CS was not followed by the original US.
- This represents what is called ***second-order conditioning.***
- Can potentially be extended to higher orders


Metronome
Followed by
Food
Elicits
Salivation
Conditioned stimulus (CS)
Unconditioned stimulus (US)
Unconditioned response (UR)
Black square
Followed by
Elicits
Salivation
Second-order stimulus
Conditioned stimulus (CS)
Conditioned response (CR)
Elicits
Salivation
Second-order stimulus
Conditioned response (CR)

https://www.chegg.com/flashcards/pscyh-150-final-78c1de23-f8f6-41a1-88a7-dde0652b3b99/deck

Classical Conditioning (Other Findings)

Stimulus Generalization

- Animals can associate new stimuli, which are similar to already CS, to produce the same or similar CR
- For instance, having survived a traumatic experience with a snake a cat will learn to avoid all snakes.

Extinction

- If a learnt CS is presented many times without the subsequent US resulting will eventually cause the animal to forget the conditioning.
- For instance if I stop opening the door when the cat knocks it will eventually stop knocking.

Spontaneous recovery

- After extinction, a CS can be recovered by reintroducing the CR.
- Hence, if I go a week not opening the door to the cat but then forget and open the door it will recover the CS.

Conditioned Inhibition

- The inverse can also be learnt – where an animal learns that CS signals the absence of US.
- For example, my cat has learnt that if it hides in the linen press when there are young visitor at the house then it can avoid being harassed.

SIT796 Reinforcement Learning

Operant Conditioning

Presented by:
Dr. Thommen George Karimpanal
School of Information Technology

Deakin University CRICOS Provider Code: 00113B

Instrumental Conditioning (Law of Effect)

Law of effect

> *Behaviours followed by favourable consequences become more likely, and behaviours followed by unfavourable consequences become less likely.*

Found that learning goes beyond the process of simply finding behaviours that result in a high reward

Found it also includes the process of *connecting* those behaviours to the situations where those actions were taken

Called this learning by “selecting and connecting”

Instrumental Conditioning

Biological and computation models of ***evolution*** are based on "*selection*".

- There is no associative component

While ***supervised learning*** is only "*associative*". Methods remember generalisation of associations between inputs and outputs based on instruction.

- There is no selection process

Reinforcement Learning is based on both the ***search*** and ***memory*** processes that are fundamental to both the Psychology and the computational models used to implement Reinforcement Learning agents.

- IC's focus on behaviours taken to reach a reward is often regarded as the basis of Reinforcement Learning for ***Control***

Reinforcement Learning agents must:

- Search for possible solutions through ***trial-and-error*** (eg applying the Law-of-effect)
- Associate situations (states) to the outcomes observed.

Instrumental Conditioning

Skinner (1938) introduced the new term into the Law of Effect - ***Reinforcement***

- Focusing on the idea of a ***reinforcement stimulus*** guiding behaviour.

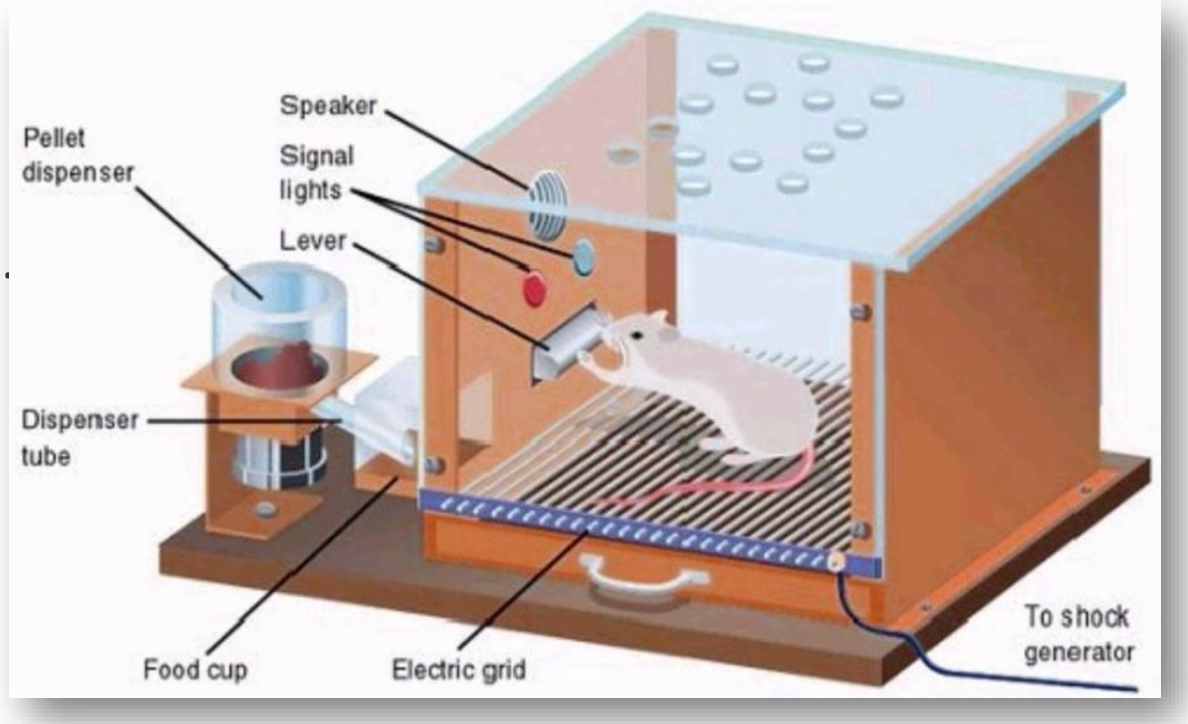
Skinner developed a operant conditioning chamber – now called a "Skinner Box", found three responses

- Positive Reinforcement
- Negative Reinforcement
- Punishment

Found how these responses can be used to train animals.

Ideas stemming from this idea

- Human psychology in early and some later learning
- Learning and teaching theories
- Changes in child rearing practices
- And of course Reinforcement learning
- …



https://www.simplypsychology.org/operant-conditioning.html

Instrumental Conditioning (Delayed Reinforcement)

Clark Hull (1943) found secondary reinforcement was possible.

- Like higher-order conditioning an animal can learn even when there is a significant time interval between action and the resulting reinforcement stimulus.

The law of effect requires a backward effect on connections.

- Relates to what Minsky (1961) referred to as the credit assignment problem
  - How do you distribute credit for success amongst many past decisions?

Trace conditioning

- Pavlov (1927) suggested that a stimulus must leave a trace in the nervous system
- Hull (1943) also suggested an animal had a goal gradient in instrumental conditioning
- Suggested that this trace reduces exponentially over time

Reinforcement Learning utilises ***eligibility traces*** and a ***value function*** to enable learning with delayed reward

Instrumental Conditioning (Delayed Reinforcement)

Clark Hull (1943) found secondary reinforcement was possible.

- Like higher-order conditioning an animal can learn even when there is a significant time interval between action and the resulting reinforcement stimulus.

The law of effect requires a backward effect on connections.

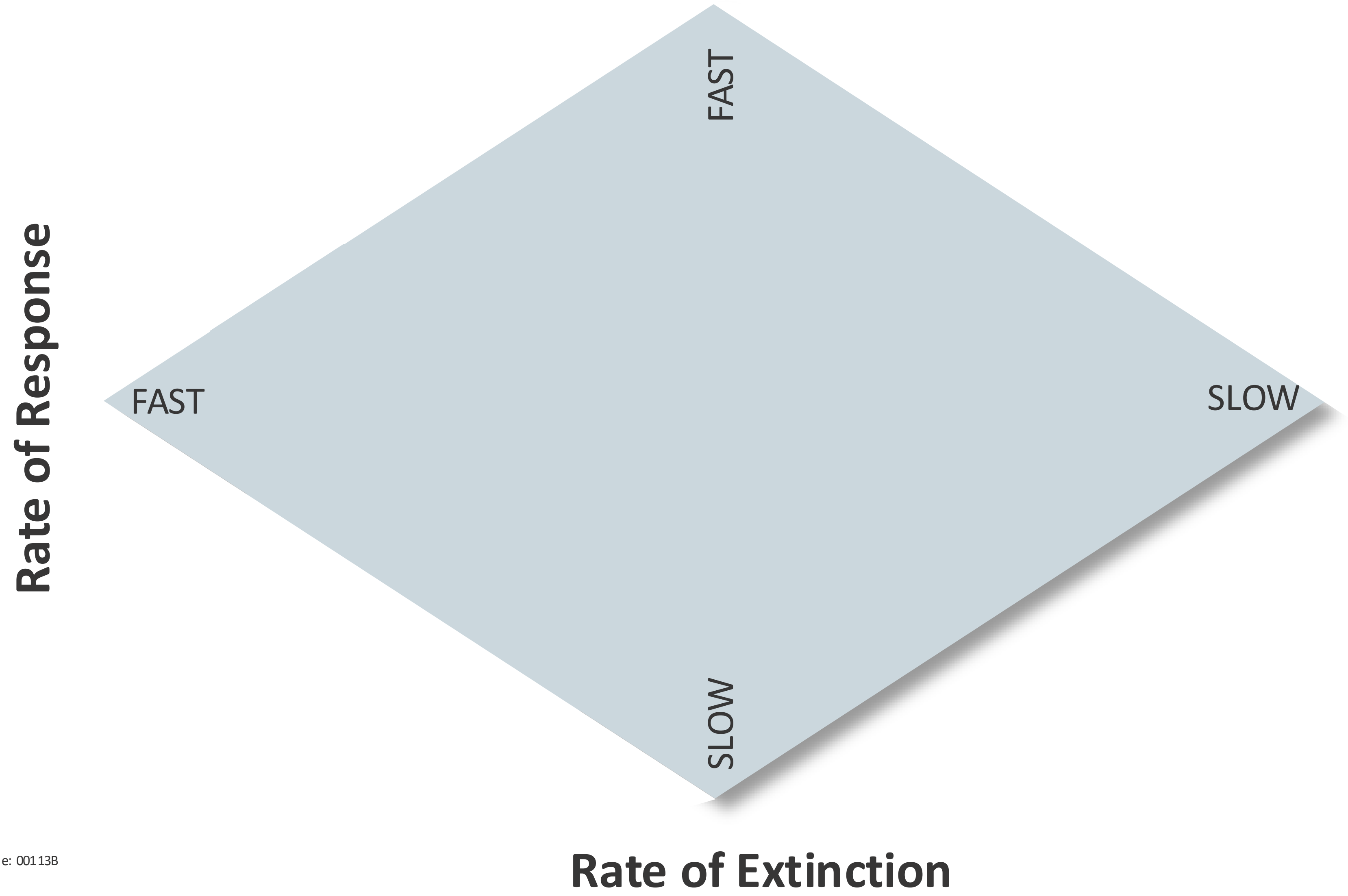
- Relates to what Minsky (1961) referred to as the credit assignment problem
  - How do you distribute credit for success amongst many past decisions?

Trace conditioning

- Pavlov (1927) suggested that a stimulus must leave a trace in the nervous system
- Hull (1943) also suggested an animal had a goal gradient in instrumental conditioning
- Suggested that this trace reduces exponentially over time

Reinforcement Learning utilises ***eligibility traces*** and a ***value function*** to enable learning with delayed reward

Instrumental Conditioning (Frequency of Reinforcement)



SIT796 Reinforcement Learning

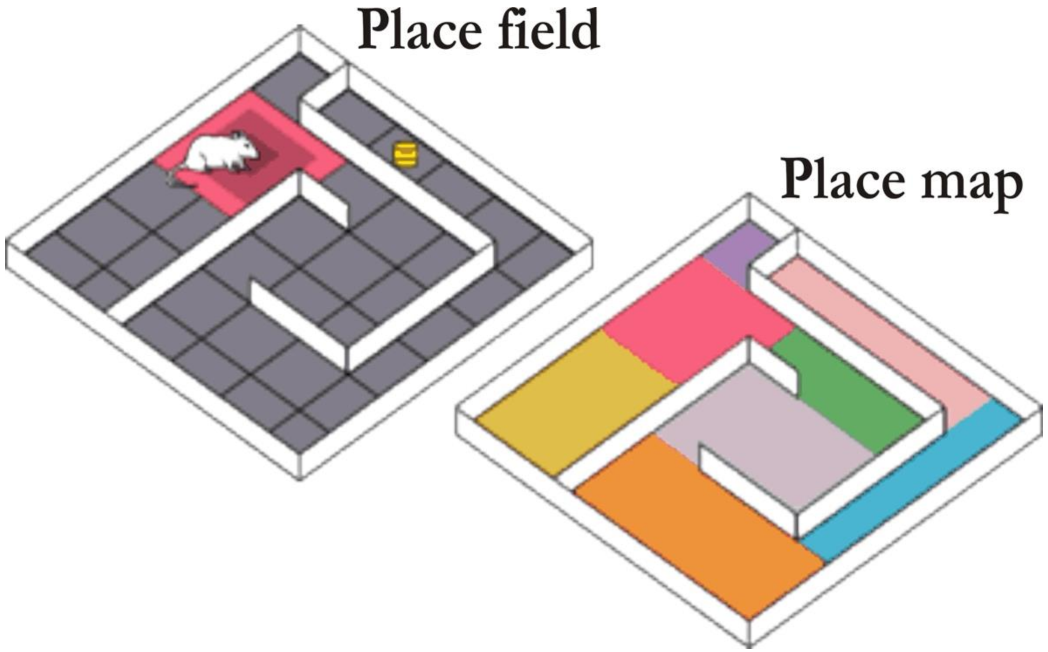
Cognitive Maps and Latent Learning

Presented by:
Dr. Thommen George Karimpanal
School of Information Technology

Cognitive Maps as Environmental Representations

The concept was introduced by Edward Tolman (1948)

- These are mental representations that increase recall and learning of information
- They are presumed to be learned by gradually acquiring elements of the world
- As cognitive, they are often presumed to differ from "true" maps of the environment.
- Cognitive maps are not restricted to spatial layouts, but can apply more generally to model an animals task space (Wilson, Takahashi, Schoenbaum, and Niv, 2014)



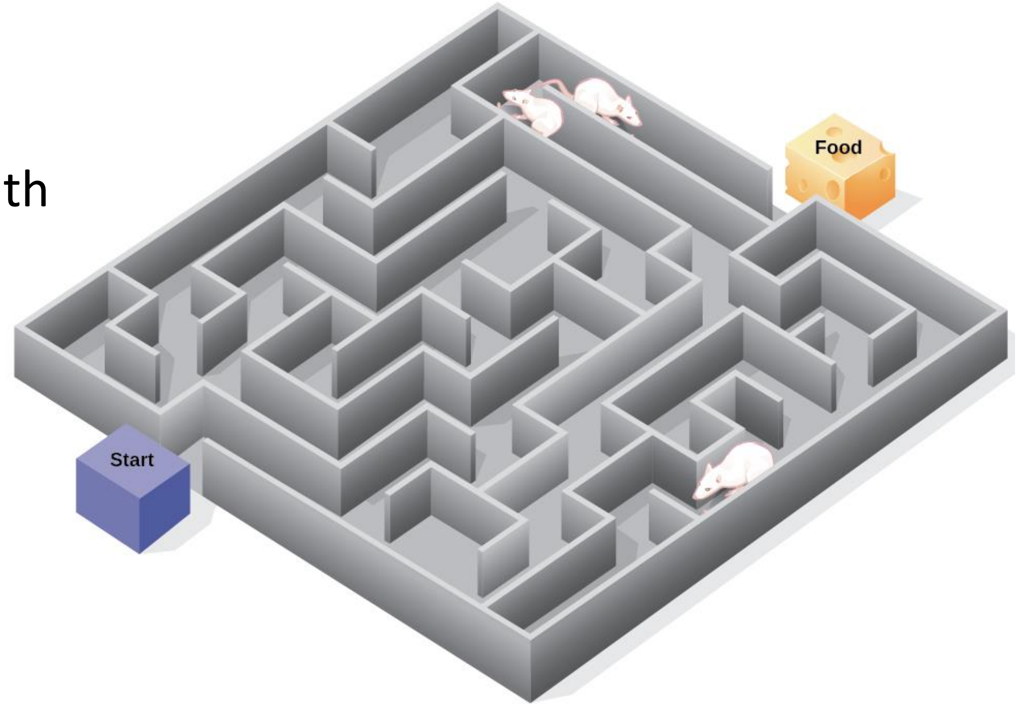
Latent Learning of Cognitive Maps

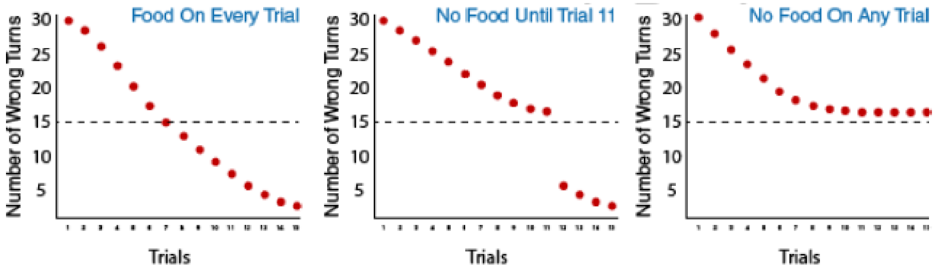
Latent Learning considers whether an animal learns a ***cognitive map*** of an environment even if there is no reason to do so.

Edward Tolman (1948) used a maze to show that this learning with 3 groups of rats

- Group 1 group always has food at the end
- Group 2 has no food at the end until trial 12 when food is now placed at the end
- Group 3 never has food at the end.

While Instrumental Conditioning is the basis of ***Value-based Reinforcement Learning***, Cognitive maps are analogous to ***Model-based Reinforcement Learning***





https://courses.lumenlearning.com/wmopen-psychology/chapter/psychology-in-real-life-latent-learning/

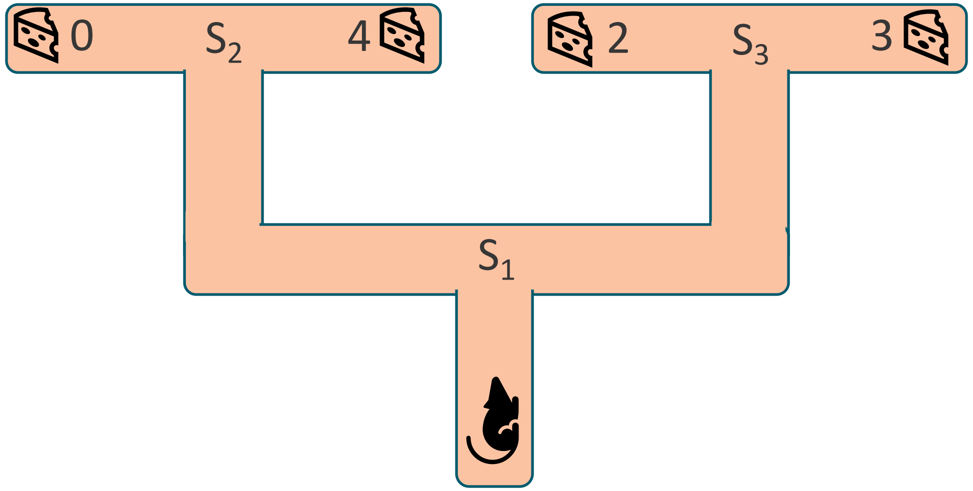
Habitual and Goal-directed Behaviour

Instrumental Conditioning (Value-based RL) corresponds to ***Habitual*** behaviour.

- Habitual behaviour is fast, automatic and reactionary
- In RL Value-Based approaches are also referred to as ***Model-Free***

Cognitive maps (model-based RL) is considered to be a ***Goal-directed*** control.

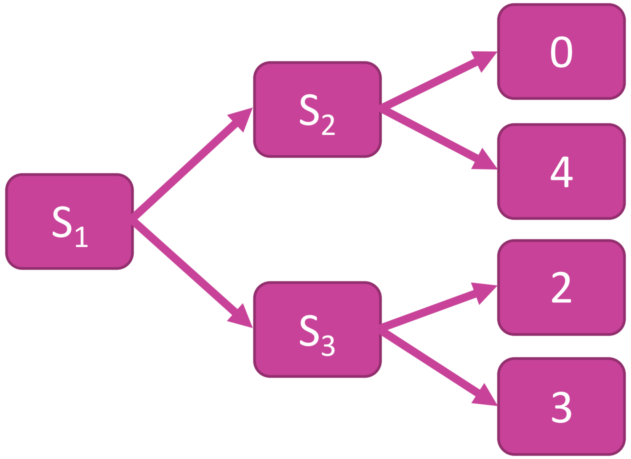
- Goal-directed behaviour is purposeful/intentional and uses knowledge of the environments



Problem

State/Action	Q-Value
S_1/Left	4
S_1/Right	3
S_2/Left	0
S_2/Right	4
S_3/Left	2
S_3/Right	3

Model-Free



Model-Based

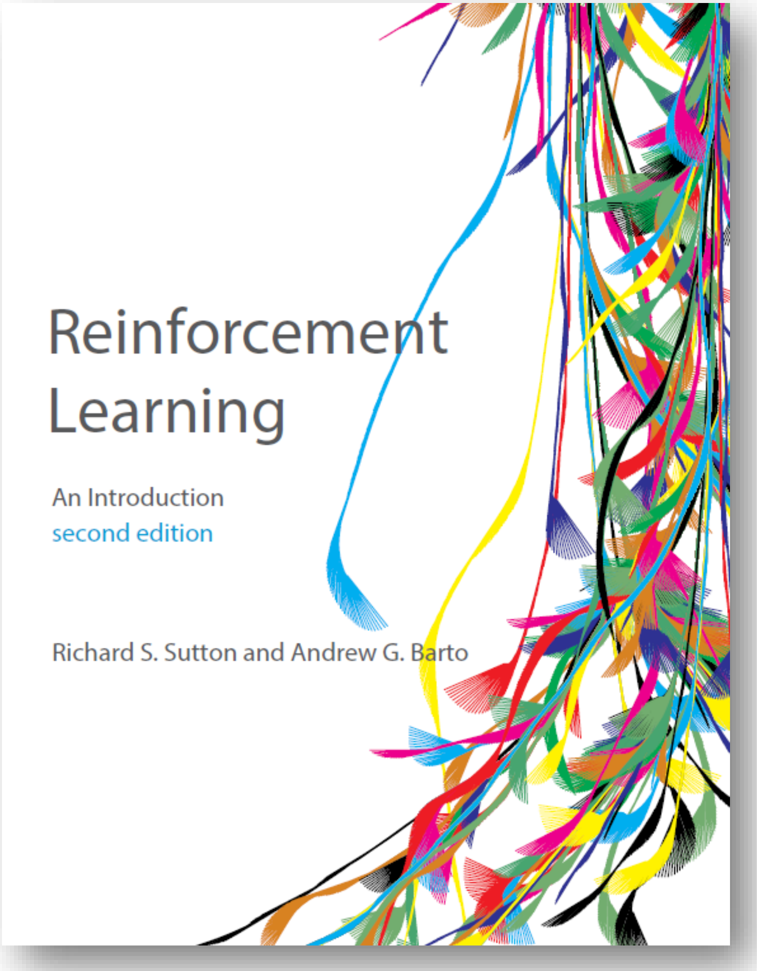
Conclusion

This was a quick overview of Psychology 101.

- Intention was to link the ideas and terminology in Psychology to those used in Reinforcement Learning
- Inspire an intuitive understanding of RL in real world terms.

For more detailed information see Sutton and Barto (2018) Reinforcement Learning: An Introduction

- Chapter 14
- http://incompleteideas.net/book/RLbook2020.pdf



Reinforcement Learning

An Introduction
second edition

Richard S. Sutton and Andrew G. Barto