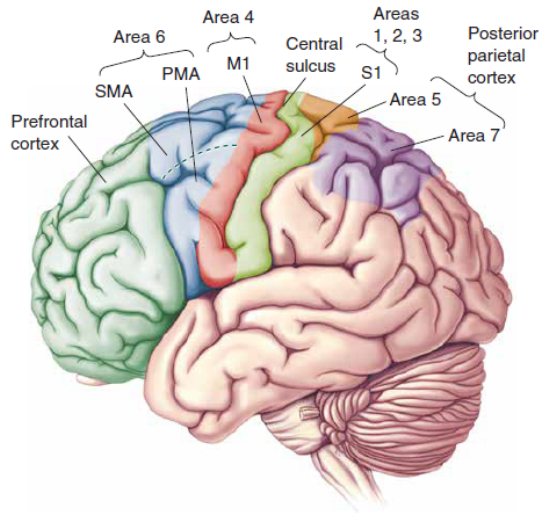
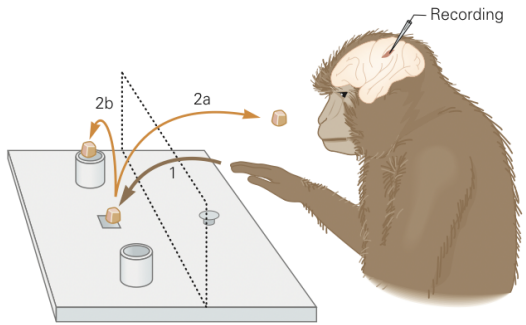
Neuromechanics of Human Motion

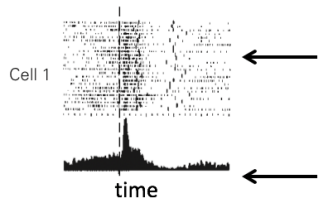
Reinforcement Learning

Joshua Cashaback, PhD

Recap — Brain Areas







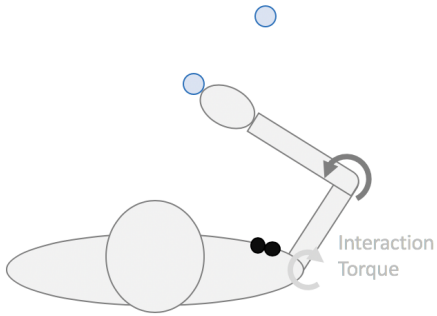
- Each line represents a single trial
- Black dot indicates neuron firing.

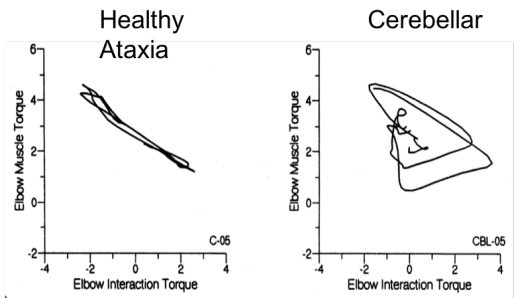
Histogram (vertical sum of all trials)

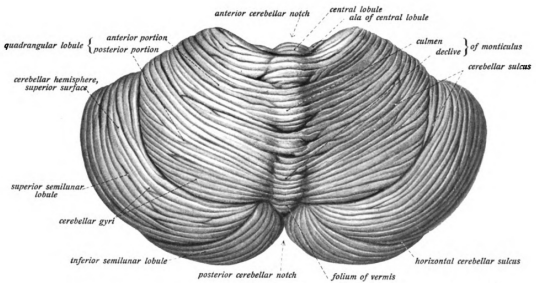
Recap — Internal Models and the Cerebellum

Internal Model

A stored or learned representation—via some pattern of neuronal activity—that accounts for environmental or internal dynamics







Lecture Objectives
Reinforcement Learning

1. Beginnings
   . Understand the Origins of Reinforcement Learning
2. Breadth
   . Knowledge on the scope of Reinforcement Learning
   . Lessons from Machine Learning
3. Behaviour
   . Neural circuits
   . Normal and diseased behaviour

BEGINNINGS

Beginnings

PSYCHOLOGY

Classical Conditioning

A learning process that occurs when two stimuli are repeatedly paired: a response which is at first elicited by the second stimulus is eventually elicited by the first stimulus alone.

Beginnings

PSYCHOLOGY

Classical Conditioning

A learning process that occurs when two stimuli are repeatedly paired: a response which is at first elicited by the second stimulus is eventually elicited by the first stimulus alone.

Pavlov's Dog

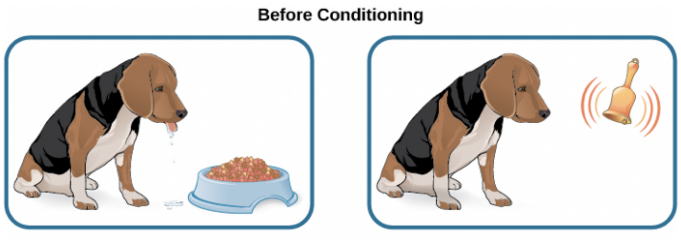
Beginnings

PSYCHOLOGY

Classical Conditioning

A learning process that occurs when two stimuli are repeatedly paired: a response which is at first elicited by the second stimulus is eventually elicited by the first stimulus alone.

Pavlov's Dog



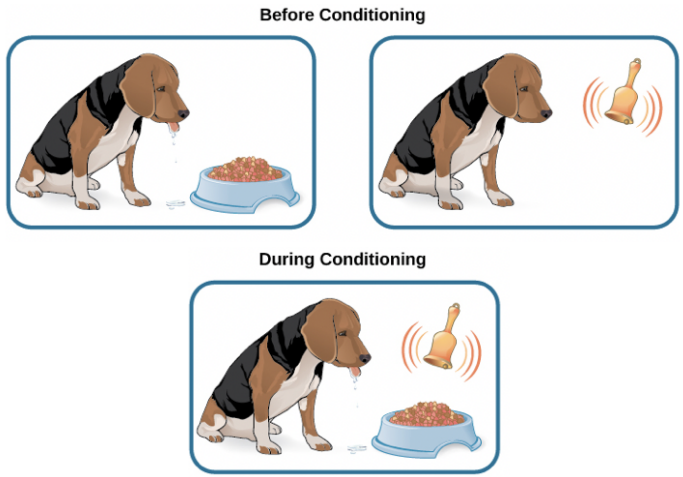
Beginnings

PSYCHOLOGY

Classical Conditioning

A learning process that occurs when two stimuli are repeatedly paired: a response which is at first elicited by the second stimulus is eventually elicited by the first stimulus alone.

Pavlov's Dog



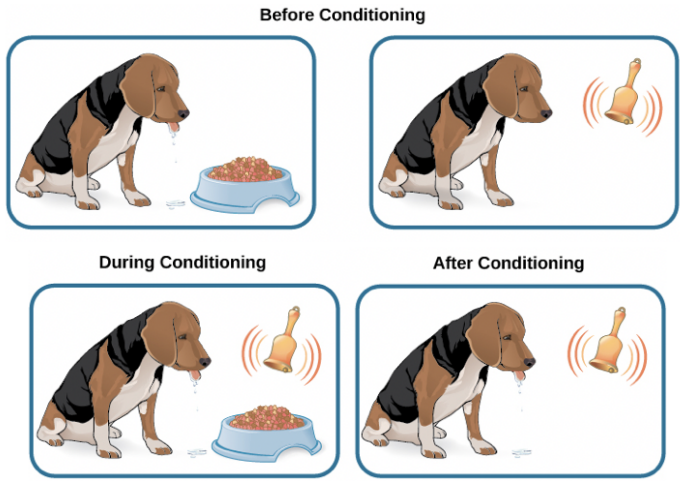
Beginnings

PSYCHOLOGY

Classical Conditioning

A learning process that occurs when two stimuli are repeatedly paired: a response which is at first elicited by the second stimulus is eventually elicited by the first stimulus alone.

Pavlov's Dog



Reinforcement and Punishment

1. Reinforcement: increase behaviour
2. Punishment: decrease behaviour
3. Positive: add something
4. Negative: take away something

Reinforcement and Punishment

1. Reinforcement: increase behaviour
2. Punishment: decrease behaviour
3. Positive: add something
4. Negative: take away something

Classroom Examples

1. Positive Reinforcement:
2. Negative Reinforcement:
3. Positive Punishment:
4. Negative Punishment:

Reinforcement and Punishment

1. Reinforcement: increase behaviour
2. Punishment: decrease behaviour
3. Positive: add something
4. Negative: take away something

Classroom Examples

1. Positive Reinforcement: Candy
2. Negative Reinforcement:
3. Positive Punishment:
4. Negative Punishment:

Reinforcement and Punishment

1. Reinforcement: increase behaviour
2. Punishment: decrease behaviour
3. Positive: add something
4. Negative: take away something

Classroom Examples

1. Positive Reinforcement: Candy
2. Negative Reinforcement: Take Away Homework
3. Positive Punishment:
4. Negative Punishment:

Reinforcement and Punishment

1. Reinforcement: increase behaviour
2. Punishment: decrease behaviour
3. Positive: add something
4. Negative: take away something

Classroom Examples

1. Positive Reinforcement: Candy
2. Negative Reinforcement: Take Away Homework
3. Positive Punishment: Writing Lines
4. Negative Punishment:

Reinforcement and Punishment

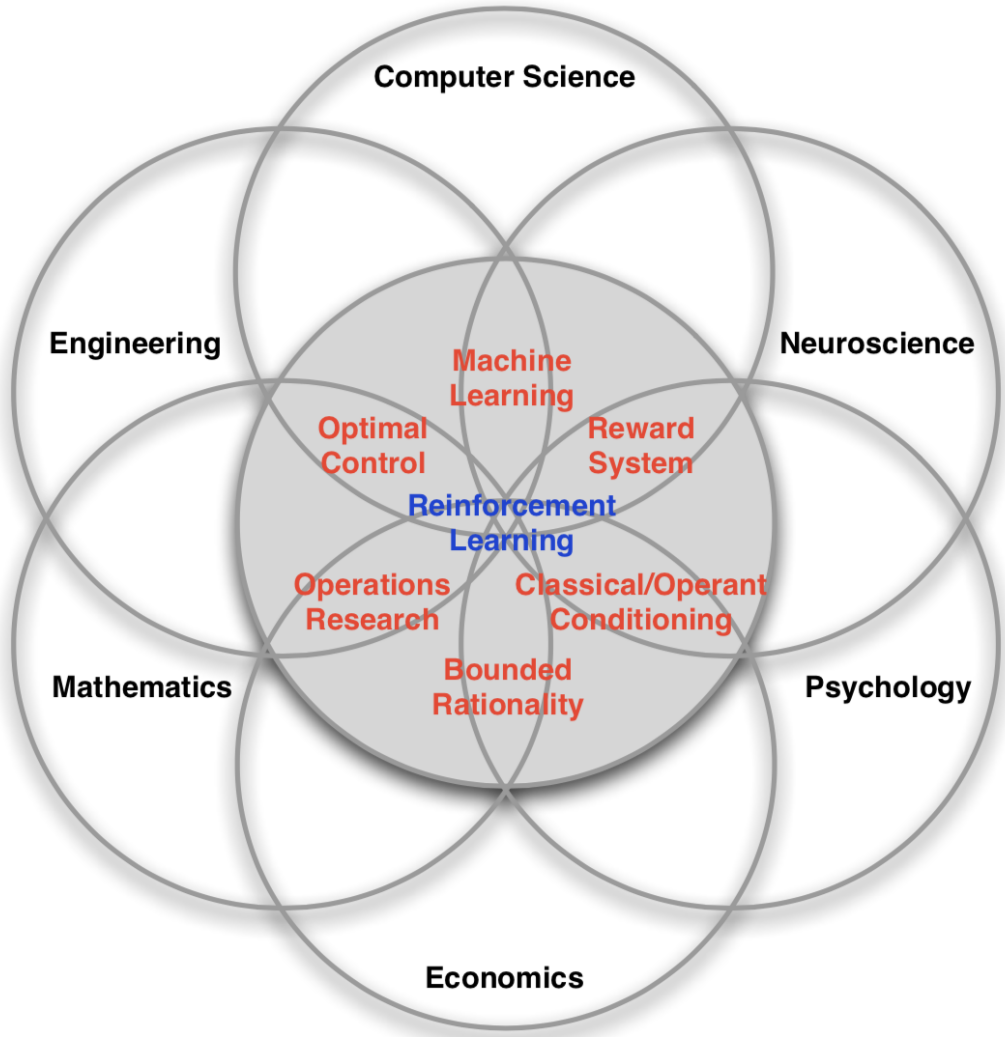
1. Reinforcement: increase behaviour
2. Punishment: decrease behaviour
3. Positive: add something
4. Negative: take away something

Classroom Examples

1. Positive Reinforcement: Candy
2. Negative Reinforcement: Take Away Homework
3. Positive Punishment: Writing Lines
4. Negative Punishment: Take Away Recess

BREADTH

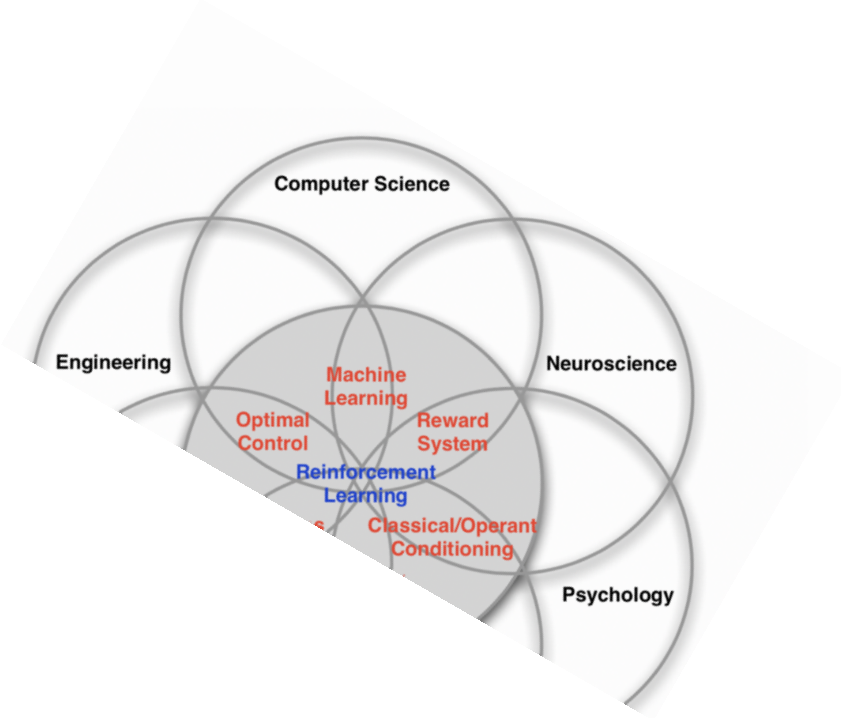
Breadth



Breadth

Neuroengineering

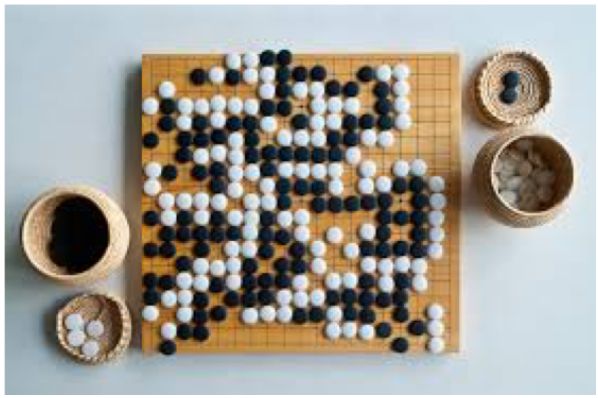
1. Classical Conditioning
2. Machine Learning / Optimal Control
   . principles of reinforcement learning
3. Reward System
   . Human (Biological) Behaviour
   . Neural Circuitry
   . Disease
4. REWARD PREDICTION ERROR

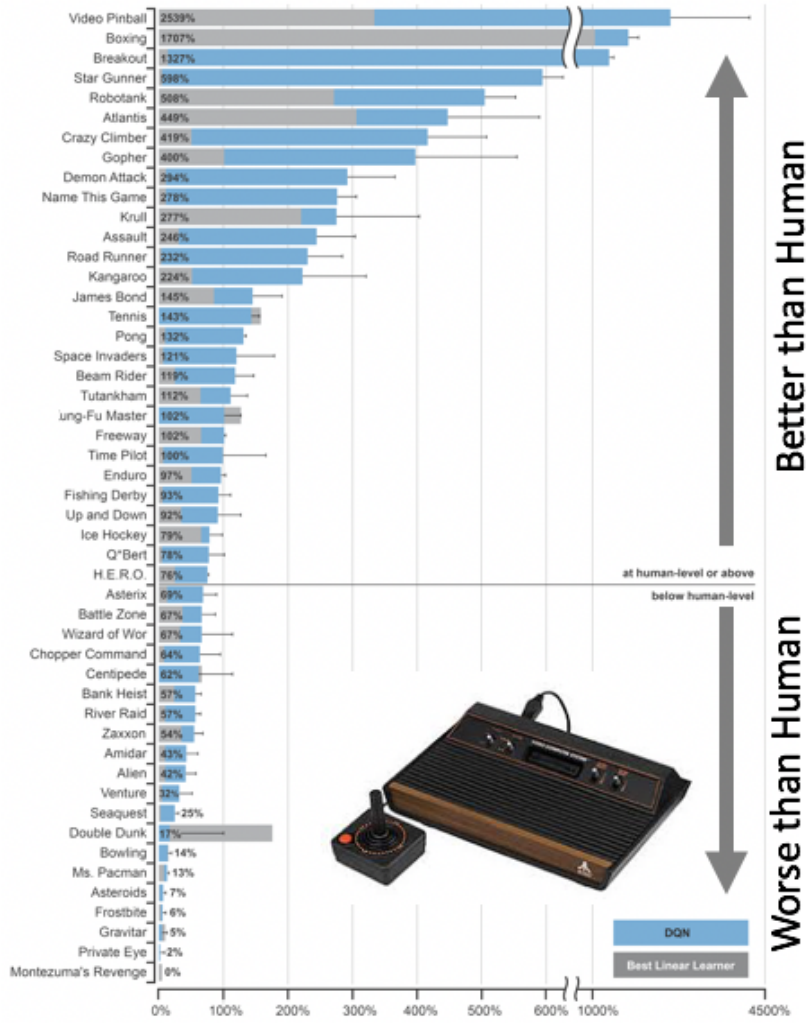


Machine (Reinforcement) Learning

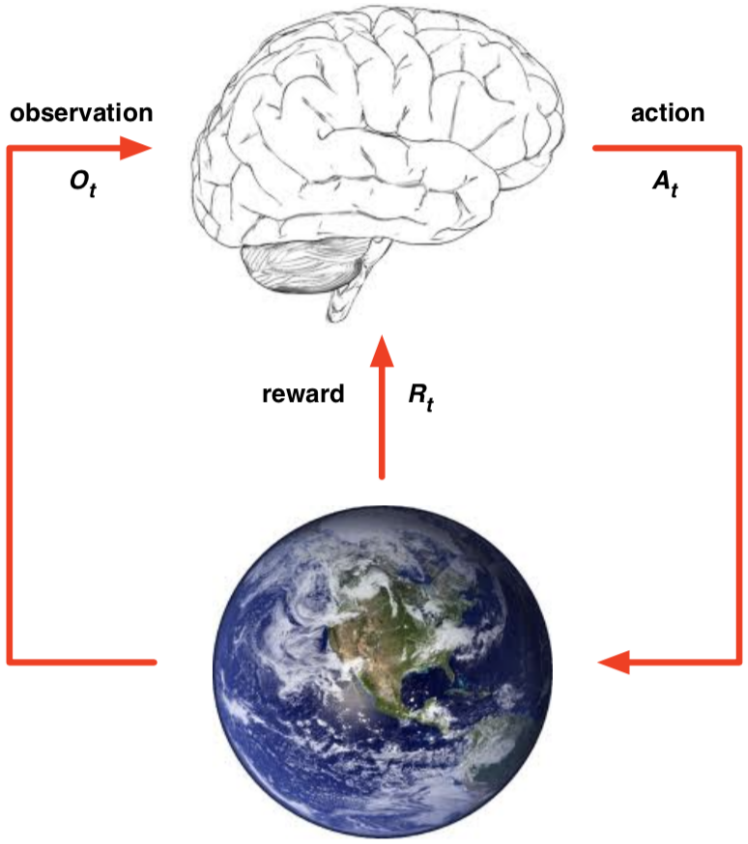
Reinforcement Algorithms

a. Extremely Powerful
b. Google DeepMind
    i. Atari
    ii. Go (poss. $= 10 * *\hat{1}0 * *48$)
    iii. Chess (4 hrs training, only taught how pieces move - Magnus Carlson; 2800 elo)
    iv.



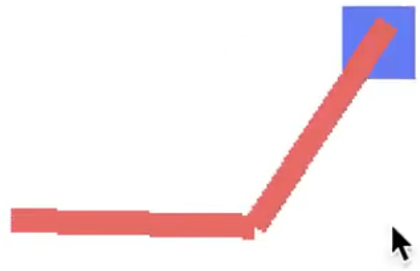


State, Action, Reward



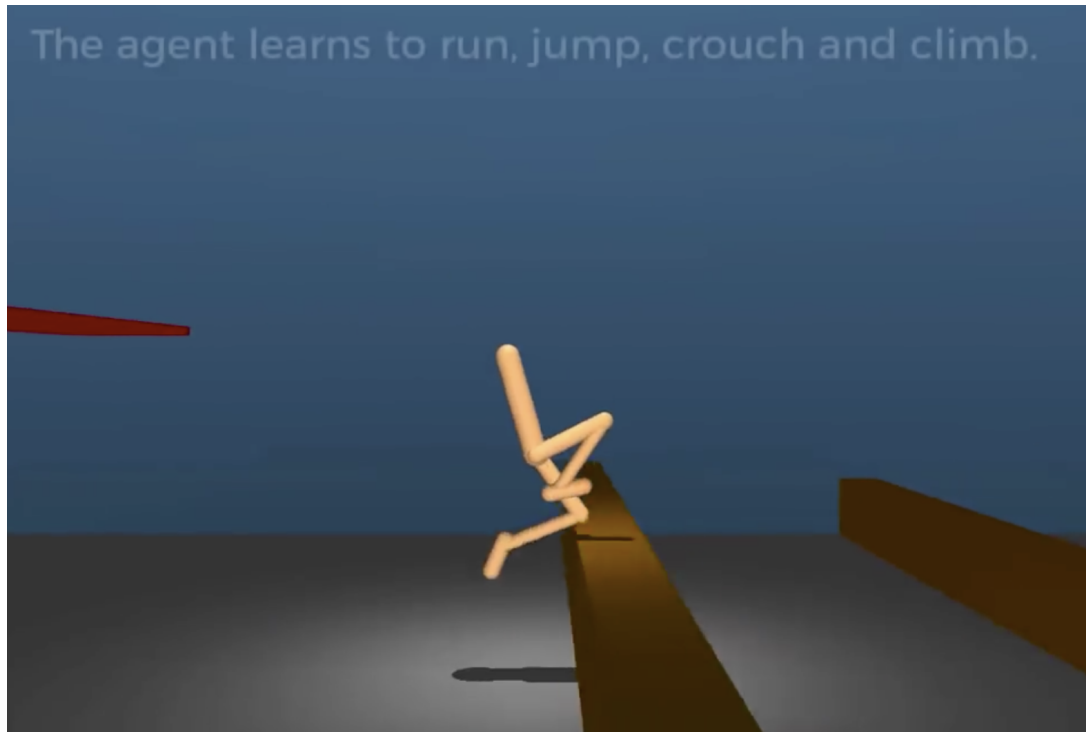
1. At each step (t) the agent:
   a. Executes action A_t
   b. Receives scalar reward R_t
   c. Receives observation O_t
2. The Environment:
   a. Receives action A_t
   b. Emits scalar reward R_t
   c. Emits observation O_t

NeuroEngineering: Optimal Control & Human(-like) Movement



Click Me: ▸ Link

NeuroEngineering: Optimal Control & Human(-like) Movement



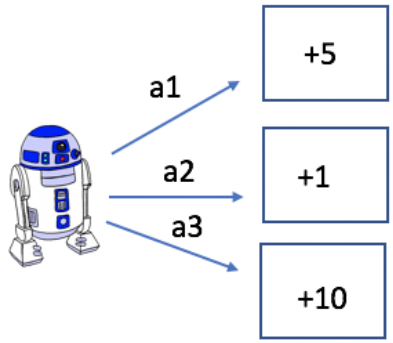
Click Me: ▸ Link

Goal of Reinforcement Learning

Maximize Reward (minimize loss)

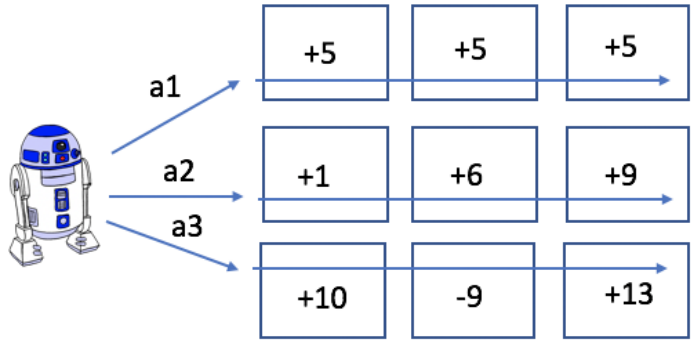
Goal of Reinforcement Learning

Maximize Reward (minimize loss)



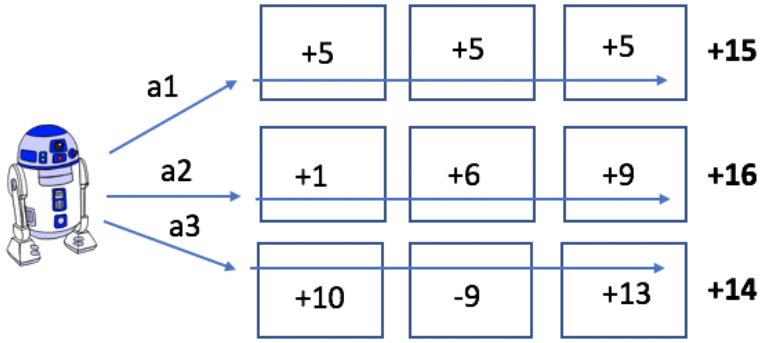
Goal of Reinforcement Learning

Maximize Reward (minimize loss)



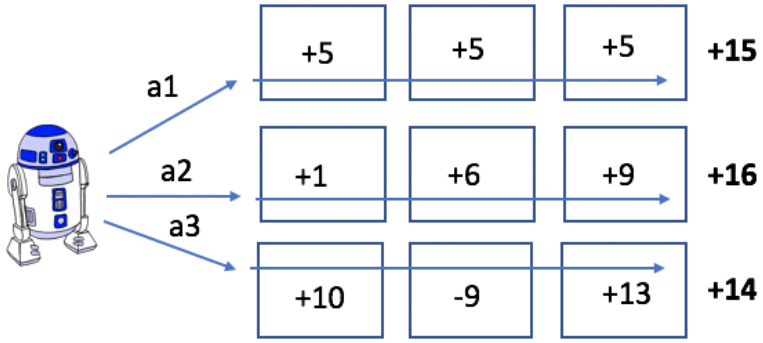
Goal of Reinforcement Learning

Maximize Reward (minimize loss)



Goal of Reinforcement Learning

Maximize Reward (minimize loss)

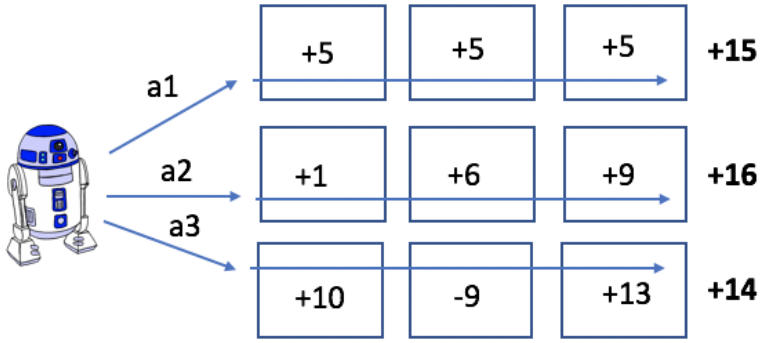


Cumulative Reward

1. Might be better to sacrifice immediate reward for long-term reward

Goal of Reinforcement Learning

Maximize Reward (minimize loss)



Cumulative Reward

1. Might be better to sacrifice immediate reward for long-term reward
    a. Investments
    b. Chess
    c. Children

Q-Learning

Q-Learning: a popular machine learning algorithm

Many others: SARSA, temporal difference, brute force, etc.

Q-Learning

Q-Learning: a popular machine learning algorithm

Many others: SARSA, temporal difference, brute force, etc.

$$Q^{new}(s_t, a_t) \leftarrow (1-\alpha) \cdot \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} \right)}^{\text{learned value}}$$

Q-Learning

Q-Learning: a popular machine learning algorithm

Many others: SARSA, temporal difference, brute force, etc.

$$Q^{new}(s_t, a_t) \leftarrow (1-\alpha) \cdot \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} \right)}^{\text{learned value}}$$

Initialized

Q-Table		Actions					
		South (0)	North (1)	East (2)	West (3)	Pickup (4)	Dropoff (5)
States	0	0	0	0	0	0	0
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	327	0	0	0	0	0	0
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	499	0	0	0	0	0	0

Training

Q-Table		Actions					
		South (0)	North (1)	East (2)	West (3)	Pickup (4)	Dropoff (5)
States	0	0	0	0	0	0	0
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	328	-2.30108105	-1.97092096	-2.30357004	-2.20591839	-10.3607344	-8.5583017
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	.	.	.	.	.	.	.
	499	9.96984239	4.02706992	12.96022777	29	3.32877873	3.38230603

Q-Learning — Paramaters

$$Q^{new}(s_t, a_t) \leftarrow (1-\alpha) \cdot \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} \right)}^{\text{learned value}}$$

Parameters:

1. r_t: reward given
2. α: Learning rate (stochastic, $0 < \alpha < 1$, deterministic)
3. γ: Temporal discounting: the value of future rewards (short-term gain, $0 < \gamma < 1$, long-term gain)
4. $Q^0(s_t, a_t)$: Initialization can influence learning

Q-Learning — Paramaters

$$Q^{new}(s_t, a_t) \leftarrow (1-\alpha) \cdot \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} \right)}^{\text{learned value}}$$

Parameters:

1. r_t: reward given
2. α: Learning rate (stochastic, $0 < \alpha < 1$, deterministic)
3. γ: Temporal discounting: the value of future rewards (short-term gain, $0 < \gamma < 1$, long-term gain)
4. $Q^0(s_t, a_t)$: Initialization can influence learning

Optional Homework*

Big Ideas of Reinforcement Learning

1. Explore vs. Exploit
2. Model-Based vs. Model-Free
3. Reward Prediction Error

Exploration

Exploration: Find out more information about the environment by searching a lot of possibilities

1. Try every possibility
2. Weakly tend to select actions that maximize reward

Exploitation

Exploitation: capitalize on known information to maximize reward

1. strongly tend to select actions that maximize reward

Exploration-Exploitation Tradeoff

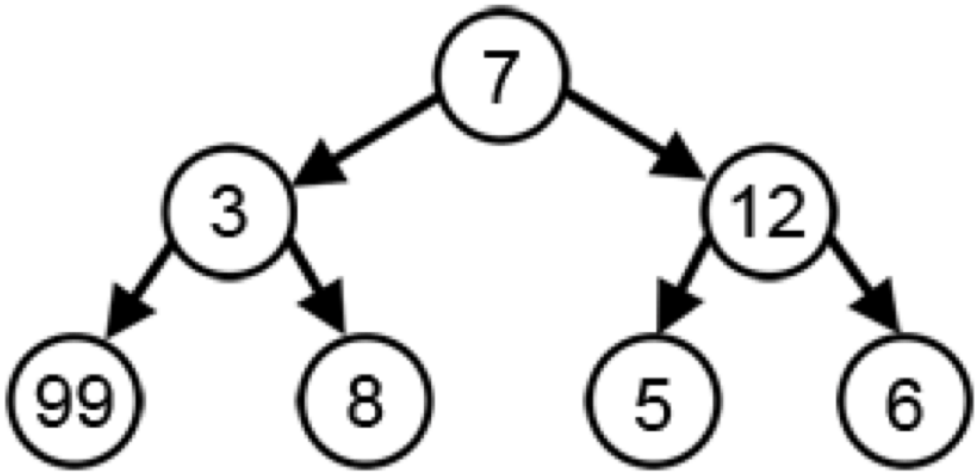
Shift emphasis from exploring to exploiting to maximize reward

1. strongly tend to select actions that maximize reward

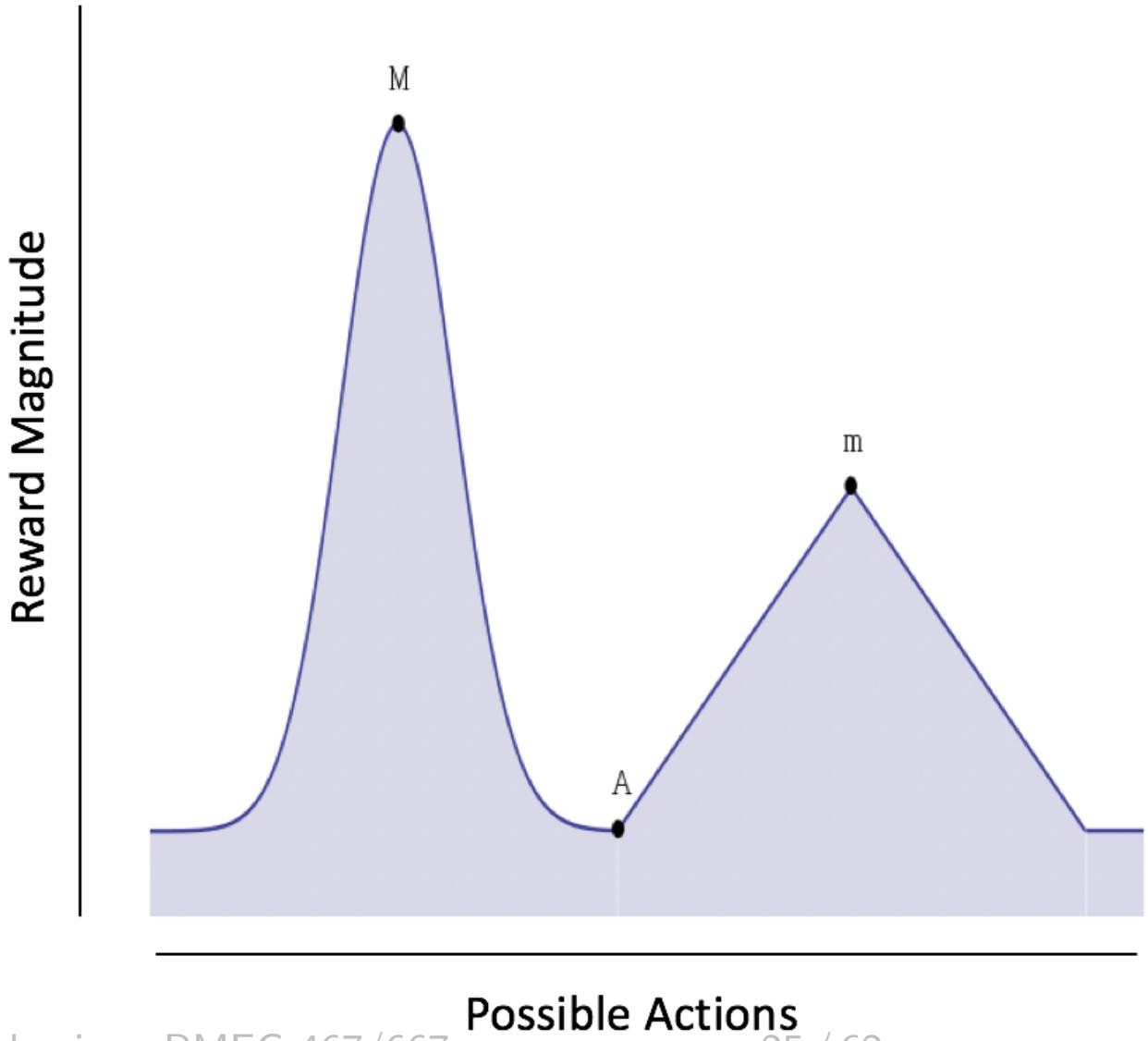
Sports Examples?

Greedy vs. Non-Greedy Algorithms

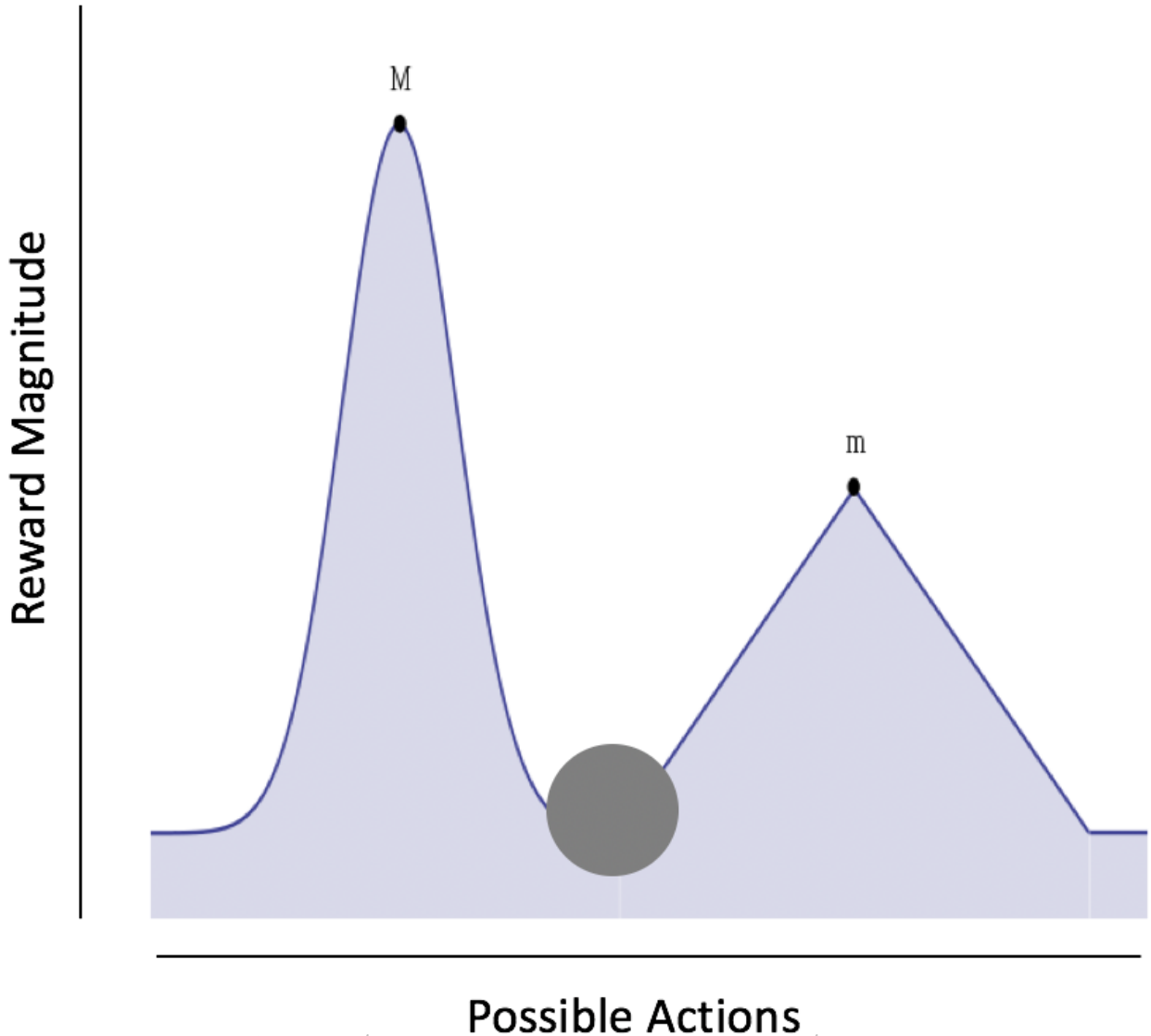
1. Greedy Algorithms Do Not Explore
2. Non-Greedy Explore



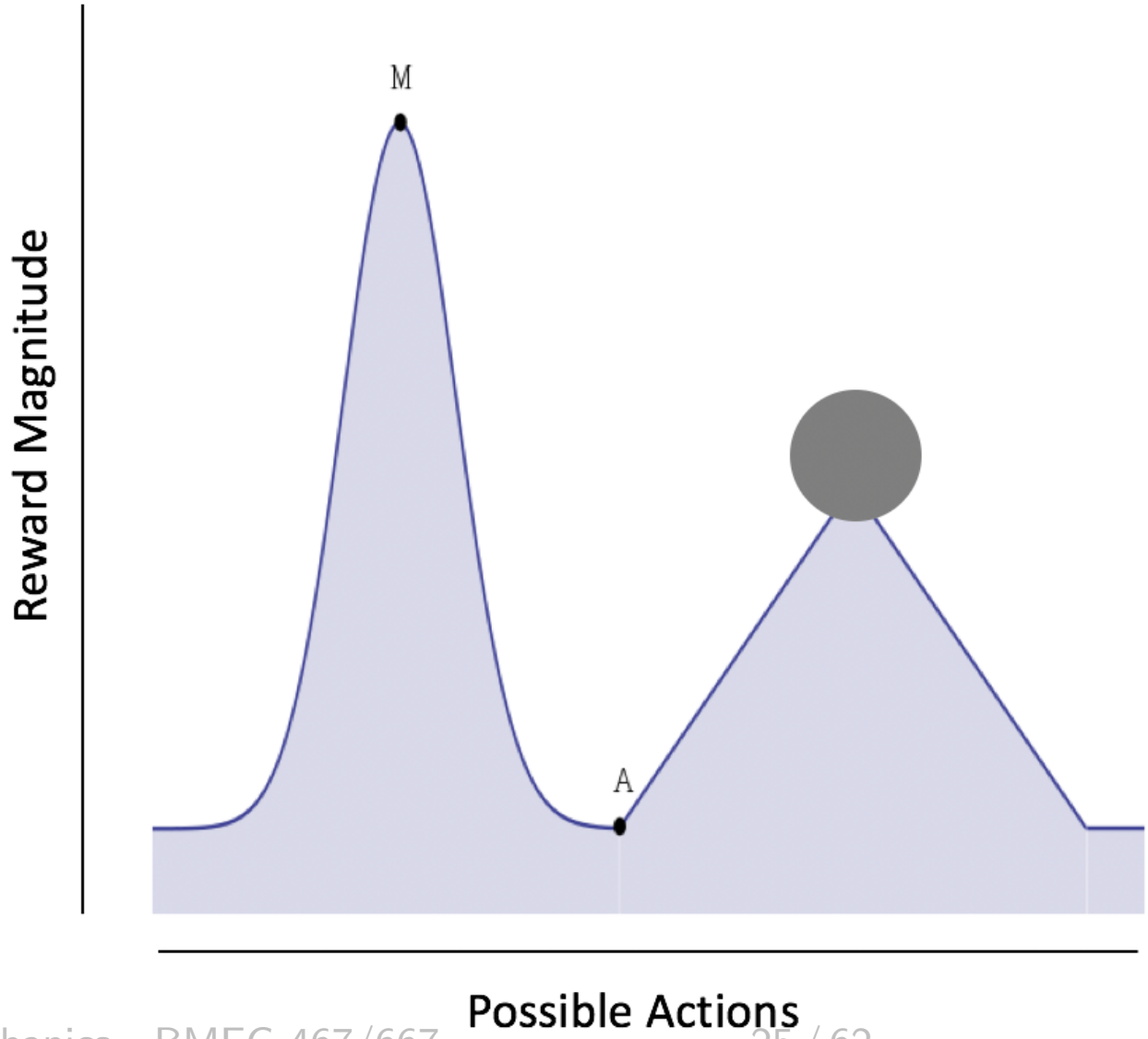
Greedy vs. Non-Greedy Algorithms



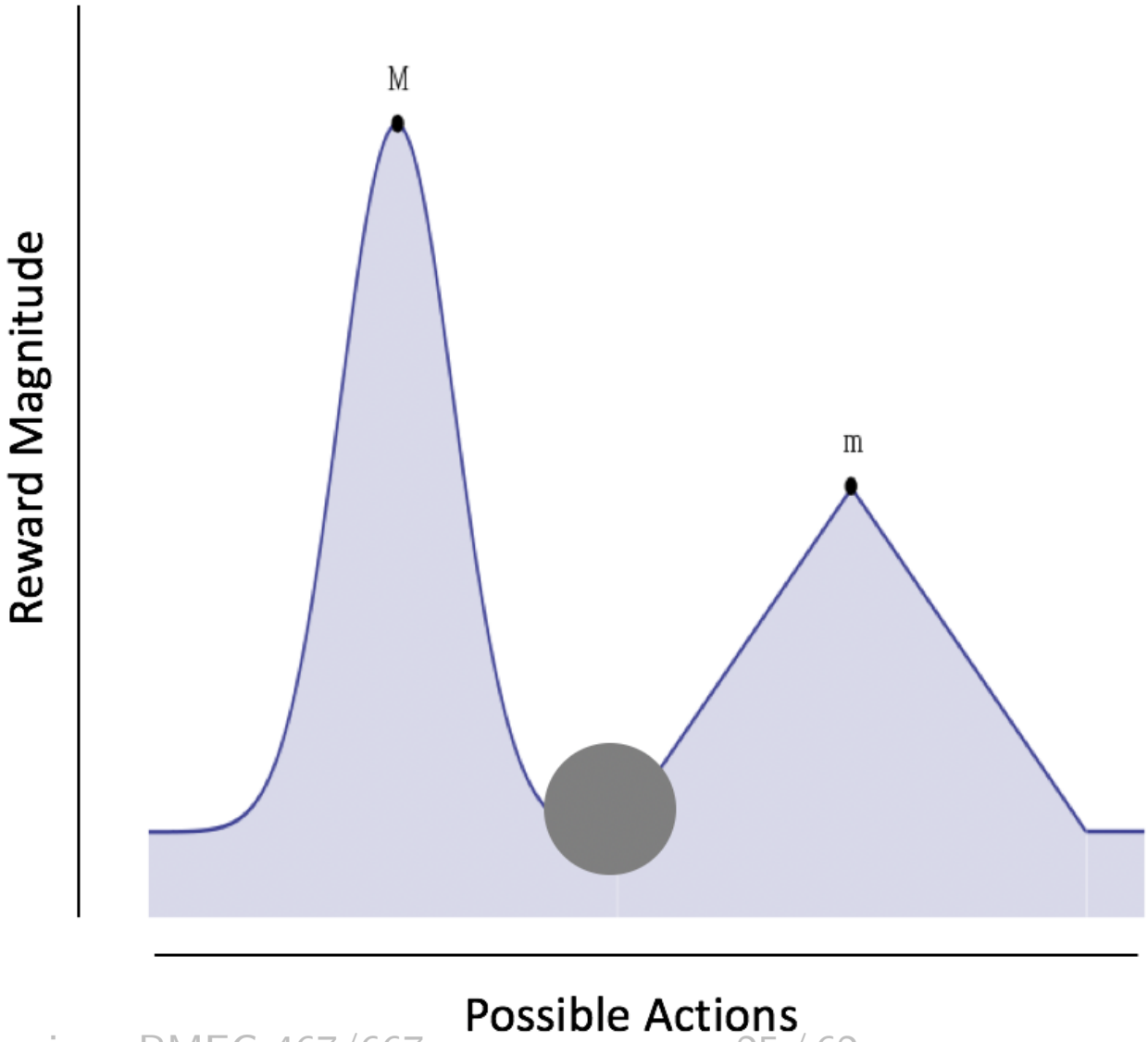
Greedy vs. Non-Greedy Algorithms



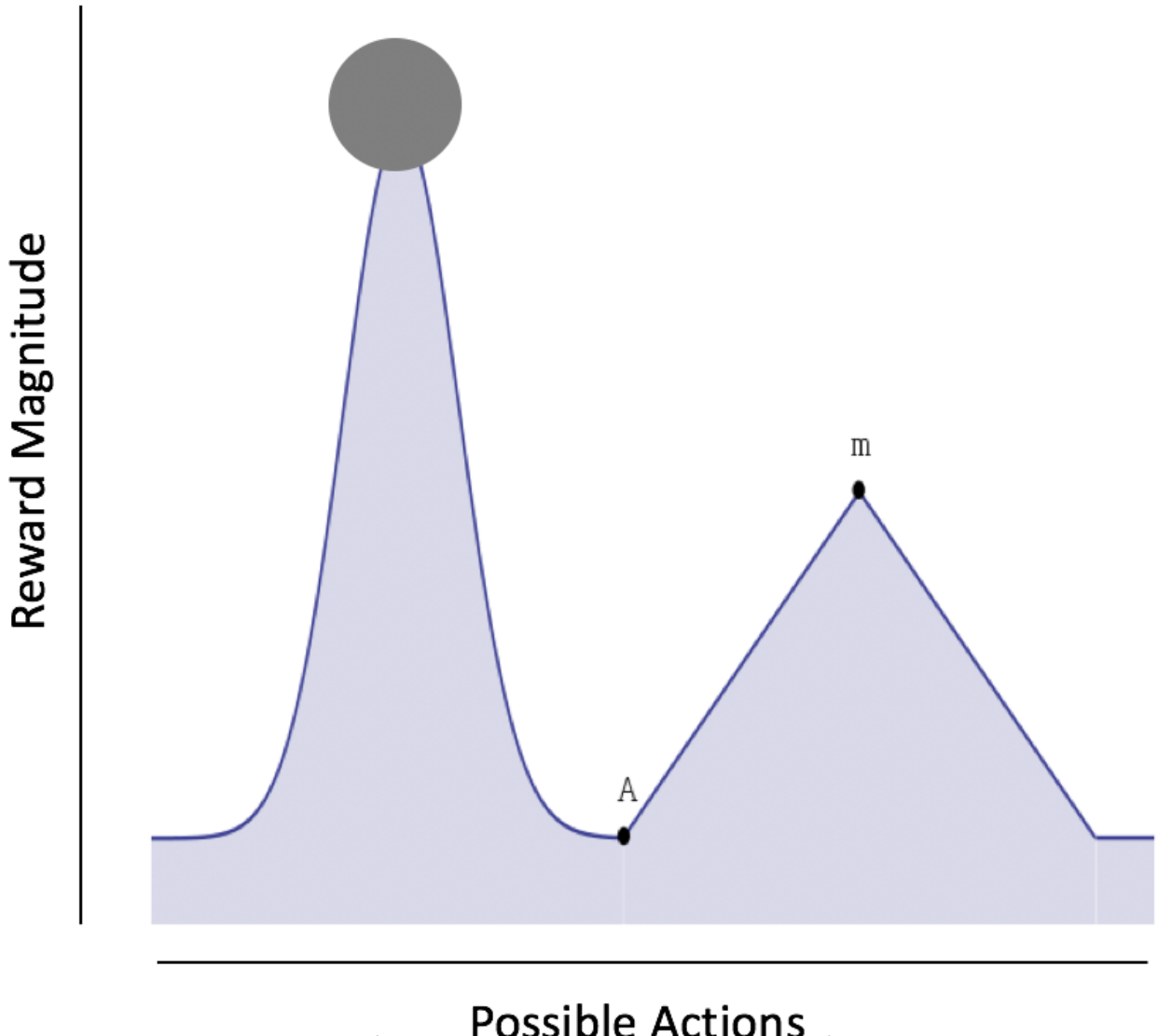
Greedy vs. Non-Greedy Algorithms



Greedy vs. Non-Greedy Algorithms



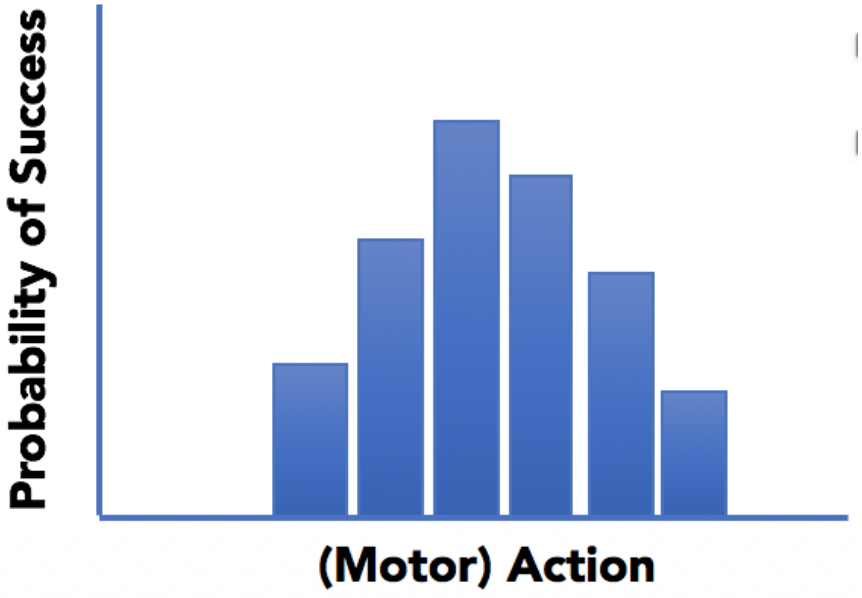
Greedy vs. Non-Greedy Algorithms



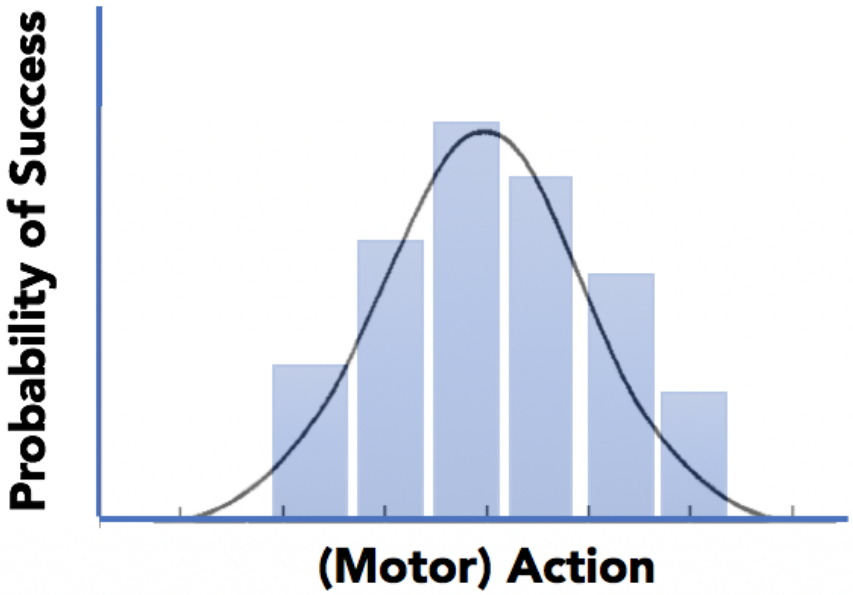
Greedy Algorithms Examples

1. Epsilon-greedy:
   - Best action selected for: $1 - \epsilon$ (e.g., $\epsilon = 0.1$)
   - uniform random actions selected for ϵ
   - can vary epsilon initially or over time
2. Soft-Max Function
   - $P_t(a) = \frac{\exp(q_t(a)/\tau)}{\sum_{i=1}^{n} \exp(q_t(i)/\tau)}$
   - $P_t(a)$ probability of action (a) selected
   - $q_t(a)$ corresponds to the expected reward of some action, a (vector)
   - i and n: some action and total number of possible actions
   - τ: 'temperature' (0 select action with most reward; ∞ select all actions with equal probability)

Model-Based and Model-Free



Model-Based and Model-Free



Model-Based and Model-Free — Example

1. Model-Based
   a. Advantages
      . Less time to learn
   b. Disadvantage
      . Only as good as model
      . Computationally complex
2. Model-Free
   a. Advantages
      . Less computation
   b. Disadvantage
      . Trial and error (Longer to Learn)

BEHAVIOUR

RL and Biological Behaviour

1. Predict Animal Foraging Patterns
   a. Explore-exploit, energy, reward, risk
2. Humans
   a. Predict Gambling Patterns (economics)
   b. Models of Depression
   c. How we acquire new motor skills!

RL and Biological Behaviour

1. Predict Animal Foraging Patterns
   a. Explore-exploit, energy, reward, risk
2. Humans
   a. Predict Gambling Patterns (economics)
   b. Models of Depression
   c. How we acquire new motor skills!

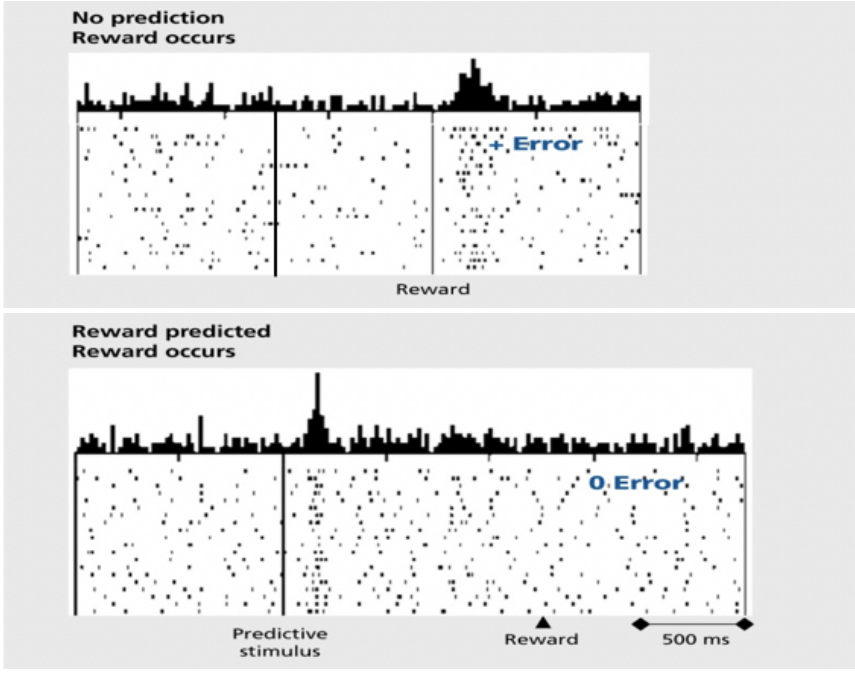
LINK = REWARD PREDICTION ERROR!

Reward Prediction Error

- monkey, basal ganglia, dopamine
- Before Learning (unconditioned)
- Positive Reward Prediction Error

Reward Prediction Error

- monkey, basal ganglia, dopamine
- Before Learning (unconditioned)
- Positive Reward Prediction Error
- After Learning (conditioned)
- No Response

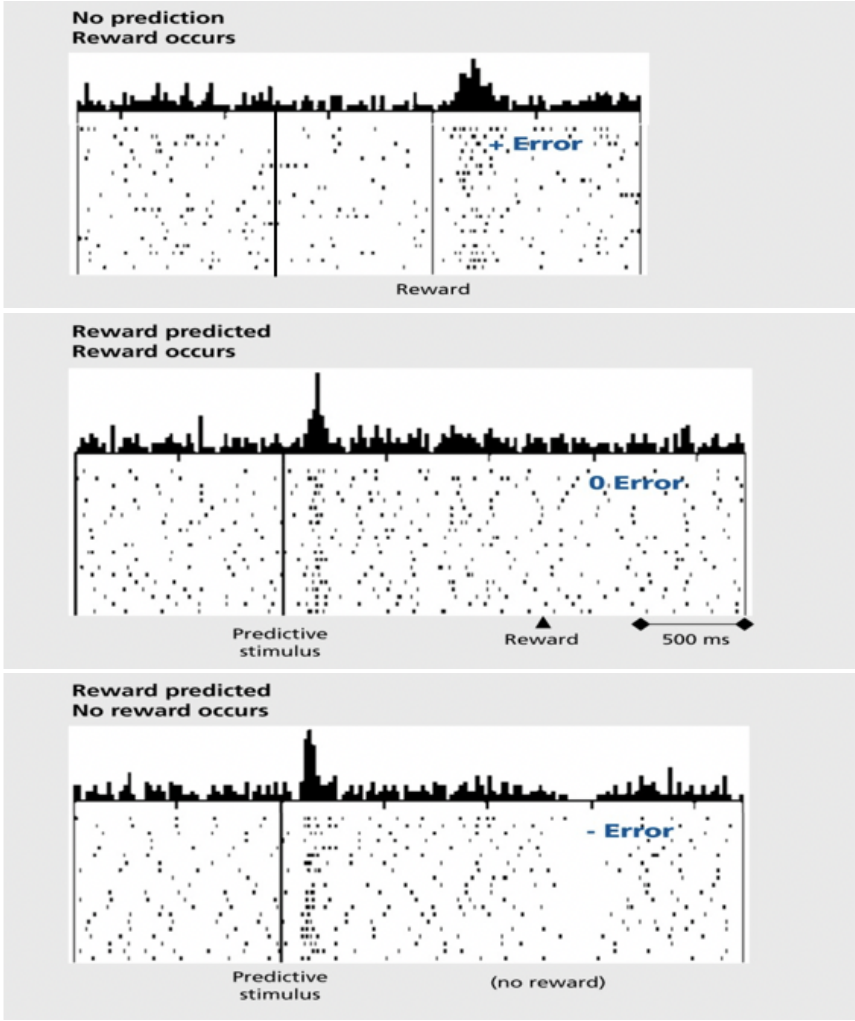


Reward Prediction Error

- monkey, basal ganglia, dopamine
- Before Learning (unconditioned)
- Positive Reward Prediction Error

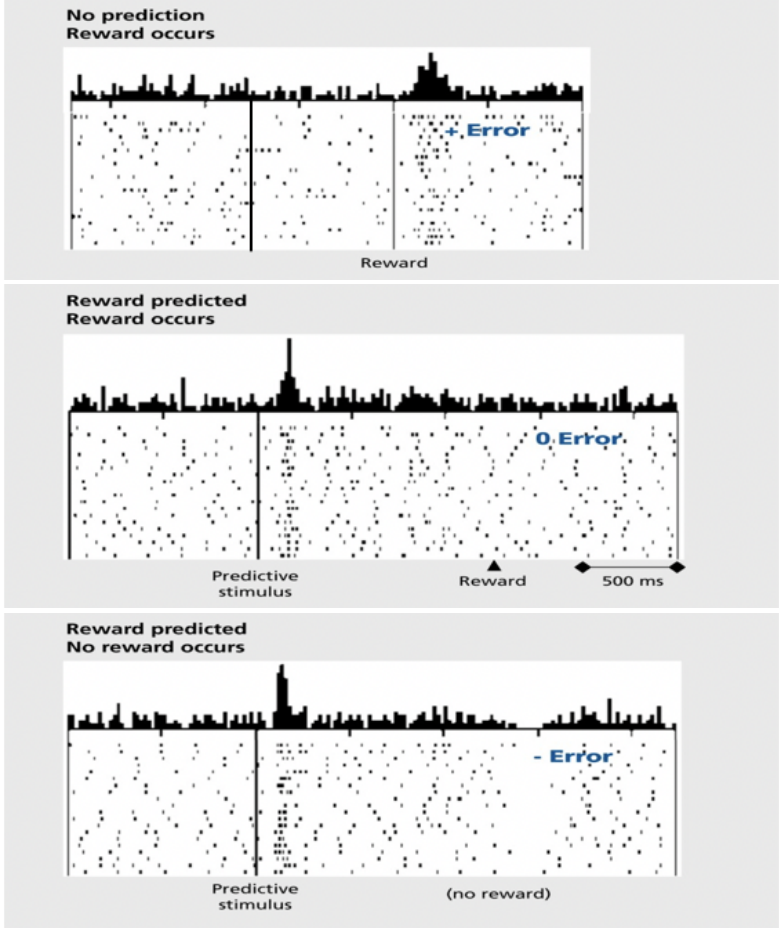
- After Learning (conditioned)
- No Response

- After Learning (conditioned)
- Negative Reward Prediction Error

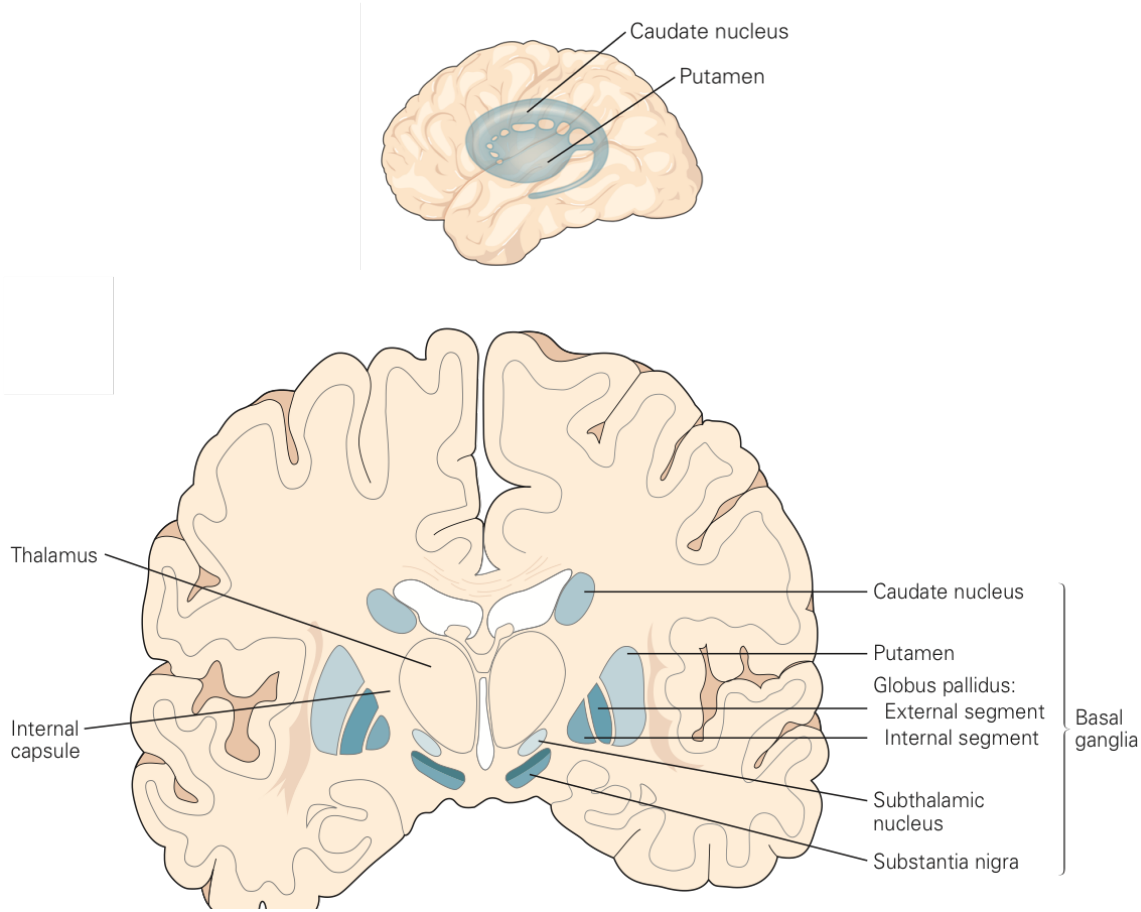


Reward Prediction Error

Dopamine Release = Observed Reward - Expected Reward



Basal Ganglia — Anatomy

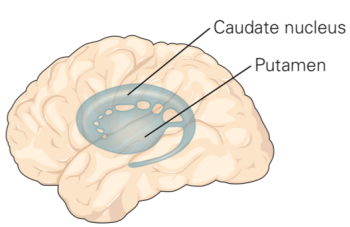


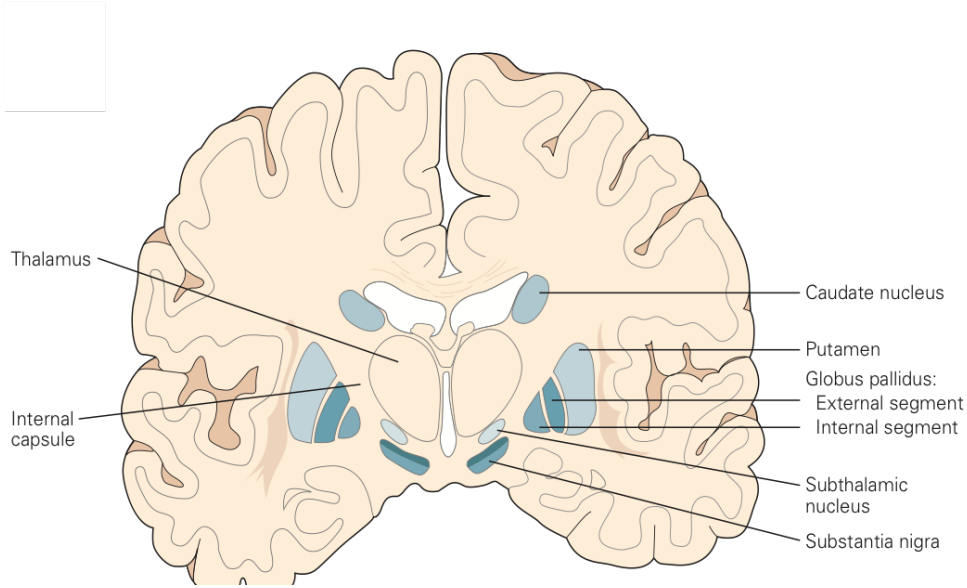
Basel Ganglia is deep in the cerebral hemispheres in the brain stem.

Basal Ganglia — Anatomy

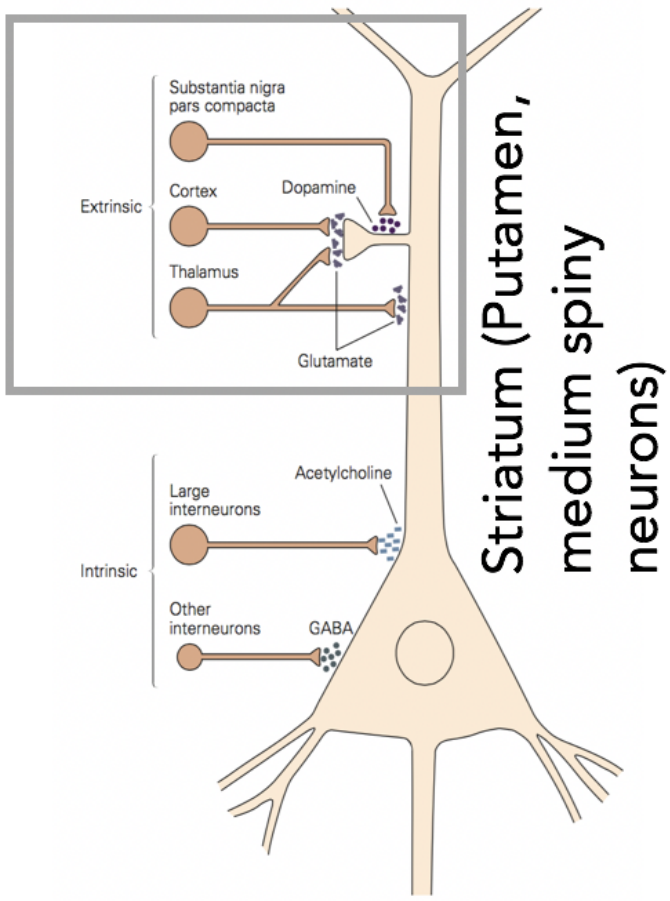
Four Major Areas

1. Striatum
   a. Caudate Nucleus
   b. Putamen
2. Globus Pallidus
   a. External segment
   b. Internal segment
3. Substantia Nigra
   a. Pars compacts (dopamine)
   b. Pars reticulata
4. Subthalamic Nucleus





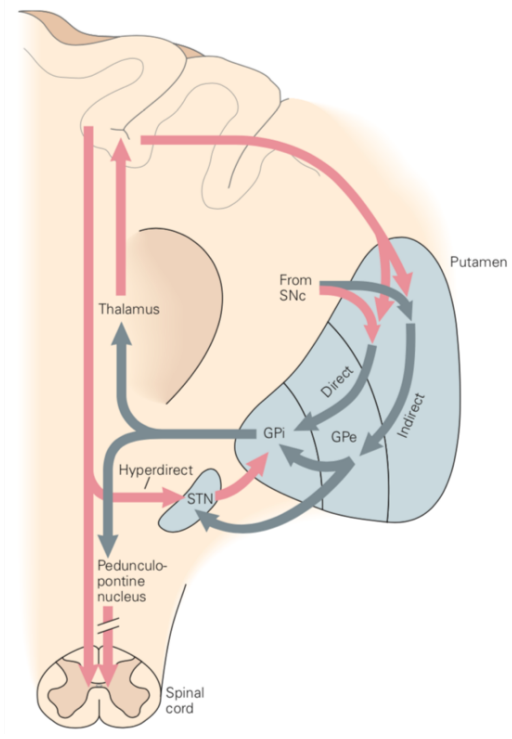
Dopamine Release



Dopamine Release

a. Dependent on conditioning
b. Reward prediction Error

Excite and Inhibit



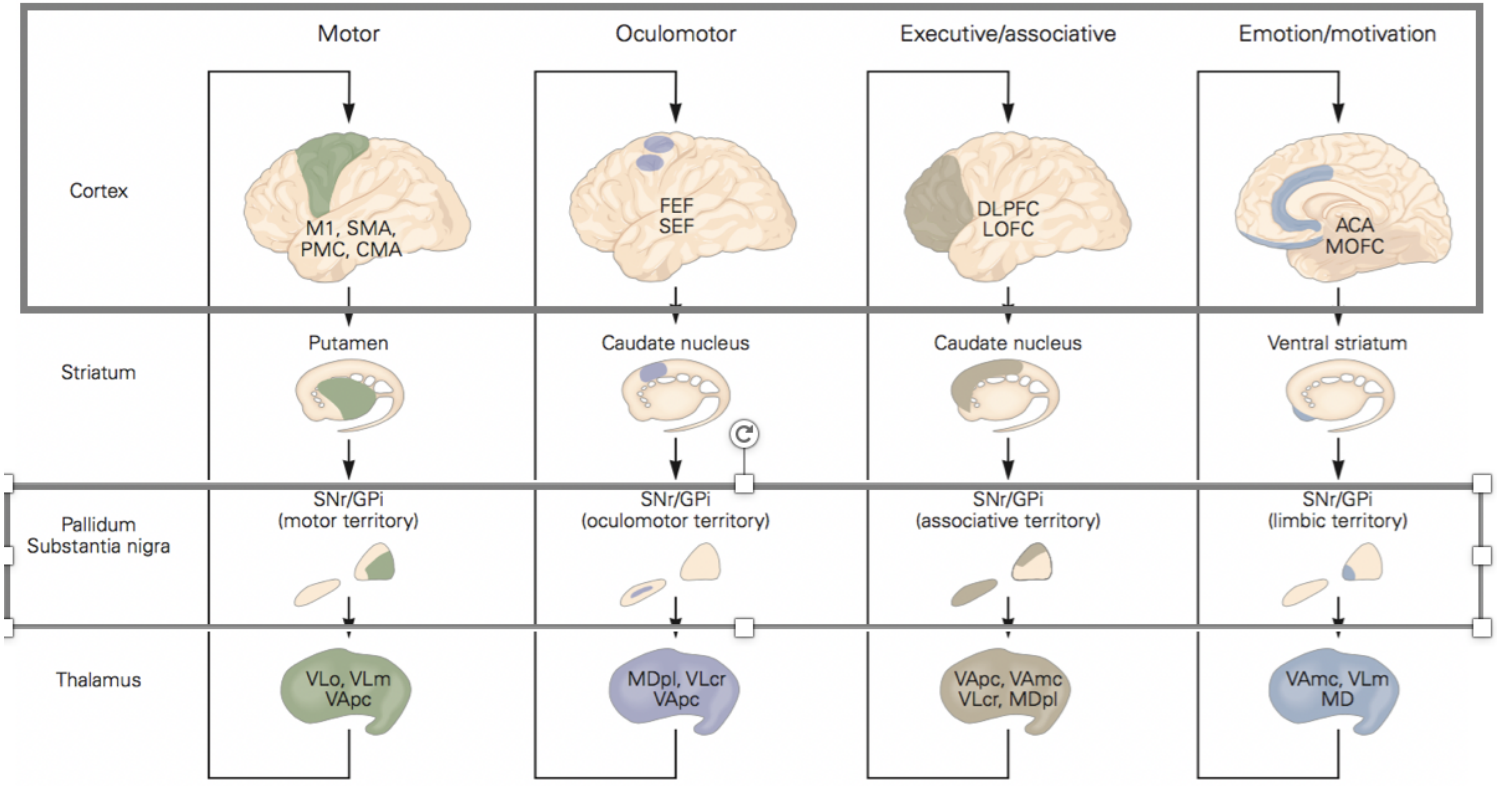
Substantia Nigra pars compacta

Dopamine Release

a. activates excitatory pathways
b. prevents inhibitory pathways from activating
c. Strengthens synapses that were involved in generating the rewarded behaviour

Red are excitatory pathways, grey are inhibitory pathways.

Cortico-Basal Ganglia-Thalamo-Cortical Loops



Motor cortex $\rightarrow$ Basal Ganglia $\rightarrow$Thalamus$\rightarrow$Motor cortex

Disruption to the Basal Ganglia

1. hypokinetic disorders (e.g., Parkinson's)
   a. inhibits motor cortex
2. hyperkinetic disorders (e.g., Huntington's)
   a. excites motor cortex

Parkinson's Disease

Click Me: ▸ Link

Parkinson's Disease — Dance Therapy

Click Me: ▸ Link

Parkinson's Disease — Symptoms

1. Dopamine Deficiency
2. Slowed movement (bradykinesia)
3. Tremor
4. Rigid muscles
5. Impaired posture and balance
6. Loss of automatic movements (e.g., blinking, smiling)
7. Speech changes
8. Writing changes
9. Cognitive and Psychological

Huntington's Disease

Click Me: ▸ Link

Huntington's Disease — Symptoms

1. Degeneration of Striatum
2. Involuntary jerking
3. Muscle rigidity
4. Rigid muscles
5. Abnormal eye movements
6. Impaired gait, posture and balance
7. Difficulty with speech or swallowing
8. Cognitive and Psychological

The Many Roles of Dopamine

Many functions depend on dopamine, thus the likely reason RL is so successful across so many domains

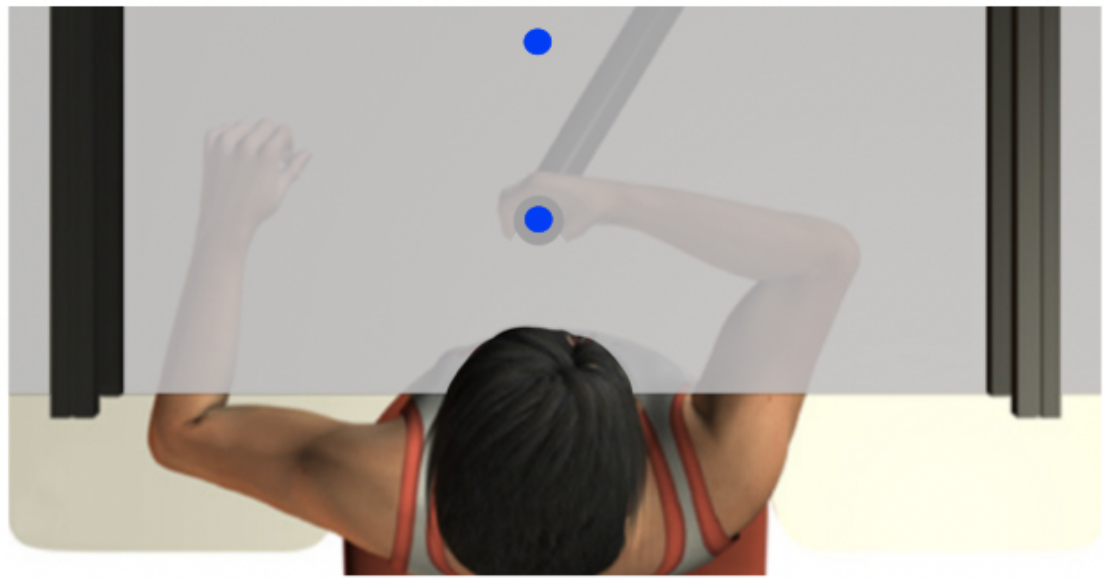
1. Motor Actions
   a. Limbs, eyes
   b. Disease states
2. Cognitive
   a. Executive
   b. Emotions (Depression)
3. Addictions
   a. Gambling
   b. Drugs
4. Lower-level functions
   a. Lactation
   b. Sexual Gratification
   c. Nausea

Sensorimotor Learning

Binary Reinforcement Feedback

Demo

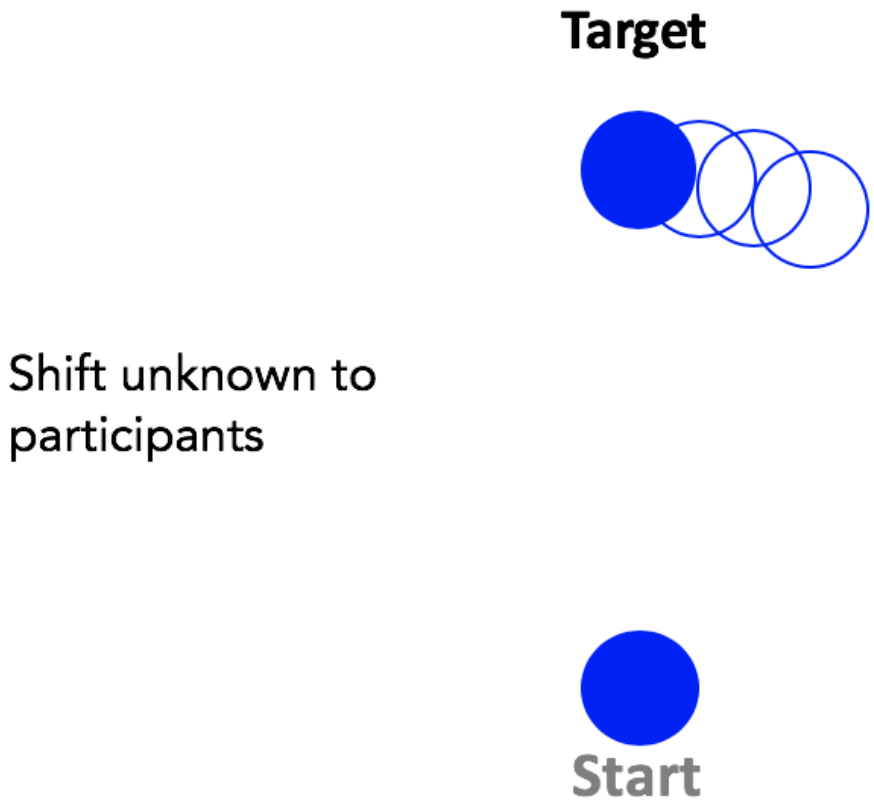
Izawa and Shadmehr (2011)



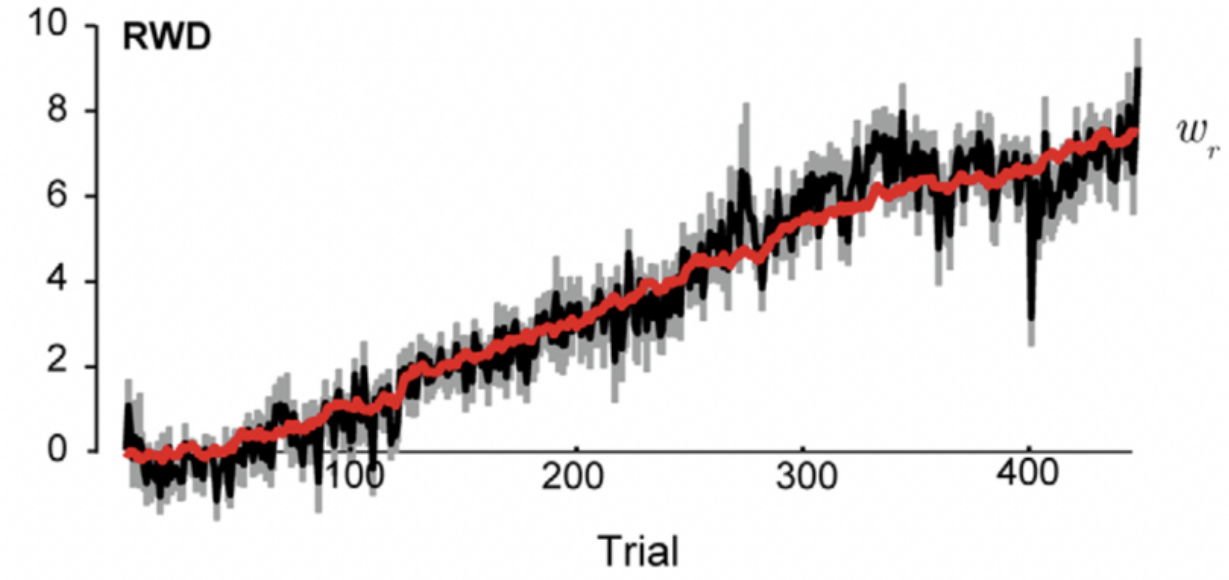
Reinforcement Feedback (Hit)

Reinforcement Feedback (Miss)

Shift Actual Target



Motor Learning

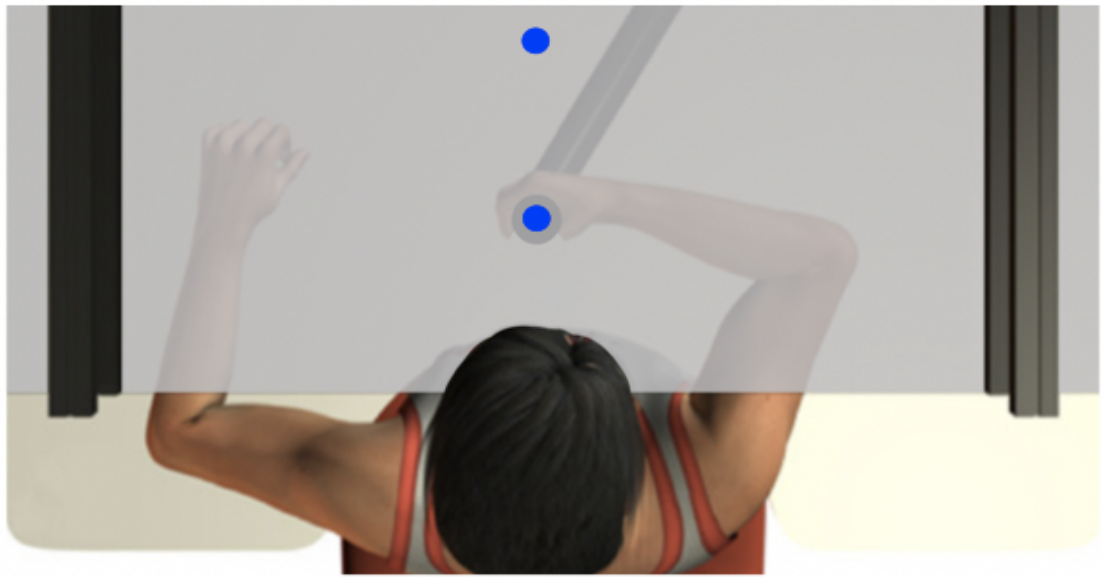


Participants Unaware of the Target shift

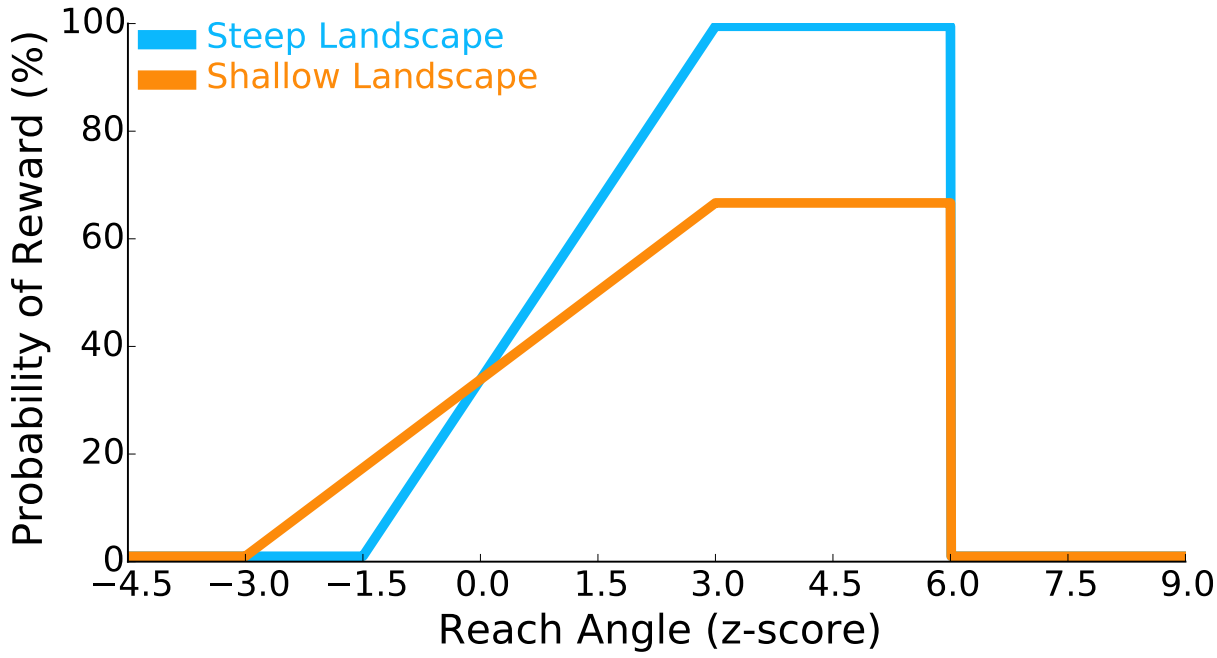
Red line represents the predicted behaviour by a RL algorithm

Cashaback et al., 2019

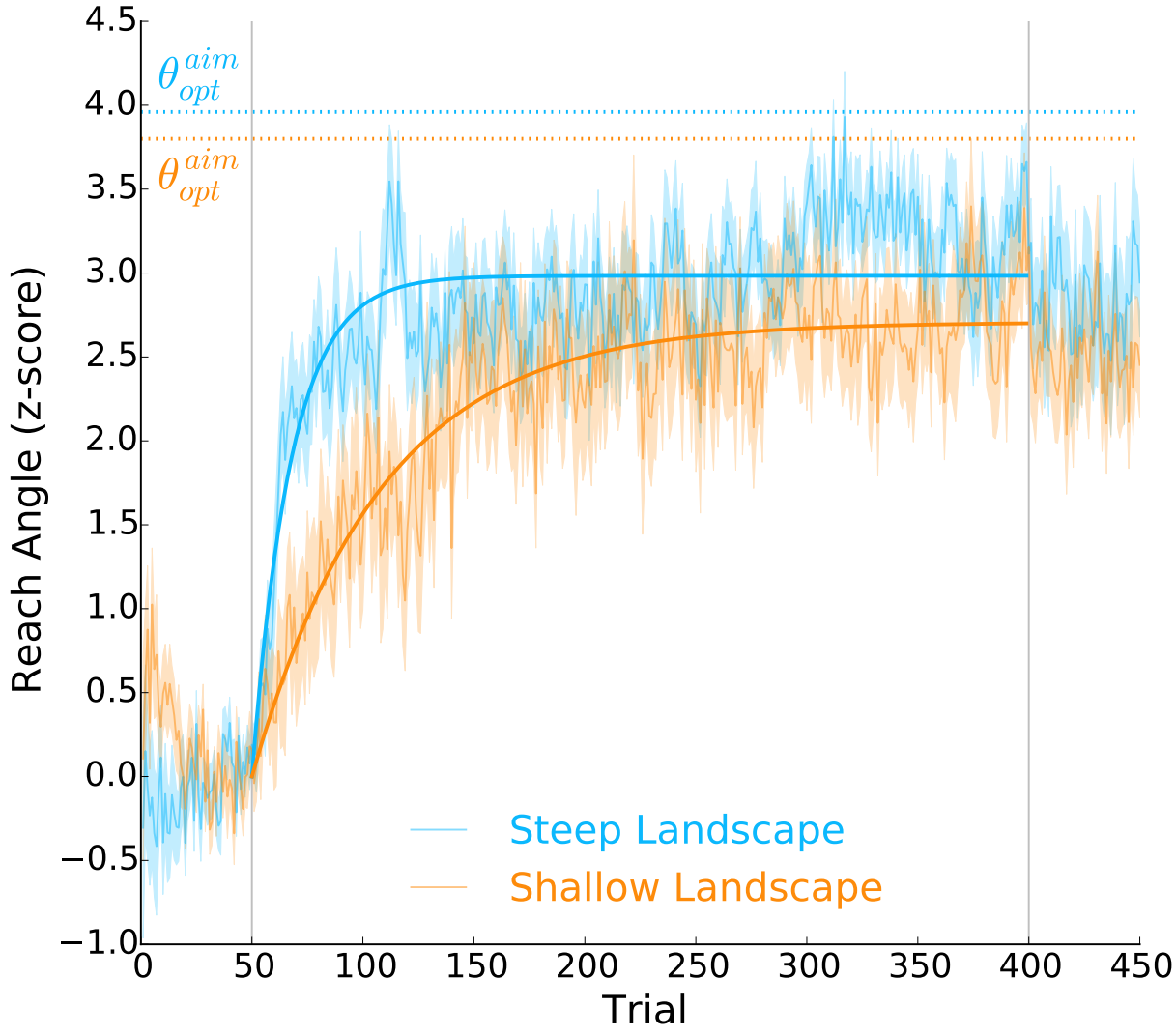
Does the Gradient Influence Learning?



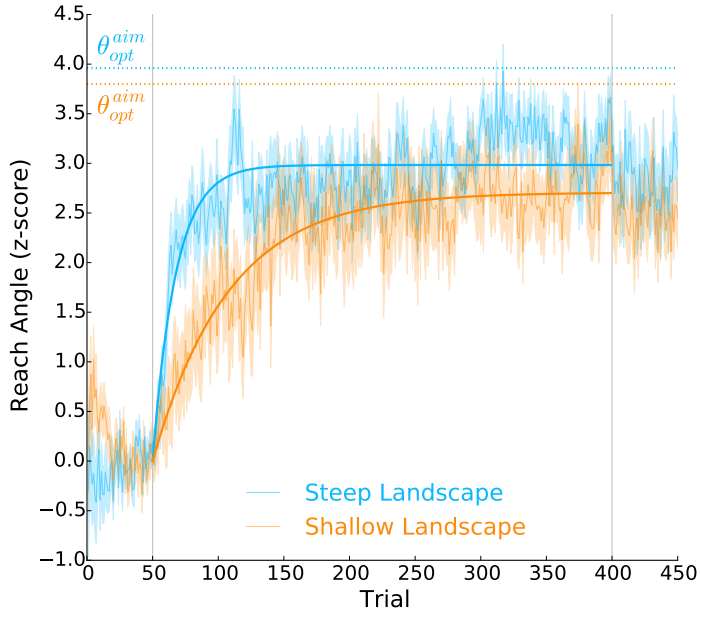
Manipulating the Gradient — E1

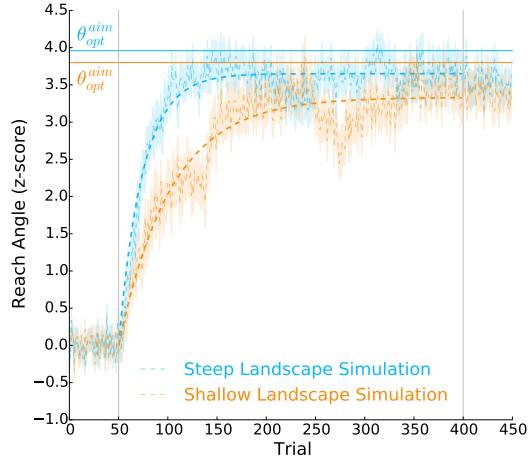


Steeper = Faster Learning

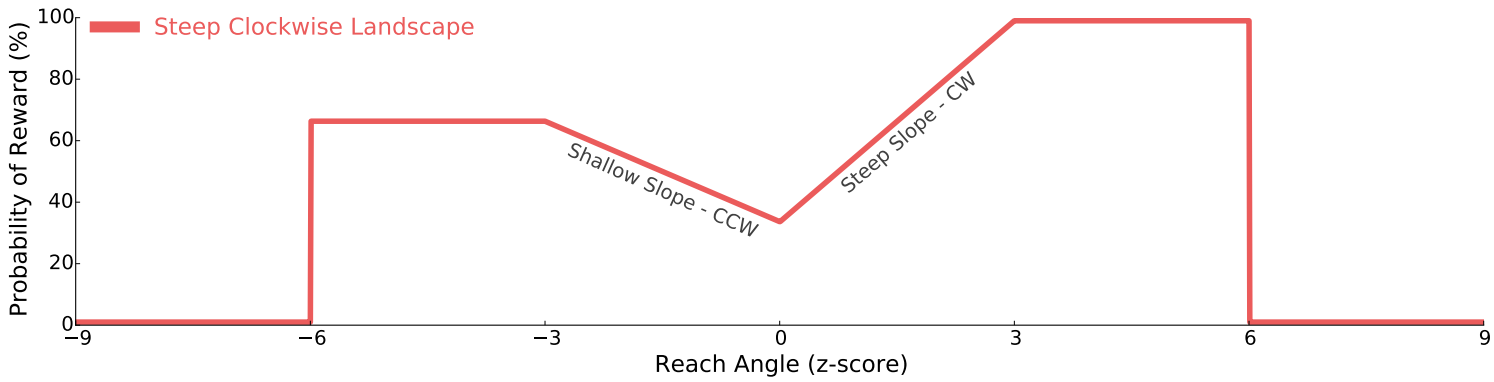


Steeper = Faster Learning

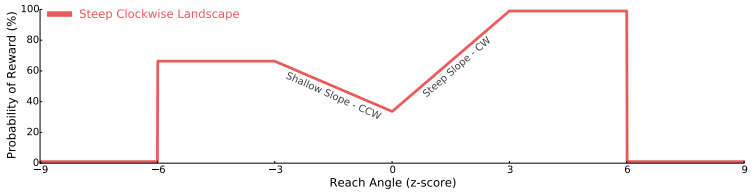


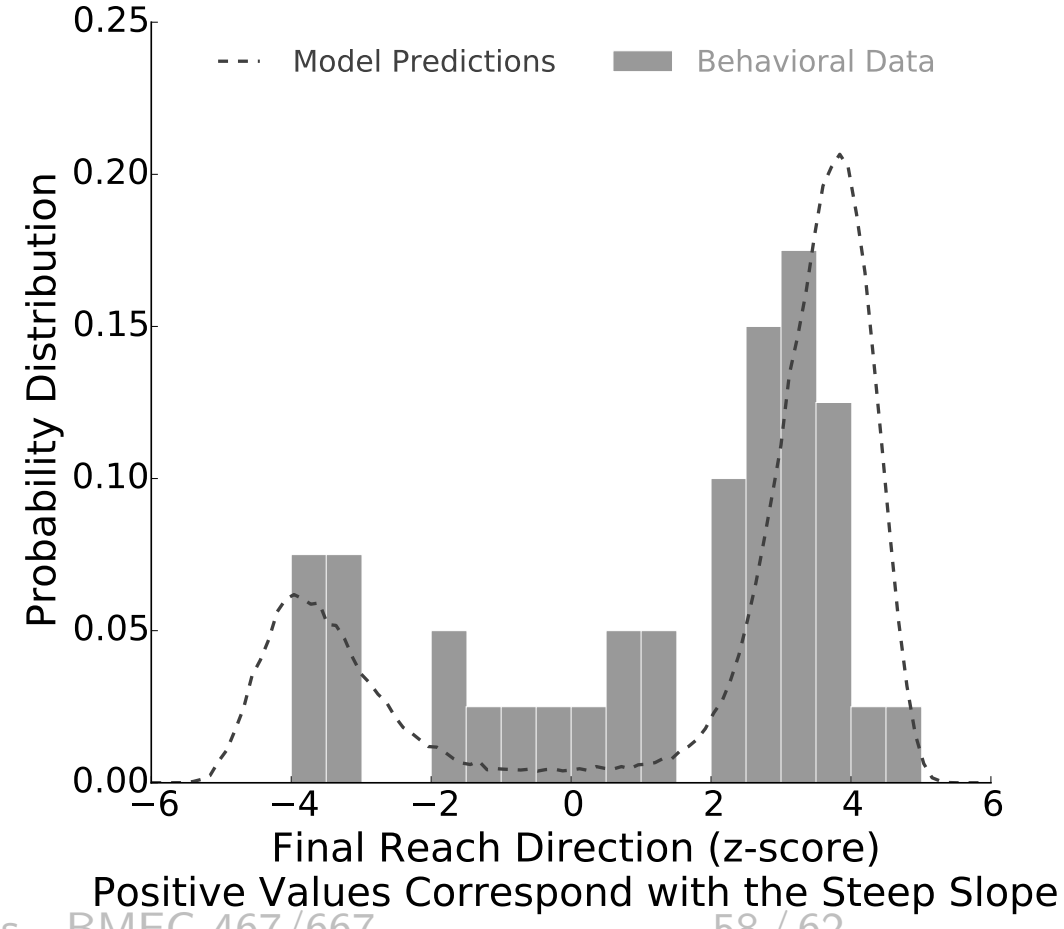


Manipulating the Gradient — E2



More Likely to Ascend the Steep Slope





Summary

1. Beginnings
   - Psychology, Classical Conditioning
   - (Positive, Negative, Reinforcement, Punishment)
2. Breadth
   - Many disciplines
   - Max. reward, exploration-exploitation, greedy, model
3. Behaviour
   - Dopamine Release and Reward-Prediction Error
   - Parkinson's, Huntington's Disease
   - Motor Learning

Summary

1. Beginnings
   - Psychology, Classical Conditioning
   - (Positive, Negative, Reinforcement, Punishment)
2. Breadth
   - Many disciplines
   - Max. reward, exploration-exploitation, greedy, model
3. Behaviour
   - Dopamine Release and Reward-Prediction Error
   - Parkinson's, Huntington's Disease
   - Motor Learning

TAKE HOME:

- Selecting Actions Can Be Optimized—Cumulative Reward
- This RL framework transcends and guides many disciplines

Questions???

Homework

1. Program a Soft-Max Function
   - $P_t(a) = \frac{\exp(q_t(a)/\tau)}{\sum_{i=1}^{n} \exp(q_t(i)/\tau)}$
   - Use 5 targets
   - set their reward value to 5, 10, 15, 10, 5.
   - manipulate the temperature, τ
   - use a histogram to plot target selection frequency give some τ
2. Great Primer on Q-learning (**optional**) : ▸ Link
   - manipulate α and γ
   - include a soft-max function

Next Class

Midterm Review

Stereotypical Human Behaviour

1. Hick's Law (reaction time and choice)
2. Fitt's Law (speed vs. accuracy tradeoff)
3. Bell-shaped Velocity Profiles
4. Behaviour of Redundant Systems