The Impact of Over-training on Cognitive Flexibility in Bayesian Reinforcement Learning Models

Submitted September 2024, in partial fulfillment of
the conditions for the award of the degree **Msc Computational Neuroscience, Cognition and AI.**

Salaar Mir
20275881

Supervised by Dr. Silvia Maggi

School of Psychology
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature: Salaar Mir

Date: 09/09/2024

Abstract

Cognitive flexibility is the ability to adapt behaviour in response to changing conditions. Despite the large body of literature examining the neural mechanisms underlying this intrinsic property, there are contradictory findings on the impact of over-training on cognitive flexibility. Recent advances in computational models provide new opportunities to better simulate such cognitive processes, offering insights that may resolve some of these conflicting findings. Here, a Bayesian reinforcement learning (RL) model was created to explore the impact of over-training on the adaptive behaviour of simulated agents in a simple reversal learning task. The results showed that while all agents successfully learned the task, over-trained agents displayed slower adaptation to rule changes, reflecting reduced cognitive flexibility. These agents were more prone to developing rigid, habitual behaviours, which impaired their ability to adjust to new conditions. Additionally, the findings suggest that the Bayesian RL models developed had limitations in incorporating meta-learning, as agents struggled to recognise patterns in the task structure and adapt accordingly.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr. Silvia Maggi, for her continuous guidance throughout the course of this dissertation. Her constant encouragement and support has been tremendously valuable in motivating me to push my boundaries and give my best effort on this important project. I am especially thankful for her patience and positivity during the summer term, which made this journey both rewarding and enjoyable.

Contents

Chapter 1

Introduction

1.1 Background

Humans face countless decisions everyday, often requiring the ability to adapt their choices based on new information. Cognitive flexibility refers to an individual's ability to adapt their thinking and behaviour in response to changing environmental stimuli and conditions (Cañas, 2006). It is a critical component of executive functions and plays a key role in problem-solving, decision-making, and adaptive behaviour in dynamic settings. In both humans and animals, cognitive flexibility is often assessed through tasks that involve set-shifting or task-switching.

Theories surrounding cognitive flexibility suggest that factors such as inhibitory control, salience detection, and working memory all contribute to an individual's ability to shift between different behavioural strategies (Diamond, 2013). Reinforcement learning (RL), particularly Bayesian RL, provides a computational framework to simulate decision-making and cognitive processes by allowing agents to learn optimal actions in environments where rewards and conditions are uncertain.

Over-training, where subjects undergo extensive training on a specific task, has been shown to impact cognitive flexibility in various ways, with some studies suggesting that it enhances task performance, while others arguing that it may reduce adaptability. The Over-training Reversal Effect (ORE) is one phenomenon that has been exhibited in a number of empirical studies looking at cognitive flexibility. This phenomenon proposes

that over-training on a task can lead to faster adaptation to changes in conditions, likely due to strengthening of neural pathways and build-up of reactive inhibition (Schmidt, De Houwer, & Moors, 2020; Reid, 1953). However, multiple scientific studies have also failed to exhibit this effect, suggesting the evidence is still inconclusive.

1.2 Research Problem

Despite the growing body of research on cognitive flexibility and over-training, the literature presents conflicting findings on the subject. Some studies report that over-training enhances adaptability, while others suggest that it leads to more rigid behaviour, reducing cognitive flexibility. These contradictions are particularly evident in studies examining the ORE. For example, Reid first demonstrated the ORE by showing that over-trained subjects reversed their learning more quickly in specific tasks (Reid, 1953). However, later studies, such as Uhl (1964), provided contradictory evidence, suggesting that over-training might be context-dependent and could actually hinder adaptability (Uhl, 1964). This inconsistency in the literature points to a gap in our understanding of how over-training influences cognitive flexibility.

Furthermore, although RL has been extensively used to model cognitive processes, there is a significant gap in research exploring how over-training affects flexibility using RL models. The potential for RL to provide a computational perspective on over-training has not been fully explored, leaving a critical gap in understanding how reinforcement impacts learning adaptability.

1.3 Objectives and Hypotheses

This study aims to serve as an exploratory investigation into whether Bayesian RL models can accurately replicate the ORE observed in empirical experiments involving cognitive flexibility. It will also examine the role of meta-learning in RL agents and how reinforcement mechanisms may lead to habitual behaviour.

Based on existing literature, the following hypotheses are proposed:

- Over-trained RL agents will not exhibit the ORE, as their ability to adapt to new rules may be hindered by excessive reinforcement of learned behaviours.
- RL agents will display an absence of meta-learning, leading to more rigid behaviour under over-training conditions.
- Continuous reinforcement will lead to habitual behaviour in over-trained agents, reducing their adaptability.

RL models will be developed to investigate whether they can replicate the adaptive behaviour of biological systems in dynamic environments. Agents will be subjected to various training conditions within a simulated reversal learning task, and the results of these simulations will be analysed to explore their potential implications.

Chapter 2

Related Work

2.1 Cognitive Flexibility

Cognitive flexibility is a key component of executive functions, allowing individuals to adapt their behaviour in response to dynamic environments. This ability is crucial for effective problem-solving and decision-making and is typically measured in laboratory settings using task-switching and set-shifting behavioural experiments (Dajani & Uddin, 2015). There has been significant work done looking at cognitive flexibility and the related brain functioning that contributes towards this intrinsic property. A review article published by Diamond emphasised that cognitive flexibility is closely linked to other executive functions such as inhibitory control and working memory, which form the foundation for adaptive behaviour in dynamic settings (Diamond, 2013). The article discusses one aspect of cognitive flexibility being the ability to change perspectives, which requires the inhibition of a previous view and the loading (or activation) of a different perspective into working memory. Further work has also highlighted the need for salience detection and mental set-shifting to allow for cognitive flexibility (Dajani & Uddin, 2015; Uddin, 2021). The salience of a stimulus is important in identifying whether it will capture attention and be further processed by the brain. Additionally, mental-set shifting (or 'shifting') is the ability to switch between different tasks (von Bastian & Druey, 2017). Such mental capacity is necessary to be able to switch between different perspectives during a specific task.

2.2 Reversal Learning

Reversal learning tasks are widely used to study cognitive flexibility in both humans and animals. These tasks require individuals to learn and then relearn associations, adapting to new rules or stimuli as they are presented. This process is similar to real-world scenarios where the ability to reverse previously learned behaviours is essential for adaptive functioning (Izquierdo, Brigman, Radke, Rudebeck, & Holmes, 2016). In typical laboratory settings, subjects are trained to distinguish between two spatial locations or visual stimuli, with one consistently being rewarded for being selected, while the other offering no reward. Once a set level of accuracy is achieved, indicating successful learning, the rewards associated with both stimuli are reversed, and subjects undergo retraining until they again reach the required performance standard (Izquierdo et al., 2016).

Studies on rodents using reversal learning tasks have provided valuable insights into the neural mechanisms underlying cognitive flexibility. Work done by Boulougouris et al. examined the role of serotonin receptors on cognitive flexibility using reversal learning tasks. They found that blocking specific receptors had dissociable effects on the ability of rats to adapt to rule changes, highlighting the role that serotonin and neural circuits play in managing behavioural flexibility (Boulougouris & Robbins, 2010). Further work by Costa et al. used a Bayesian perspective to highlight the role of dopamine in reversal learning (Costa, Tran, Turchi, & Averbeck, 2015).

2.3 Over-training

Over-training has been an area of growing interest in the context of cognitive flexibility. It is defined as excessive mental or physical training of a specific task, often leading to a change in task performance (Symons, Bruce, & Main, 2023). Over-training syndrome (OTS) is the extreme end of this spectrum, where prolonged over-training causes significant performance decline and psychological symptoms such as cognitive impairments and mood disturbances (Kreher & Schwartz, 2012). In typical laboratory settings that measure the impact of over-training, it is most often done by providing a certain subset

of subjects with an extended number of trials, while a control group receive the standard amount. This is effective in analysing how excessive training impacts a subjects' ability to correctly switch between strategies within a given task.

Recent studies have produced differing conclusions on the impact of over-training on cognitive flexibility. For example, research on endurance athletes shows that those in a state of over-training exhibit slower reaction times and impaired decision-making abilities (Symons et al., 2023). Contrarily, further work, particularly in tasks involving reversal learning, has demonstrated that over-training can lead to improved cognitive flexibility in certain contexts (Dhawan, Tait, & Brown, 2019).

Reversal learning tasks have been widely used to assess the impact of over-training on cognitive flexibility. Significant work has been done to see whether extended amounts of training have facilitated reversal learning compared to normally trained subjects within task-switching experiments. Interesting work by Sitterley et al. showed that, in a successive discrimination task done on human participants, over-trained subjects displayed enhanced reversal learning in comparison to participants trained only to criterion (Sitterley & Capehart, 2014). Further work done looking at rodents have produced similar results (Dhawan et al., 2019; Reid, 1953).

These are examples of ORE, which suggests that an over-trained subject is likely to adapt more quickly to a habit reversal learning task, in comparison to subjects who have not received extended training (Colman, 2015). This appears somewhat paradoxical, as traditional learning theories would suggest that more training of a particular task would 'glue in' the original habit, causing for reversal to be more difficult to adapt to (Wood & Rünger, 2016; Dickinson, 1985).

Several theories have been proposed to explain this phenomenon. Work done by Schmidt et al. suggests that over-training strengthens specific neural circuits involved in a task, making them more efficient and better prepared for adaptation when rules change (Schmidt et al., 2020). Another theory, based on the initial findings by Reid, proposes that over-training can lead to a build-up of reactive inhibition, where repeated responses to the same task increase the likelihood of subjects changing behaviours more rapidly after extended

training (Reid, 1953). However, contradictory findings from other empirical studies highlight the need for further research to better understand the mechanisms behind the ORE and under what conditions it occurs.

2.4 Reinforcement Learning Models

To better understand the impact of over-training on cognitive flexibility, computational models could be developed to simulate learning under different training conditions. RL models are computational frameworks that simulate how agents learn to make decisions by maximising cumulative rewards. Such models are being used more and more often to understand various cognitive processes, including cognitive flexibility. Sutton and Barto's seminal work on RL provides a comprehensive introduction to these models, emphasising their application in various cognitive tasks (Sutton & Barto, 1999).

Bayesian RL is another form of RL that integrates principles of Bayesian inference into traditional RL frameworks. In Bayesian RL, agents maintain a probabilistic model of the environment, updating their beliefs about the system as they gather more data (Ghavamzadeh, Mannor, Pineau, & Tamar, 2015). This approach allows for a balance of exploration and exploitation, as agents consider the uncertainty of their knowledge, leading to more efficient learning.

A large number of neuroscientific studies have linked RL to the dopamine system in the brain (Wanjerkhede, Surampudi, & Mytri, 2014; Daw & Tobler, 2014; Babayan, Uchida, & Gershman, 2018). Specifically, the basal ganglia, which is critical for learning rewards and punishments (Young & Sonne, 2018), seems to play a big role. Dopamine neurons encode a prediction error signal, which is the difference between expected and received rewards (Schultz, 1998). This error is similar to the reward signal in RL, which guide the agent's learning process. Interesting work done by Qü et al. shows that dopamine signalling in the nucleus accumbens seems to match Bayesian updating of probabilistic beliefs (Qu et al., 2023).

Studies on over-training have shown that when subjects undergo extensive training, the prediction error signal diminishes as the behaviour becomes more habitual (Deng, Song,

Ni, Qing, & Quan, 2023). This causes a shift from goal-directed to habitual learning, where the process becomes more automated, decreasing the need for constant prediction error based adjustments (Wickens, Horvitz, Costa, & Killcross, 2007). Integrating such insights with RL models provides a deeper understanding of how neural mechanisms govern learning and decision-making processes.

2.5 Gaps in Existing Research

While there is substantial evidence linking over-training with improved cognitive flexibility, this is still an on-going topic of discussion. Existing studies often produce conflicting results regarding the impact of extended training on flexibility. While some research suggests that over-training has a positive effect on cognitive flexibility, particularly in the context of reversal learning (Sitterley & Capehart, 2014; Dhawan et al., 2019), other work has shown a potential decline in cognitive performance due to the extended training (Symons et al., 2023; Hill, Spear, & Clayton, 1962; Gabriel, Freer, & Finger, 1979). Such work demonstrates that the impact of the ORE is still inconclusive, and the effects of over-training might be context-dependent, varying across different tasks, species and environment settings.

Another significant gap lies in the application of reinforcement learning (RL) models to study over-training. Although such models have been previously used to simulate decision-making processes (Sandbrink & Summerfield, 2024; Verbeke, Pieter and Verguts, Tom, 2024), there are challenges in accurately modeling over-training within these frameworks. For instance, as over-training often leads to habitual behaviour with diminished prediction error signals (Deng et al., 2023), it is unclear how well current RL models can adequately capture the switch from goal-directed to habitual learning. This limitation suggests the need for more sophisticated RL models that can better replicate the over-training process and its effects on cognitive flexibility.

Addressing these gaps would provide a more holistic understanding of the effects of over-training on cognitive flexibility, with potential implications for optimising decision-making behaviour.

Chapter 3

Methodology

3.1 Overview of Experimental Design

To address the gap in the literature examining the effects of over-training using RL models, this study explores the mechanisms that influence adaptability through RL simulations. The design of these simulations draws heavily from the work of Maggi et al. (Maggi et al., 2024), which tracked subjects' behavioural strategies at a trial-by-trial resolution. In that study, laboratory rats performed a binary lever-press task to receive sucrose rewards for correct lever presses. The experiment analysed the rats' behavioural strategies for spatial or visual cue-based rules, and to assess cognitive flexibility, the rules were continuously switched after 10 consecutive correct trials. This allowed researchers to observe how rats adjusted their behavioural strategies in response to changing reward conditions.

Within the context of this study, a Bayesian RL model will be used to see if it can replicate the animal behaviour within normal conditions, as well as given extensive training. It will focus on the spatial component of this experiment, requiring RL agents to either choose to go left or right before receiving a reward for the correct decision. Similar to the original study, rules were reversed once agents had correctly selected the correct action 10 consecutive times. However, as an extension to the original study, a certain subset of agents underwent extensive training, to see how this impacted their ability to adapt to rule changes. These agents received an additional 100 trials after they had correctly learned the current rule.

3.2 RL Framework

3.2.1 RL Overview

RL is a branch of machine learning where agents learn to make decisions by interacting with their environment to maximise cumulative rewards. An agent observes the current state of an environment, selects an action, and receives a reward depending on its decision. The agent's objective is to learn an action policy that maximises long-term rewards.

A fundamental challenge agents face in RL is maintaining a good balance between exploration (trying new actions) and exploitation (taking actions that have previously led to high rewards) (Yogeswaran & Ponnambalam, 2012). This is known as the exploration-exploitation trade-off and can be addressed by implementing action selection functions that guide agents in deciding whether to explore or exploit. One method is using sampling algorithms, which allow agents to sample actions based on uncertainty and potential reward estimates.

Bayesian RL models use probability distributions to model the agent's uncertainty about rewards and outcomes. By updating these distributions as new data is observed, this framework provides a structured way for agents to adapt to dynamic environments. In this study, the Bayesian RL framework is applied to a cognitive flexibility task, where agents are trained to adapt to changes in reward conditions. The simulation models both normal and over-training conditions to analyse how different training strategies impact learning and adaptability.

3.2.2 Bayesian Inference

The two essential distributions that allow agents to maintain and update their belief system as they gather more data are the prior distribution and the posterior distribution. The prior distribution of a model represents the initial beliefs or assumptions about the environment. For example, an agent may have a prior belief about the probability of receiving a reward for a specific action. In this study, the agents were initialised with a uniform prior distribution, indicating that they assumed all actions initially had equal

probabilities of success. These prior beliefs are then updated by the likelihood of the observed data, which represents how probable the new data is under the current model. This update creates the posterior distribution (Gelman, 2006). Thus, the posterior distribution represents the agent's updated understanding of the environment, using the initial beliefs combined with further evidence gathered by observed outcomes (Schoot et al., 2013).

In order to maintain a probabilistic model of the agents beliefs, a Beta distribution was used as the prior distribution, being continuously updated as new data was observed. These distributions are often used for two-armed bandit tasks such as this where there are only two possible outcomes. Such experiments that only have a binary 'success' or 'failure' outcome are called Bernoulli experiments (Schneider, 2005). Beta distributions are often used for such experiments as the resulting posterior distribution is also a Beta distribution, simplifying the process of updating beliefs.

The Beta distribution is defined as:

$$\text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{3.1}$$

Where:

- x is a random variable representing a probability (ranging from 0 to 1).
- α and β are shape parameters that control the distribution's shape.
- $\Gamma(\cdot)$ is the Gamma function, which generalises the factorial function for real numbers.

The parameters α and β are often attributed to the successes and failures of an agents actions resulting in a reward. They are updated as new data is observed:

$$\alpha_{\text{new}} = \alpha_{\text{old}} + \text{successes} \tag{3.2}$$

$$\beta_{\text{new}} = \beta_{\text{old}} + \text{failures} \tag{3.3}$$

Bayes' theorem was used to update the probability estimates of possible actions based on the observed data. This theorem is mathematically expressed as:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} \tag{3.4}$$

Where:

- $P(\theta|D)$ is the posterior probability of the hypothesis θ given the data D.
- $P(D|\theta)$ is the likelihood of observing the data D given the hypothesis θ.
- $P(\theta)$ is the prior probability of the hypothesis θ before observing the data.
- $P(D)$ is the marginal likelihood, representing the total probability of observing the data under all possible hypotheses.

Within the context of this study, D represents the observed outcomes from the simulation. Furthermore, θ is the probability of receiving a reward for a specific action, which the model aims to estimate and update based on new data.

3.2.3 Sampling Algorithms

The RL models implemented in this study used two different Bayesian sampling algorithms to compare how their action selection policies influenced agent behaviour. Unlike traditional action selection policies in RL models (such as softmax or Q-learning), which rely on a single value to make decisions, Bayesian sampling methods consider the entire distribution of possible outcomes. This allows agents to take into account the variability in their decision-making process, making these methods particularly effective for handling uncertain environments. The two algorithms that were implemented in this study were

Thompson Sampling (Thompson, 1933) and Bayes Upper Confidence Bound (Bayes-UCB) (Kaufmann, Cappe, & Garivier, 2012).

The Thompson Sampling algorithm selects actions by sampling from the posterior distribution of the expected reward. It can be mathematically expressed as:

$$a = \arg\max_{a'} \left(\text{Beta}(\alpha_{a'}, \beta_{a'}) + \eta\right) \tag{3.5}$$

Where:

- a is the selected action that maximises the sampled value from the Beta distribution.
- a' represents a candidate action from the set of all possible actions.
- $\text{Beta}(\alpha_{a'}, \beta_{a'})$ is the posterior Beta distribution for action a', with parameters $\alpha_{a'}$ and $\beta_{a'}$.
- $\alpha_{a'}$ and $\beta_{a'}$ are the shape parameters of the Beta distribution for action a'.
- η is the noise level.

As can be seen in (3.5), the noise level (η) introduces variability in the action selection process by adding randomness to the sampled values from the Beta distribution. This encourages more exploration, allowing the agent to occasionally choose sub-optimal actions to discover potentially better strategies.

The Bayes-UCB sampling algorithm selects actions using the upper confidence bound of the expected reward distribution. The expected reward is typically the mean of the reward distribution for an action, calculated based on the posterior distribution. Thus, the algorithm uses the upper quantile of the posterior distribution, rather than the posterior mean (Ghavamzadeh et al., 2015). This provides an 'optimistic' approach to action selection, as the algorithm deliberately overestimates the potential reward of less certain actions or those explored less. The action selection rule can be mathematically expressed as:

$$a = \arg\max_{a'} Q_{1-\frac{1}{t}} \left(\text{Beta}(\alpha_{a'}, \beta_{a'}) + \eta\right) \tag{3.6}$$

Where:

- a is the action selected by the algorithm.
- a' represents a candidate action from the set of all possible actions.
- $\text{Beta}(\alpha_{a'}, \beta_{a'})$ is the posterior Beta distribution for action a', with parameters $\alpha_{a'}$ and $\beta_{a'}$.
- $Q_{1-\frac{1}{t}}(\text{Beta}(\alpha_{a'}, \beta_{a'}))$ is the upper quantile (at level $1 - \frac{1}{t}$) of the Beta distribution, representing the optimistic estimate of the action's expected reward (Ghavamzadeh et al., 2015).
- t is the current time step.
- $\alpha_{a'}$ and $\beta_{a'}$ are the shape parameters of the Beta distribution for action a'.
- η is the noise level.

As can be seen in (3.6), the time step t affects the level of exploration. As t increases, the quantile level $1 - \frac{1}{t}$ approaches the posterior mean, leading to increased exploitation over time (Kaufmann et al., 2012).
Unless explicitly stated, the RL simulations discussed in this study will use the Thompson sampling algorithm by default.

3.3 Simulation Setup

3.3.1 Environment and Task Design

The RL simulation in this study was implemented with Python, using key libraries such as NumPy for efficient matrix operations, Pandas for data manipulation, and Matplotlib

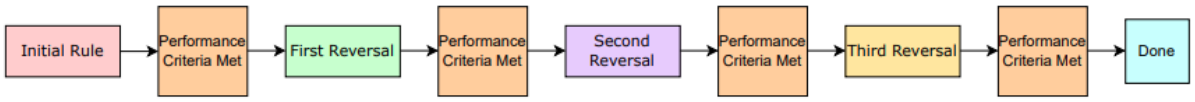
for visualisation. The environment simulated a simple two-choice task where agents interacted with states that represented binary decisions. The data and analysis code to replicate all figures can be found at the following GitHub repository: `https://github.com/salaarmir/CognitiveFlexibilityBayesianRL`.

Agents were trained on a reversal learning task, where they were required to learn a reward structure that would periodically reverse once a specific performance criterion was met. This criterion was defined as 10 consecutive trials where the correct decision was consistently made. In each state, the agent was required to choose between going left or right in a virtual decision-making task. The agent selected an action and received a reward based on whether the correct rule was followed. Upon reaching the performance criterion, the rules were reversed to assess the agent's ability to adapt to the change in conditions. The study included a control group that received only the standard training, with the rule being reversed immediately after the performance criterion was met. Additionally, a subset of agents underwent over-training, which involved an additional 100 trials under the same rule after reaching the performance criterion. The simulation featured a total of four rules, with three reversals that agents had to adapt to after learning each rule. Figure 3.1 shows a timeline of the simulation task for both the control group (Panel a) as well as the over-trained group (Panel b). A total of 50 iterations of each simulation were averaged to account for variability in the agent's learning process and ensure the reliability of the results.

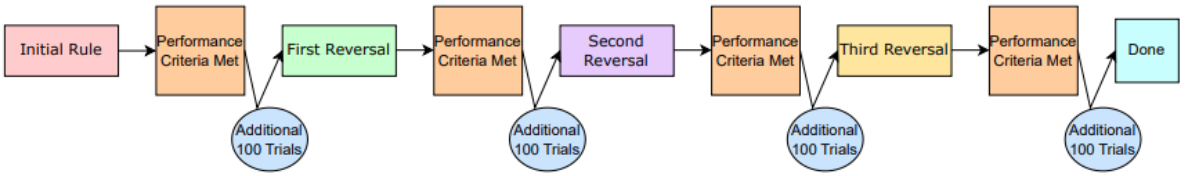
3.3.2 Parameter Settings

In addition to comparing the sampling algorithms that were implemented, certain other parameters were also varied systematically to understand their effect on replicating the empirical findings. The two key parameters examined were the decay factor and noise level.

The decay factor γ is a parameter that controls the rate at which the influence of past experiences on the agent's behaviour diminishes over time. A higher γ would result in slower adaptation to new information, as the impact of historical experiences will be



(a)



(b)

Figure 3.1: **Timeline of the simulation task.** (a) Timeline for the control group. (b) Timeline for the over-trained group. This group received an additional 100 trials of training after they had correctly learned the current rule.

preserved. This can allow for more stable learning, however, reduces the agent's ability to quickly adapt to changes. Conversely, a lower γ might result in more sporadic learning but accelerates adaptation, favoring more recent experiences compared to older ones.

The noise level η refers to the degree of variability or randomness in an agent's decision making process. A higher η introduces more randomness, causing for increased exploration and more variable behaviour patterns. This can prevent overfitting to a specific strategy and the discovery of more optimal actions. On the other hand, lower η values result in more consistent behaviour, as agents rely heavily on their learned policies.

Within the context of cognitive flexibility, decay factors are important in modelling how individuals update their beliefs in response to new information. Furthermore, noise levels can represent an individuals propensity for exploration versus exploitation in decision-making under uncertainty.

3.3.3 Data Collection

The data collection process in this study was designed to capture the key metrics related to the learning performance of the RL agents. The primary data collected included the Maximum A Posteriori (MAP) estimates of action probabilities, cumulative rewards, cumulative choice distributions, and the number of trials required for agents to learn the individual rules. These metrics were used to evaluate the effects of over-training, noise levels, and decay factors on learning adaptability in RL agents.

Calculating Maximum A Posteriori (MAP)

The MAP estimate was calculated for each action at every decision point in the simulation. The MAP estimate represents the mode of the posterior distribution, providing a point estimate of the most likely value of the parameter given the observed data. In the context of this study, the MAP is the maximum probability of choosing the selected action at any given trial.

To compute the MAP estimate, the posterior Beta distribution of each action a' was evaluated. The MAP is the mode of this posterior distribution and represents the most probable value of the action's success probability given the agent's observed data. For a Beta distribution, the MAP is calculated as:

$$\text{MAP} = \frac{\alpha - 1}{\alpha + \beta - 2} \tag{3.7}$$

Where α and β are the shape parameters of the Beta distribution for a given action.

In this study, the MAP estimate was used as a post-hoc measure to analyse the agent's decision-making process over time, rather than for selecting actions during the simulation. By computing the MAP estimate, it was possible to track how the agent's confidence in each action evolved across different training conditions, providing insight into how different factors like over-training and noise level affected the learning process.

Cumulative Reward and Cumulative Choice Distributions

The cumulative reward was calculated as the running total of rewards accumulated by the RL agent at each trial over the course of the simulation. At each trial, the reward for that specific trial was added to the sum of rewards from all previous trials. This metric provided a more detailed view of the agent's ability to maximise rewards over time under different training conditions.

The cumulative choice distributions were computed to visualise the agent's decision-making behaviour across trials. This involved recording the frequency of each action chosen by the agent over time and plotting the cumulative totals. By analysing these distributions, patterns in the agent's behaviour could be observed, providing insights into how these patterns were influenced by different variables.

Number of Trials for Learning

The number of trials required for learning was recorded for each condition tested in the simulations, allowing us to compare the learning efficiency of agents under normal and over-training conditions, as well as across different noise levels and decay factors. Tracking the number of trials required for learning provided insights into the impact of these parameters on the flexibility of the RL agents.

Chapter 4

Results

4.1 Cognitive Flexibility Under Different Training Conditions

To understand the impact of training conditions on the cognitive flexibility of RL agents, bar plots were created to compare the learning performances of different rule stages in the simulation. Figure 4.1 shows subplots of the average number of trials it took for normally trained vs over-trained agents to successfully learn the initial rule as well as the first, second and third reversal. Each plot corresponds to the learning performances for a specific rule. Results are compared across different η values for better analysis. It is evident that, while both groups were able to learn the initial rule in a comparable number of trials, over-trained agents required significantly more trials to adapt to rule reversals. Thus, there seems to be no ORE displayed within the agents in the RL simulations.

4.2 Parameter Adjustment

To better understand the effects of the parameter settings and training conditions on reversal learning performances, simulation data was processed and visualised for analysis. Different decay factors and noise levels were compared against both normally trained and over-trained agents to assess the impact on learning performances. Figure 4.2 compares

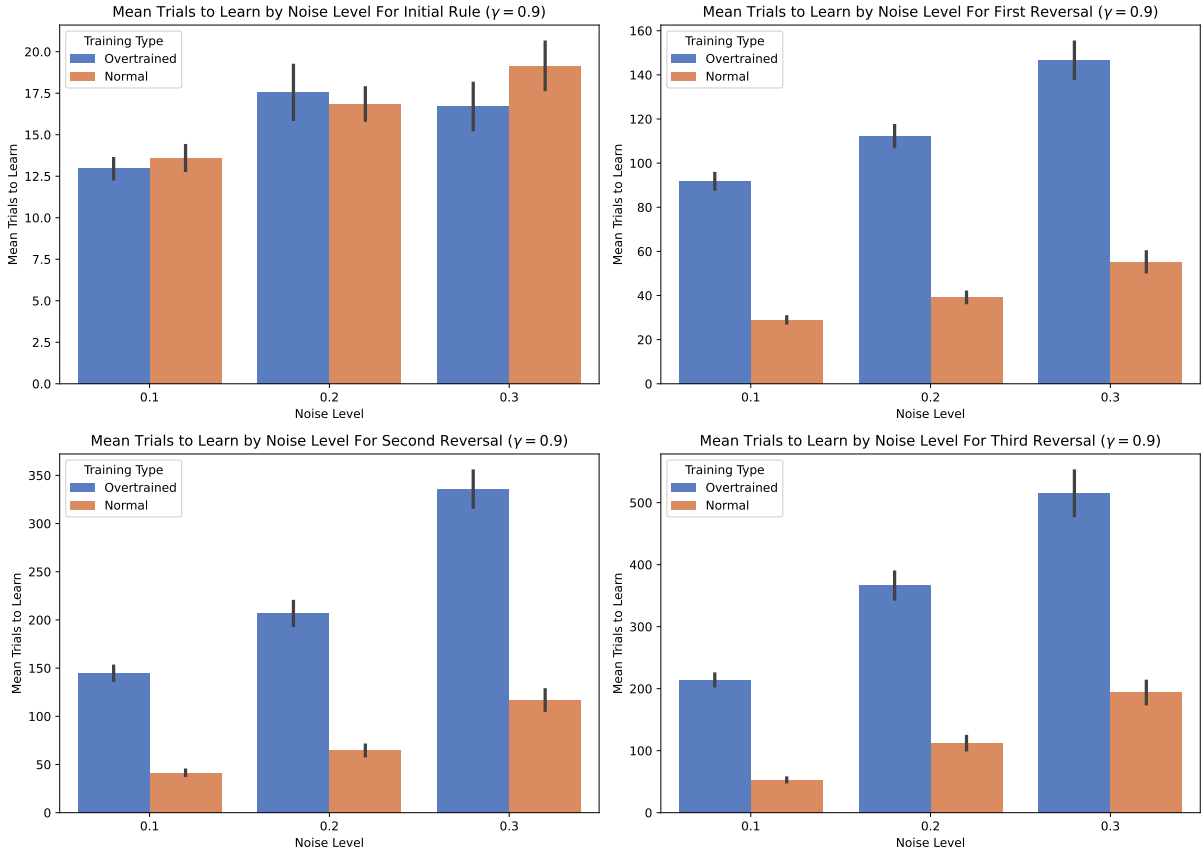


Figure 4.1: **Average learning time for initial rule and reversals under different training conditions for various η values.** Top left image: average time it took to learn the initial rule. Top right image: average learning time for first reversal. Bottom left image: average learning time for second reversal. Bottom right image: average learning time for third reversal. For all plots, γ was set to 0.9.

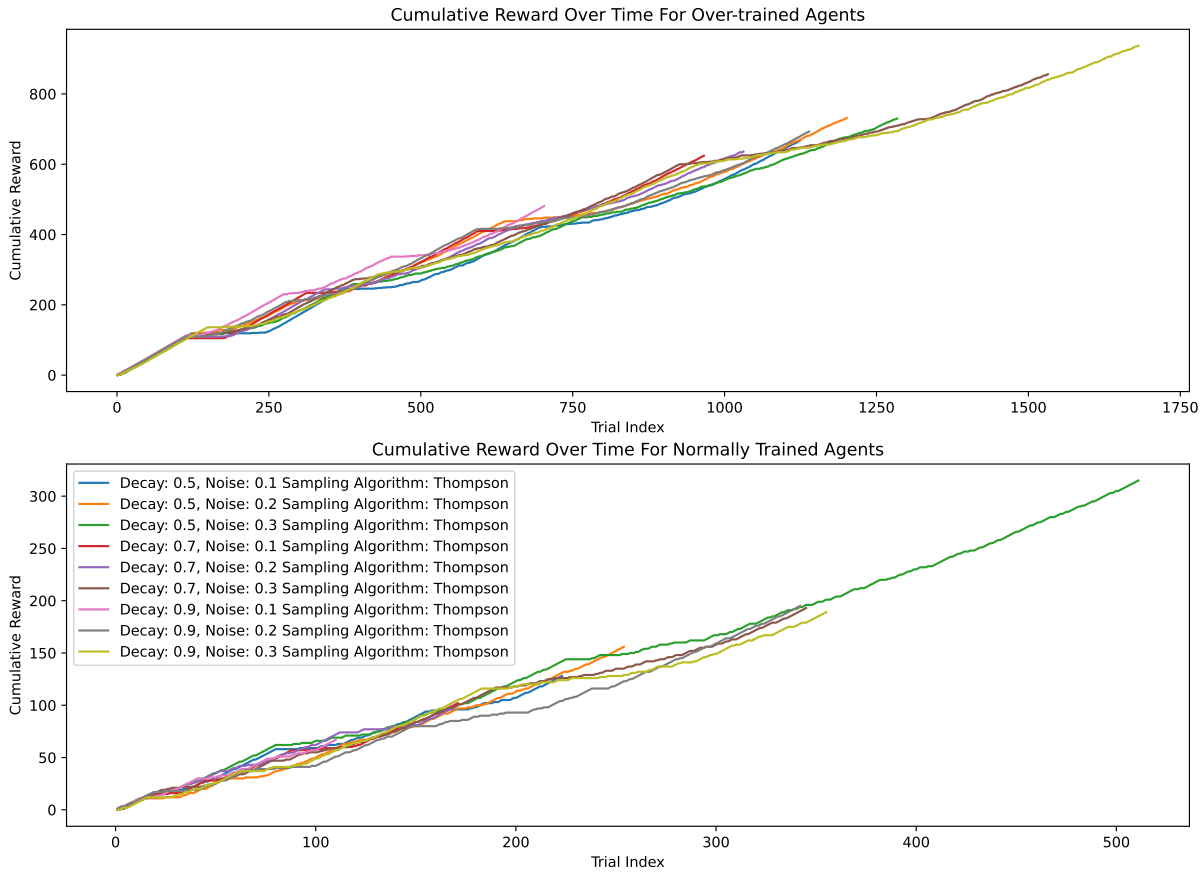


Figure 4.2: **Cumulative reward for decay factors from 0.5 to 0.9 and noise levels from 0.1 to 0.3 for both normally trained and over-trained agents.** Top image: cumulative reward for over-trained agents. Bottom image: cumulative reward for agents with normal training. Plot legend is the same across both figures.

the cumulative reward of both normally trained (bottom image) as well as over-trained (top image) agents for different decay factors and noise levels. Looking at these graphs, we can see that over-trained agents had considerably longer learning times in comparison to those that received only normal amounts of training. The fastest agent to be able to correctly learn all the rules was normally trained with $\gamma = 0.9$ and $\eta = 0.1$. This agent managed to learn all rules in 110 trials. To compare, the fastest over-trained agent learned all the rules in 603 trials (disregarding the trials of additional training). This agent also had $\gamma = 0.9$ and $\eta = 0.1$.

To further analyse the impact of decay factor γ and noise level η on reversal learning performances, additional visualisations were created to compare their effects on the agents' behaviour. Figure 4.3 displays the cumulative distribution of choices and rewards for over-trained agents across a range of γ values from 0.5 to 0.9 and η values from 0.1 to 0.3.

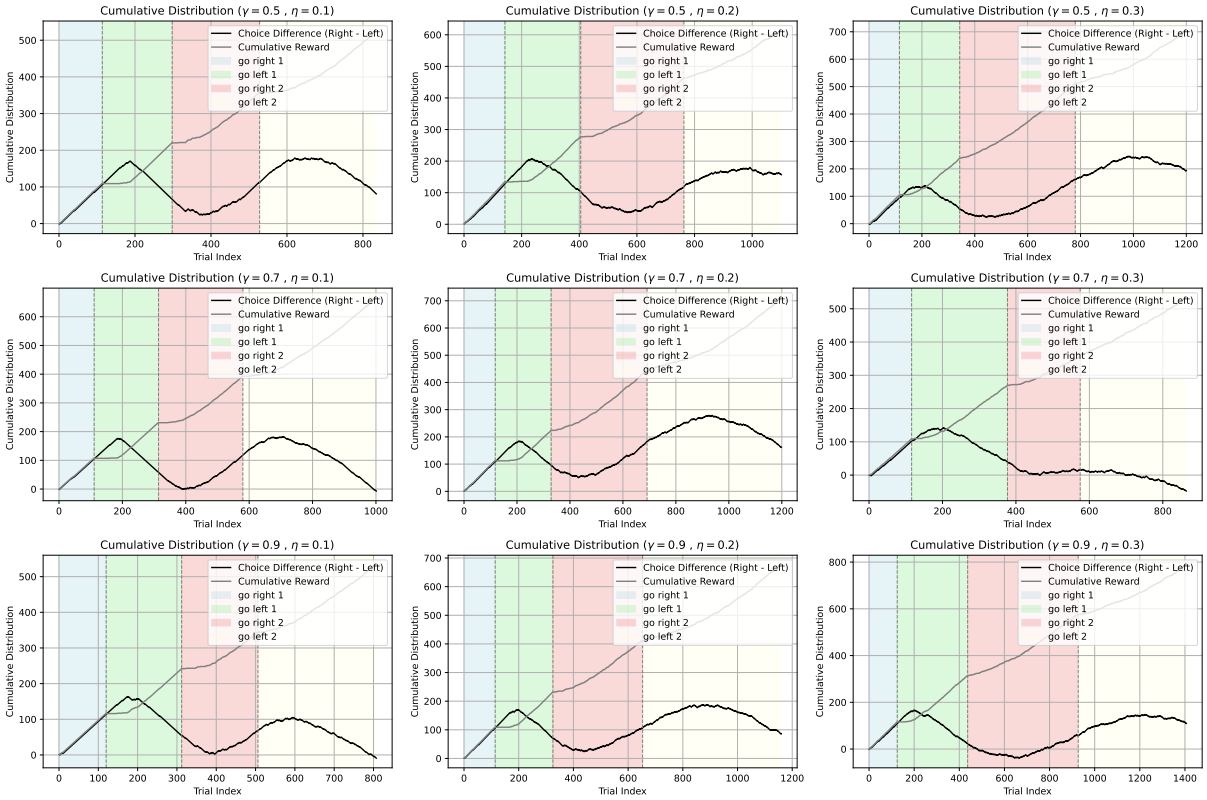


Figure 4.3: **Cumulative distribution of choices for decay factors from 0.5 to 0.9 and noise levels from 0.1 to 0.3.** The plots are organised in a grid format, with rows representing different γ values and columns representing different η values. Shaded regions represent different rule periods. Results are shown for over-trained agents.

The plots are organised in a grid format, with rows representing different γ values and columns representing different η values.

As γ increases from top to bottom, the cumulative distribution of choices (represented by the lines) appears to maintain a similar shape. This suggests that the agents' behaviour, in terms of their distribution of choices between "Go Right" and "Go Left," is relatively consistent across different decay factors, indicating that the agents' memory of past events does not significantly affect their overall choice distribution in these simulations. This could imply that for over-trained agents, the effect of γ on reversal learning is minimal, potentially due to a saturation point where agents have already optimised their strategy and additional memory retention does not lead to further improvement.

Comparing the plots horizontally, each column shows a different η value, increasing from left to right. As η increases, the distributions become more flattened. At lower noise levels ($\eta = 0.1$), the cumulative choices are more distinct, with sharper peaks indicating

more deterministic behaviour. Agents are more likely to commit to a specific strategy (either "Go Right" or "Go Left"). In contrast, higher noise levels ($\eta = 0.3$) lead to more exploration and less deterministic behaviour, as indicated by the flattened curves. This suggests that as noise increases, agents are less certain about their choices and tend to explore more, resulting in a more uniform distribution of actions over time. The increased noise introduces more variability in decision-making, making agents less likely to stick to a single choice pattern.

By examining the plots diagonally, we can analyse the combined effects of γ and η on the agents' learning behaviour. For example, in the bottom-right plot ($\gamma = 0.9$, $\eta = 0.3$), the distribution is flatter and more spread out, indicating that higher noise levels, even with a high decay factor, encourage more exploration. In contrast, lower noise levels paired with any γ value tend to produce more distinct choice distributions, showing more exploitation based on observed data.

These visualisations highlight that while the decay factor γ alone does not significantly alter the cumulative choice distribution of over-trained agents, the noise level η plays an important role in modulating the trade-off between exploration and exploitation. Higher η values promote exploration, leading to more varied choice distributions. This analysis suggests that, in these simulations, noise level is a more influential parameter than the decay factor in determining agents' behaviour.

4.3 Statistical Analysis

To further assess the statistical significance of different training conditions and parameter settings on reversal learning performance, a three-way ANOVA test was conducted. The goal of this analysis was to determine whether decay factor γ, noise level η, and over-training significantly affected the learning performance of agents, as well as to explore any potential interactions between these variables. The results, summarised in Table 4.1, indicate significant effects of both training conditions and noise levels on the learning outcomes ($p < 0.05$).

There is a significant effect of noise level on learning performance ($F = 3.363, p = 0.042$),

Source	Sum of Squares	df	F-Value	p-Value
C(Q("Decay Factor"))	15287.694	2	0.584	0.561
C(Q("Noise Level"))	88011.444	2	3.363	0.042
C(Q("Over-trained"))	217910.014	1	16.654	0.000149
C(Q("Decay Factor")):C(Q("Noise Level"))	21955.556	4	0.419	0.794
C(Q("Decay Factor")):C(Q("Over-trained"))	6350.861	2	0.243	0.785
C(Q("Noise Level")):C(Q("Over-trained"))	3192.111	2	0.122	0.885
C(Q("Decay Factor")):C(Q("Noise Level")):C(Q("Over-trained"))	10832.889	4	0.207	0.933
Residual	706559.750	54	NaN	NaN

Table 4.1: **Three-way ANOVA results for the effects of decay factor, noise level, and over-training on learning performance.**

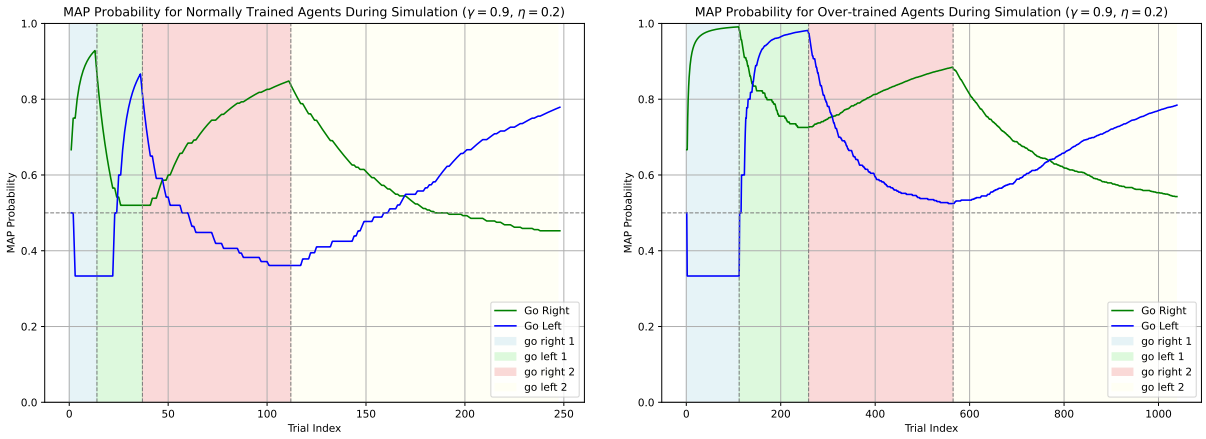
indicating that changes in noise level significantly influence the agents' ability to learn and adapt to new rules. Furthermore, a significant effect of over-training was also observed ($F = 16.654, p = 0.000149$), suggesting that additional training beyond the initial learning significantly impacts reversal learning performance.

However, The decay factor did not show a significant effect on learning performance ($p = 0.561$), nor did it significantly interact with other variables. This suggests that, within the range of γ values tested, the decay factor alone does not substantially influence the agents' learning outcomes. Moreover, there were no significant interactions between any of the factors. Specifically, the interaction between noise level and over-training ($p = 0.885$) was not significant, indicating that the effect of noise level on learning performance does not depend on whether the agent was over-trained.

4.4 Comparison of MAP Probabilities

To better understand the effects of over-training on learning and decision-making strategies, the MAP probabilities for over-trained and normally trained agents were compared over time. The MAP probabilities provide insights into how certain an agent is about the best action to take at any given trial, reflecting the learning and adaptation processes under different training conditions. Figure 4.4 presents the MAP probabilities for "Go Right" and "Go Left" actions over time for normally trained agents (Panel a) and over-trained agents (Panel b).

In the case of normally trained agents, rapid shifts in MAP probabilities corresponding to changes in task rules (denoted by the shaded background regions) can be observed.



(a) (b) .

Figure 4.4: **MAP probabilities for normally trained and over-trained agents.** (a) MAP probabilities for normally trained agents over time. (b) MAP probabilities for over-trained agents. For both simulations decay factor γ and noise level η were set to 0.9 and 0.2, respectively.

Initially, the probability of choosing "Go Right" increases sharply, indicating that the agent quickly learns the optimal action based on immediate rewards. As the task rules change, the MAP probabilities adjust rapidly, showing the agent's ability to adapt to new conditions without excessive reliance on previous training. This flexibility is reflected in the sharp transitions between high probability for "Go Right" and "Go Left," suggesting that normally trained agents maintain a balanced approach between exploration and exploitation.

For over-trained agents, the learning dynamics are notably different. The MAP probabilities for "Go Right" and "Go Left" actions show more gradual changes over time. After initially learning the optimal action, the probability of switching to the correct new action following a rule change is slower and less decisive than in normally trained agents. This behaviour suggests a stronger influence of the previous training on current decision-making, indicative of habitual responses rather than adaptive strategies. The extended flat regions and slower transitions in MAP probabilities indicate that over-trained agents have a harder time unlearning previously reinforced actions and adapting to new rules.

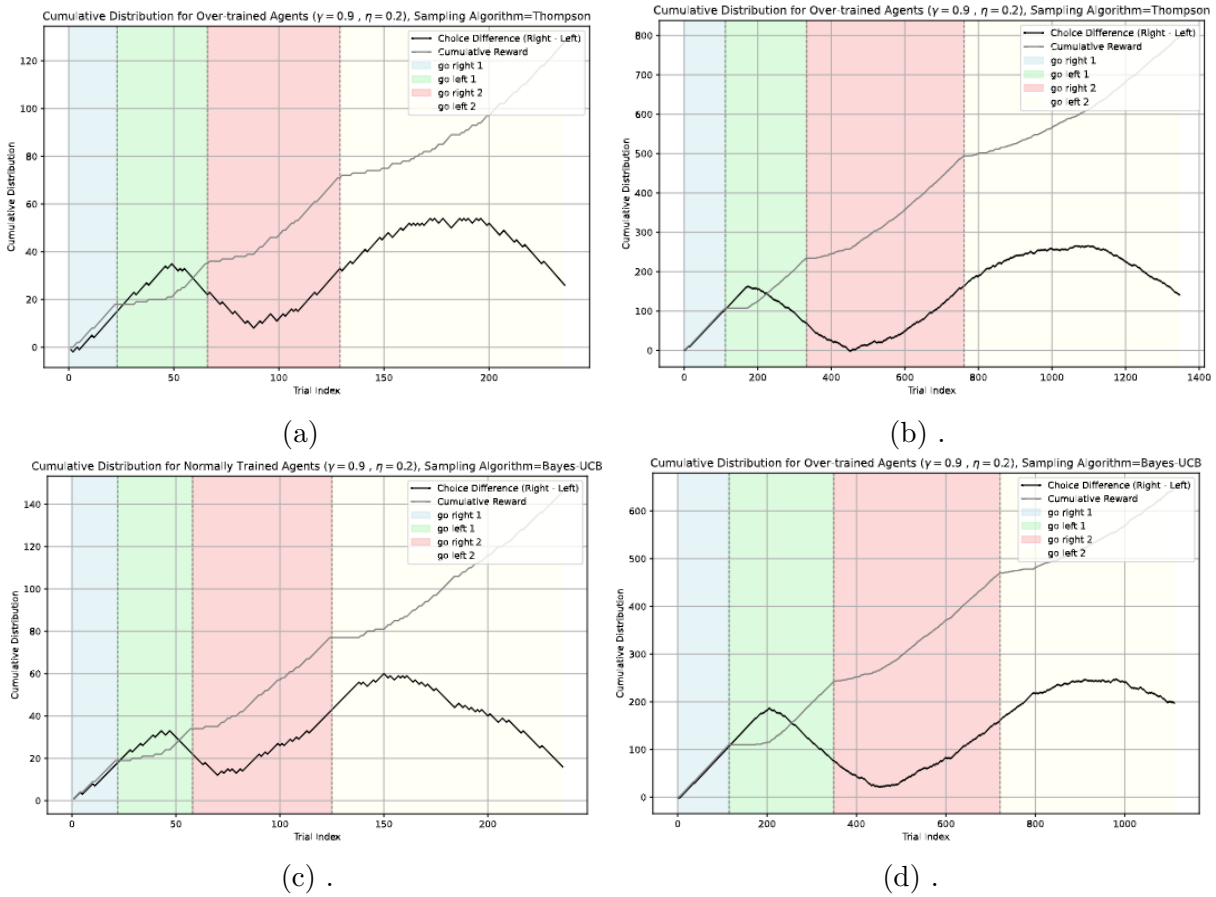


Figure 4.5: **Comparison of Thompson sampling and Bayes-UCB using cumulative choice distributions for normally trained and over-trained agents.** (a) Cumulative choice distributions using Thompson sampling for normally trained agents. (b) Cumulative choice distributions using Thompson sampling for over-trained agents. (c) Cumulative choice distributions using Bayes-UCB for normally trained agents. (d) Cumulative choice distributions using Bayes-UCB sampling for over-trained agents. For all simulations, decay factor γ and noise level η were set to 0.9 and 0.2, respectively.

4.5 Comparison of Bayesian Sampling Algorithms

Despite the Thompson sampling algorithm being used for all simulations throughout the study, the Bayes-UCB algorithm was also implemented for the purposes of comparison. Figure 4.5 compares both sampling algorithms for normally trained and over-trained agents using the cumulative choice distributions.

Both Thompson Sampling (Panels a and b) and Bayes-UCB (Panels c and d) exhibit similar cumulative choice distributions for normally trained and over-trained agents. The learning curves show comparable slopes and adaptation rates, suggesting that under the parameters used (decay factor $\gamma = 0.9$ and noise level $\eta = 0.2$), the agents' performance

is not substantially affected by the choice of sampling algorithm. This could be due to the task environment and parameter settings, which might have created conditions where the algorithms performed similarly. Despite being different algorithms, they are both designed to balance exploration and exploitation efficiently. In this study, the agents may not have faced enough variability or change in the environment to produce distinct behaviours between the two methods. These outcomes suggest that, within the context of this study, the impact of the sampling algorithm on learning performance is minimal.

Chapter 5

Discussion

5.1 Interpretation of Findings

This study investigated the effects of different training conditions and parameter settings on cognitive flexibility in RL agents. As hypothesised, The primary findings indicate that over-training significantly hinders the ability of agents to adapt to changes. This interpretation is supported by the statistically significant interaction between training condition and learning performance ($p < 0.05$), where over-trained agents showed slower adaptation to rule changes, compared to normally trained agents. This absence of the ORE in the RL simulations suggests that the benefits of over-training observed in certain empirical studies (Dhawan et al., 2019; Eimas, 1966; Nakagawa, 1992) do not translate directly to computational models.

Normally trained agents also exhibited a more dynamic adjustment in MAP probabilities, quickly adapting to changes in the task rules. Alternatively, over-trained agents showed a tendency towards slower adaptation and less flexibility in their learning behaviours. These findings align with previous studies that suggest over-training can lead to rigid, habitual behaviours, reducing cognitive flexibility (Lesage & Verguts, 2021).

Furthermore, the results demonstrated significant effects of noise level on learning outcomes, highlighting the role of exploration in reinforcement learning. Higher noise levels were associated with greater variability in agent behaviour, promoting further exploration rather than sticking with a particular action. However, increased exploration also resulted

in longer learning times. This can be attributed to the fact that agents with higher noise levels do not consistently exploit the best-known actions, instead often choosing to explore other possibilities. As a result, these agents may require more trials to converge on an optimal strategy compared to agents with lower noise levels who focus more on exploiting the optimal actions.

5.2 Implications for Cognitive Flexibility

5.2.1 Absence of Meta-learning

The results of this study show that over-trained agents require more time to learn each rule reversal, in comparison to agents trained only to criterion. As shown in Figure 4.1, over-trained agents took significantly longer to adapt to each successive rule change. In contrast, empirical data from certain biological studies have showed the opposite trend: over-trained subjects tend to become more efficient in learning new reversals, suggesting they are able to learn task structures over time (Reid, 1953; Sitterley & Capehart, 2014; Dhawan et al., 2019).

This disparity suggests that the RL model in this study fails to capture the improvement in performance often seen in humans and animals during repeated rule reversals. In these biological systems, over-training leads to an increased ability to learn new rules faster, as subjects develop an understanding of the task structure through experience. This ability to "learn how to learn", known as meta-learning, enables individuals to refine their learning processes based on prior experiences (Huisman, van Rijn, & Plaat, 2021; Vilalta, Giraud-Carrier, & Brazdil, 2010). This helps them anticipate changes and adjust more quickly when new rules are introduced. Unlike humans and animals, the RL agents in this study lacked this capability, as they were unable to effectively adapt their learning strategies based on past experience.

The absence of meta-learning in RL agents may explain why they do not exhibit the same improvements over time as observed in empirical data. While RL agents are effective at learning individual tasks, they are limited in their ability to generalise from

past experiences to improve performance in new conditions. This limitation suggests that RL models, as implemented in this study, do not fully replicate the adaptability inherent in biological learning systems. Incorporating meta-learning capabilities into RL models could potentially improve their performance, allowing them to better mimic the cognitive flexibility observed in humans and animals.

5.2.2 Habitual Behaviour

Furthermore, the slower adaptation to new rules exhibited by the over-trained agents suggests a reliance on more rigid, habitual learning. Contrarily, normally trained agents displayed a more flexible approach, adapting quickly to changes and showing a more dynamic adjustment in MAP probabilities.

The Beta distribution seems to have played a significant role in shaping these differences in behaviour across training conditions. This distribution models the success probability for a given action, using the counts of successes and failures to update its parameters (α and β). Over-training increases the number of samples (i.e., trials) experienced by an RL agent, reinforcing its distribution parameters. As over-trained agents accumulate more data, their Beta distributions become more peaked, with higher concentrations around the most frequently observed actions. The extended number of trials causes agents to have more certainty in their action-selection policies, reducing the variability in their behaviour and leading to more habitual responses. In comparison, normally trained agents have fewer samples and a less peaked Beta distribution. This allows them to be more flexible in their action choices, resulting in more rapid adaptation to rule changes.

The reinforcement learning mechanism caused through the Beta distribution mirrors the process of habit formation in biological systems, where consistent reinforcement of specific behaviours strengthens neural pathways (Smith & Graybiel, 2016). The continuous action repetition makes these behaviours more automatic and less sensitive to changes in conditions. Further work done by Luque et al. involving human participants also produced similar results, showing that over-training causes a switch from goal-directed to habitual behaviour (Luque, Molinero, Watson, López, & Pelley, 2019). Thus, as hypothesised, this

study showed that over-training not only leads to a decrease in cognitive flexibility of RL agents, but also simulates the development of habitual learning through the statistical properties of the Beta distribution.

5.3 Limitations

While this study provides valuable insights into the effects of over-training and parameter settings on cognitive flexibility in RL agents, several limitations must be acknowledged. First, as mentioned previously, the RL models used in this study do not incorporate meta-learning capabilities. This may limit their ability to adapt to new tasks or changes in the environment. This lack of meta-learning likely contributed to the absence of the ORE in the simulations, as the agents were unable to adjust their learning strategies based on accumulated experience.

Furthermore, the experimental conditions tested in this study, including the range of decay factors and noise levels, may not fully capture the complexity of real-world learning scenarios. While these parameters provide a useful framework for analysing cognitive flexibility, they may not incorporate the full range of factors that influence learning and adaptability in biological systems.

Finally, the generalisability of the findings may be limited by the specific tasks and environments used in the simulations. The simulations created for this study were simple binary decision tasks in which agents were required to make one of two possible choices. The results may not necessarily apply to other types of tasks or more complex decision-making environments, where different learning dynamics might be involved.

5.4 Future Directions

The current study provides an initial exploration of cognitive flexibility and the impact of over-training on learning adaptability using Bayesian RL models. However, there are several avenues for future research that could enhance the understanding of cognitive flexibility and improve the performance of RL models in mimicking human and animal

learning processes.

Further work could incorporate better meta-learning capabilities into the RL model, allowing agents to more effectively adjust their learning strategies based on experience. Significant work has already been done exploring how such models can be developed to more accurately replicate biological systems (Wang et al., 2017; Nagabandi et al., 2019). Future models could explore the development of algorithms that allow agents to recognise patterns in task structures, thereby accelerating the learning process over multiple reversals.

Another important consideration is the dynamic adjustment of RL parameters, such as noise levels and decay factors, based on the agent's performance. Recent studies have shown that adaptive parameter tuning can lead to improved long-term performance in uncertain environments (Kim, Kim, Choi, & Park, 2022; Xu, Honda, & Sugiyama, 2018). Such adaptive learning strategies could allow RL agents to automatically fine-tune their exploration-exploitation balance as the task evolves, further improving their flexibility.

Moreover, future simulations could implement more complex tasks involving multiple rules or environmental changes for a better approximation of real-word scenarios. These tasks could include stochastic reversal tasks, where rewards are provided probabilistically (for example, only 75% of the time). The unpredictability of rewards could force RL agents to adjust their strategies more frequently, potentially allowing for faster adaptation. Some work has already been done using RL models to interpret human behaviour in stochastic reversal tasks (Eckstein, Master, Dahl, Wilbrecht, & Collins, 2022). It would be interesting to explore whether the added uncertainty of stochastic rewards might enhance the ORE.

5.5 Conclusion

In summary, this study provides valuable insights into the role of over-training on cognitive flexibility in Bayesian RL models, highlighting both the strengths and limitations of the current approach. While the RL agents demonstrated clear learning patterns, the absence of meta-learning and the development of habitual behaviour highlight the need for more

sophisticated models. The results indicate that the models developed in this study were unable to replicate the ORE observed in other empirical findings. Future work should focus on enhancing model flexibility through the incorporation of stochastic elements, adaptive learning strategies, and implementation of meta-learning algorithms to better simulate the ability to adapt learning strategies based on prior experience. Addressing these problems will further bridge the gap between biological and computational models, offering deeper insights into cognitive processes.

References

Babayan, B., Uchida, N., & Gershman, S. (2018, 05). Belief state representation in the dopamine system. *Nature Communications*, *9*. doi: 10.1038/s41467-018-04397-0

Boulougouris, V., & Robbins, T. W. (2010). Enhancement of spatial reversal learning by 5-ht2c receptor antagonism is neuroanatomically specific. *Journal of Neuroscience*, *30*(3), 930–938. Retrieved from `https://www.jneurosci.org/content/30/3/930` doi: 10.1523/JNEUROSCI.4312-09.2010

Cañas, J. (2006, 03). Cognitive flexibility. In (p. 297-300). doi: 10.13140/2.1.4439.6326

Colman, A. (2015). *A dictionary of psychology*.

Costa, V. D., Tran, V. L., Turchi, J., & Averbeck, B. B. (2015). Reversal learning and dopamine: A bayesian perspective. *Journal of Neuroscience*, *35*(6), 2407–2416. Retrieved from `https://www.jneurosci.org/content/35/6/2407` doi: 10.1523/JNEUROSCI.1989-14.2015

Dajani, D. R., & Uddin, L. Q. (2015). Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends in neurosciences (Regular ed.)*, *38*(9), 571 - 578. doi: 10.1016/j.tins.2015.07.003

Daw, N., & Tobler, P. (2014, 01). Value learning through reinforcement: The basics of dopamine and reinforcement learning. *Neuroeconomics: Decision Making and The Brain*, 283-298.

Deng, Y., Song, D., Ni, J., Qing, H., & Quan, Z. (2023). Reward prediction error in learning-related behaviors. *Frontiers in Neuroscience*, *17*. doi: 10.3389/fnins.2023.1171612

Dhawan, S. S., Tait, D. S., & Brown, V. J. (2019). More rapid reversal learning following

overtraining in the rat is evidence that behavioural and cognitive flexibility are dissociable. *Behavioural Brain Research*, *363*, 45-52. Retrieved from https://www.sciencedirect.com/science/article/pii/S0166432818315687 doi: https://doi.org/10.1016/j.bbr.2019.01.055

Diamond, A. (2013). Executive functions [Journal Article]. *Annual Review of Psychology*, *64*(Volume 64, 2013), 135-168. doi: https://doi.org/10.1146/annurev-psych-113011-143750

Dickinson, A. (1985, 02). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London*, *B 308*, 67-78. doi: 10.1098/rstb.1985.0010

Eckstein, M. K., Master, S. L., Dahl, R. E., Wilbrecht, L., & Collins, A. G. (2022). Reinforcement learning and bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal. *Developmental Cognitive Neuroscience*, *55*, 101106. Retrieved from https://www.sciencedirect.com/science/article/pii/S1878929322000494 doi: https://doi.org/10.1016/j.dcn.2022.101106

Eimas, P. D. (1966). Effects of overtraining, irrelevant stimuli, and training task on reversal discrimination learning in children. *Journal of Experimental Child Psychology*, *3*(4), 315-323. Retrieved from https://www.sciencedirect.com/science/article/pii/0022096566900750 doi: https://doi.org/10.1016/0022-0965(66)90075-0

Gabriel, S. M., Freer, B., & Finger, S. (1979). Brain damage and the overlearning reversal effect. *Psychobiology*, *7*, 327-332. Retrieved from https://api.semanticscholar.org/CorpusID:54016642

Gelman, A. (2006). Prior distribution. In *Encyclopedia of environmetrics.* John Wiley Sons, Ltd. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470057339.vap039 doi: https://doi.org/10.1002/9780470057339.vap039

Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2015). Bayesian reinforcement

learning: A survey. *Foundations and Trends® in Machine Learning*, *8*(5-6), 359-483. Retrieved from http://dx.doi.org/10.1561/2200000049 doi: 10.1561/2200000049

Hill, W. F., Spear, N. E., & Clayton, K. N. (1962). T maze reversal learning after several different overtraining procedures. *Journal of Experimental Psychology*, *64*(5), 533. doi: 10.1037/h0045869

Huisman, M., van Rijn, J., & Plaat, A. (2021, 08). A survey of deep meta-learning. *Artificial Intelligence Review*, *54*. doi: 10.1007/s10462-021-10004-4

Izquierdo, A., Brigman, J., Radke, A., Rudebeck, P., & Holmes, A. (2016, 03). The neural basis of reversal learning: An updated perspective. *Neuroscience*, *345*. doi: 10.1016/j.neuroscience.2016.03.021

Kaufmann, E., Cappe, O., & Garivier, A. (2012, 21–23 Apr). On bayesian upper confidence bounds for bandit problems. In N. D. Lawrence & M. Girolami (Eds.), *Proceedings of the fifteenth international conference on artificial intelligence and statistics* (Vol. 22, pp. 592–600). La Palma, Canary Islands: PMLR. Retrieved from https://proceedings.mlr.press/v22/kaufmann12.html

Kim, M., Kim, J.-S., Choi, M.-S., & Park, J.-H. (2022, 09). Adaptive discount factor for deep reinforcement learning in continuing tasks with uncertainty. *Sensors*, *22*, 7266. doi: 10.3390/s22197266

Kreher, J., & Schwartz, J. (2012, 03). Overtraining syndrome: A practical guide. *Sports health*, *4*, 128-38. doi: 10.1177/1941738111434406

Lesage, E., & Verguts, T. (2021, 10). Contextual overtraining accelerates habit formation in new stimuli.
doi: 10.31234/osf.io/7m6bh

Luque, D., Molinero, S., Watson, P., López, F., & Pelley, M. (2019, 11). Measuring habit formation through goal-directed response switching. *Journal of Experimental Psychology: General*, *149*. doi: 10.1037/xge0000722

Maggi, S., Hock, R. M., O'Neill, M., Buckley, M., Moran, P. M., Bast, T., ... Humphries, M. D. (2024, mar). Tracking subjects' strategies in behavioural choice experiments

at trial resolution. *eLife*, *13*, e86491. Retrieved from https://doi.org/10.7554/eLife.86491 doi: 10.7554/eLife.86491

Nagabandi, A., Clavera, I., Liu, S., Fearing, R. S., Abbeel, P., Levine, S., & Finn, C. (2019). *Learning to adapt in dynamic, real-world environments through meta-reinforcement learning.* Retrieved from https://arxiv.org/abs/1803.11347

Nakagawa, E. (1992). Effects of overtraining on reversal learning by rats in concurrent and single discriminations. *The Quarterly Journal of Experimental Psychology Section B*, *44*(1), 37–56. Retrieved from https://www.tandfonline.com/doi/abs/10.1080/02724999208250601 doi: 10.1080/02724999208250601

Qu, A., Tai, L.-H., Hall, C., Tu, E., Eckstein, M., Mischanchuk, K., ... Wilbrecht, L. (2023, 11). *Nucleus accumbens dopamine release reflects bayesian inference during instrumental learning.* doi: 10.1101/2023.11.10.566306

Reid, L. S. (1953). The development of noncontinuity behavior through continuity learning. *Journal of experimental psychology*, *46 2*, 107-12. Retrieved from https://api.semanticscholar.org/CorpusID:34977738

Sandbrink, K., & Summerfield, C. (2024). Modelling cognitive flexibility with deep neural networks. *Current Opinion in Behavioral Sciences*, *57*, 101361. Retrieved from https://www.sciencedirect.com/science/article/pii/S2352154624000123 doi: https://doi.org/10.1016/j.cobeha.2024.101361

Schmidt, J. R., De Houwer, J., & Moors, A. (2020, 04). Learning Habits: Does Overtraining Lead to Resistance to New Learning? *Collabra: Psychology*, *6*(1), 21. doi: 10.1525/collabra.320

Schneider, I. (2005, 12). Jakob bernoulli, ars conjectandi (1713). In (p. 88-104). doi: 10.1016/B978-044450871-3/50087-5

Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J., Neyer, F., & Aken, M. (2013, 10). A gentle introduction to bayesian analysis: Applications to developmental research. *Child development*, *85*. doi: 10.1111/cdev.12169

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1-27. Retrieved from https://doi.org/10.1152/jn.1998.80.1.1

(PMID: 9658025) doi: 10.1152/jn.1998.80.1.1

Sitterley, T., & Capehart, J. (2014, 08). Human successive discrimination reversal: Effects of overtraining and reinforcement. *Psychonomic Science*, *4*, 293-294. doi: 10.3758/BF03342302

Smith, K., & Graybiel, A. (2016, 03). Habit formation. *Dialogues in Clinical Neuroscience*, *18*, 33-43. doi: 10.31887/DCNS.2016.18.1/ksmith

Sutton, R., & Barto, A. (1999, 01). Reinforcement learning. *Journal of Cognitive Neuroscience*, *11*, 126-134.

Symons, I., Bruce, L., & Main, L. (2023, 08). Impact of overtraining on cognitive function in endurance athletes: A systematic review. *Sports Medicine - Open*, *9*. doi: 10.1186/s40798-023-00614-3

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*, 285-294. Retrieved from `https://api.semanticscholar.org/CorpusID:120462794`

Uddin, L. (2021, 02). Cognitive and behavioural flexibility: neural mechanisms and clinical considerations. *Nature Reviews Neuroscience*, *22*. doi: 10.1038/s41583-021-00428-w

Uhl, C. N. (1964). Effects of overtraining on reversal and nonreversal discrimination shifts in a free operant situation. *Perceptual and Motor Skills*, *19*(3), 927-934. (PMID: 14238242) doi: 10.2466/pms.1964.19.3.927

Verbeke, Pieter and Verguts, Tom. (2024). Reinforcement learning and meta-decision-making. *CURRENT OPINION IN BEHAVIORAL SCIENCES*, *57*, 6. Retrieved from `{http://doi.org/10.1016/j.cobeha.2024.101374}`

Vilalta, R., Giraud-Carrier, C., & Brazdil, P. (2010, 07). Meta-learning - concepts and techniques. In (p. 717-731). doi: 10.1007/978-0-387-09823-4_36

von Bastian, C., & Druey, M. (2017, 06). Shifting between mental sets: An individual differences approach to commonalities and differences of task switching components. *Journal of Experimental Psychology: General*, *146*, 1266-1285. doi: 10.1037/xge0000333

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... Botvinick, M. (2017). *Learning to reinforcement learn.* Retrieved from https://arxiv.org/abs/1611.05763

Wanjerkhede, S., Surampudi, B., & Mytri, V. (2014, 08). Reinforcement learning and dopamine in the striatum: A modeling perspective. *Neurocomputing*, *138*, 27–40. doi: 10.1016/j.neucom.2013.02.061

Wickens, J. R., Horvitz, J. C., Costa, R. M., & Killcross, S. (2007). Dopaminergic mechanisms in actions and habits. *Journal of Neuroscience*, *27*(31), 8181–8183. Retrieved from https://www.jneurosci.org/content/27/31/8181 doi: 10.1523/JNEUROSCI.1671-07.2007

Wood, W., & Rünger, D. (2016, 01). Psychology of habit. *Annual Review of Psychology*, *67*, 289-314. doi: 10.1146/annurev-psych-122414-033417

Xu, L., Honda, J., & Sugiyama, M. (2018, 09–11 Apr). A fully adaptive algorithm for pure exploration in linear bandits. In A. Storkey & F. Perez-Cruz (Eds.), *Proceedings of the twenty-first international conference on artificial intelligence and statistics* (Vol. 84, pp. 843–851). PMLR. Retrieved from https://proceedings.mlr.press/v84/xu18d.html

Yogeswaran, M., & Ponnambalam, S. (2012). Reinforcement learning: Exploration–exploitation dilemma in multi-agent foraging task. *Opsearch*, *49*, 223–236.

Young, C., & Sonne, J. (2018, 12). Neuroanatomy, basal ganglia. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK537141/