Universiteit Leiden

Simulating Depression with Reinforcement Learning

Noam Dunsky (s4009495)

Supervisors:
Joost Broekens & Steven Miletić

Bachelor Thesis – Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 8, 2026

Abstract

Depression is a mood disorder affecting attention, judgment, perception and memory, and uncovering its mechanisms is an active field of research. One way to gain insight into the mechanism underlying depression is reinforcement learning (RL), a paradigm in which agents learn by trial and error to make optimal choices. RL can be used to study concepts like cognition, behaviour and affect, and prior work has reproduced the choice behaviour of depressed patients versus healthy controls in clinical studies. It employed an asymmetric learning rate model between the two groups, but it is unclear whether asymmetric learning is the only RL mechanism that can account for this behaviour, or whether alternative mechanisms fit equally well or better. This thesis extends this research by implementing models that vary action selection, learning rate and reward perception, inspired by psychological theories of depression. The results show that prediction-error-dependent and regret-dependent learning models offer alternative methods and interpretations for replicating the human data. The conclusion proposed that these mechanisms motivate future experimental research, such as multi-armed bandit tasks and tasks with non-binary or delayed rewards, to further discriminate between the models.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Joost Broekens and Steven Miletić, for their guidance, feedback, patience and support during the thesis process. I had the opportunity to learn from two enthusiastic and professional researchers, who inspired me and enabled me to explore the field of cognitive modelling, which I will continue to explore further in my master's. I am grateful for the opportunity to work on the intersection of two fields, my biggest interests: psychology and computer science. Thank you for making this stepping stone in my academic career enjoyable and valuable.

Contents

1 Introduction

1.1 Motivation

Depression is a psychiatric condition characterised by affective, motivational, and cognitive symptoms, including impairments in decision-making and learning [1]. A central symptom of depression is anhedonia [2], defined as the diminished ability to experience pleasure from positive stimuli. This connection between rewards, decision-making, and learning can be modelled and studied purely computationally, for example, through reinforcement learning (RL).

In RL, reward mechanisms are a central component to guide behaviour. An agent making decisions learns an optimal policy to maximise expected reward [3]. In this sense, RL provides an intuitive computational framework to investigate decision-making in depression. During simulated experiments, components of the framework, such as action selection and value calculation, can be tuned to observe changes in learning. In order to study learning and decision-making under uncertainty, experimental tasks such as probabilistic multi-armed bandit tasks are becoming increasingly popular [4]. In these tasks, participants repeatedly choose between options with unknown reward probabilities and must learn from trial and error. Such tasks allow researchers to develop and fit RL models to confirm or disprove theories, as well as provide insight into participants' learning processes during the task.

So far, empirical findings on RL performance in depressed individuals have been inconsistent. Some studies report memory sensitivity to negative prediction errors [5], whereas others find differences in emotional rather than reward processing [6]. It has also been shown that depressed individuals have higher variability of action selection, which is characterised by difficulty choosing actions [7]. The inconsistencies between individual studies are explained by Vandendriessche et al [8], who demonstrate that learning impairments in depression are context-dependent. Their paper describes a task performed in two different contexts. A "rich" context, in which positive rewards are more probable, and a "poor" context, in which they are less probable. Specifically, depressed participants show worse learning performance in the poor context, which the authors attribute to a negativity bias. Their computational model explains this effect by applying differences in learning rates between the "healthy" and "depressed" simulated agents. They suggest that depressed individuals have a negativity bias, which means that they learn more slowly in a reward-poor context and have higher learning rates for negative outcomes, whereas the "healthy" control subjects showed the opposite tendency. The negativity bias in reward processing is one possible mechanism that can produce this behavioural pattern. However, [8] do not explore other explanations in terms of RL mechanisms, such as differences in exploration-exploitation trade-off, which might also explain the experimental findings. This thesis will show that the same data may also be explained by other computational models.

Computational models of emotion and decision-making already exist. Broekens [9] argues that such models developed by computer scientists should be used by psychologists to enrich theories and explanations in affective science, and examples of modelling emotions within an RL framework can be found in [10, 11]. However, these latter models have not yet been used to replicate the observed experimental data. By reproducing the behaviour observed in depressed individuals using

a range of alternative RL mechanisms, the thesis offers a bridge between these models and human data. Testing whether the context-dependent learning bias of [8] is uniquely explained by altered learning rates or arises from other mechanisms may provide additional insight into the underlying mechanisms of depression.

1.2 Research Question and Outline

This thesis aims to reproduce the results of human participants in a computational RL setup. In particular, it will investigate whether the behavioral results reported in [8], interpreted as arising from altered learning rates, can also be reproduced using alternative RL mechanisms.

Research Question
Are the context-dependent learning biases in depression observed by [8] uniquely explained by altered learning rates, or can they emerge from alternative reinforcement learning mechanisms?

Sub-question
How might alternative computational mechanisms relate to plausible cognitive or neural processes associated with depression?

1.3 Thesis Overview

The thesis consists of the following sections:

- Section 2 provides the necessary background and related work;
- Section 3 connects the psychological theory to implications for reinforcement learning;
- Section 4 introduces the baseline model upon which this thesis builds;
- Section 5 discusses the methodology;
- Section 6 describes the experiments and their outcomes;
- Section 7 discusses limitations and future work;
- Section 8 concludes the thesis;
- Appendix A contains mathematical equivalence proof;
- Appendices B and C contain supplemental experiment data for the simple beta model.

The thesis starts with the background necessary to connect the psychological phenomenon of depression to its computational framework, starting with the disorder itself.

2 Background and Related Work

2.1 Depression

Major Depressive Disorder (MDD) is a mood disorder that affects attention, judgment, perception, and memory [12]. According to the Diagnostic and Statistical Manual of Mental Disorders [2], MDD is diagnosed in the case of having at least 5 symptoms of a defined list of 9 symptoms. In this section, some of the main symptoms and relevant psychiatric and psychological theories are reviewed: anhedonia, learned helplessness, negative beliefs and mood congruency. These inspired the computational models discussed later on in the thesis.

2.1.1 Anhedonia

Anhedonia is the reduced pleasure from positive stimuli [2, 13], and it is directly related to reduced motivation and anticipation in social interactions. This is one of the main symptoms of depression [14]

When diving deeper into the brain's reward-processing mechanisms, there are two distinct aspects, each associated with its own neural circuits: "liking" and "wanting" [15]. The "liking" aspect refers to reward perception, how a reward is experienced in the moment. Translated into human experience, it is tasting ice cream and realising that it is enjoyable. This aspect is related to the opioid system, specifically the Ventral Pallidum (VP), Ventromedial Prefrontal Cortex (vmPFC), OFC (Orbitofrontal Cortex), Anterior Cingulate Cortex (ACC) and the amygdala [16]. In contrast, the "wanting" aspect relates to motivation. Continuing with the analogy, it is the craving for ice cream that eventually causes you to leave the house and get it. This aspect is connected to dopamine.

Based on these two aspects, anhedonia can be decomposed into three types: consummatory, motivational and decisional [16]. Consummatory anhedonia is the inability to experience pleasure in the moment itself, i.e. tasting the ice cream but not enjoying it. Motivational anhedonia is the absence of anticipation of a positive stimulus, which is knowing ice cream makes you happy, but lacking the drive to go and get it. Decisional anhedonia is the inability to correctly weigh the costs and benefits of a reward, not recognising that a five-minute walk and three euros are worth the tasty treat. The different sub-components of reward processing are illustrated in Figure 1. These distinctions between perceiving, anticipating, and acting on rewards, map naturally onto distinct components of a reinforcement learning agent, as developed in Section 3.

2.1.2 Learned Helplessness

The learned helplessness hypothesis can be illustrated by an example. A small elephant is tied to an anchor in the ground, preventing it from escaping. It tries again and again, but the rope is strongly mounted to the ground. Later, as the elephant grows, he becomes much stronger and can easily pull the anchor out, allowing it to be released. But because it was trying so many times and failed, it quit trying. This is called "learned helplessness", a term introduced by Seligman [18]- failing to try after repeated attempts that end in failure. It was first observed in animal experiments, in which dogs stopped trying to escape after receiving sufficient electrical shocks. In his paper, Seligman proposed learned helplessness as a model for understanding depression, as both share

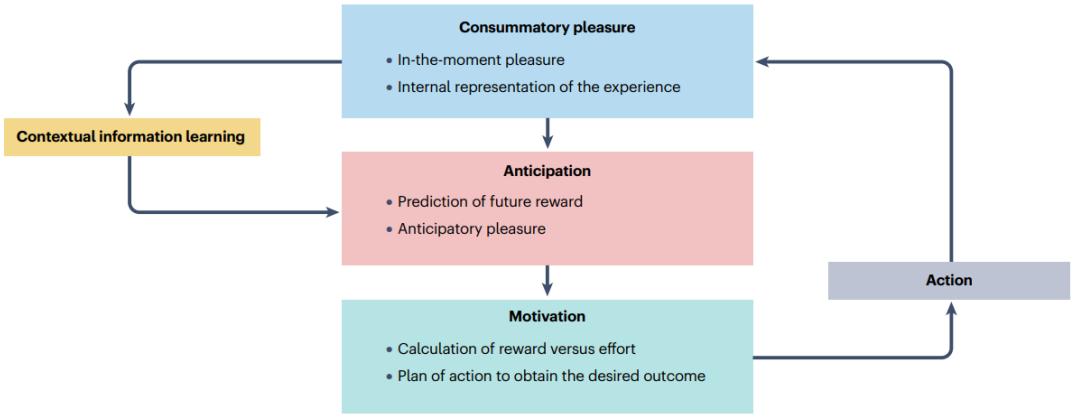

Figure 1: Sub-components of the reward processing cycle, adapted from [17]. Consummatory pleasure (blue) forms the value representation of a stimulus; anticipation (pink) predicts future reward; the transition between the consummatory pleasure and the anticipation is the learning process of the value of an action (yellow); the motivation component (green) weighs reward against effort before action (grey).

the same symptoms of passiveness and negative thoughts of one's capabilities. This is called "the learned helplessness model of depression". In his paper, [19], Miller supports this hypothesis, this time with a group of students. Later on, [20] discusses the neural insights and suggests that it might contribute to developing a treatment for depression. In computational terms, "giving up" after repeated failures suggests a shift in how an agent balances exploration and exploitation. This will be elaborated in the experiments section (Section 6).

2.1.3 Negative Beliefs, Mood Congruency

An influential model of depression is Beck's cognitive model of depression [21]. His model focuses on cognitive aspects of depression, making it relevant for the computational models discussed later in the thesis, as it highlights the differences in information processing between the healthy and depressed populations. In his model, he described the thinking patterns of depressed people and introduced the core concept of the cognitive triad: depressed people have a negative view about the experience (past and present), about the future, and about the self, which are the causes for other symptoms. This causes a bias in perception (filtering out the positive experience), in memory (encoding and retrieval), and in the meaning attributed to the situation. This concept is called "mood congruency". As a consequence, a depressed person who is constantly in a bad mood filters, remembers and thinks about information in a way that is congruent with their mood. Such a bias in perception and memory can be represented computationally either as a distortion in how reward is perceived or as a negative prior belief about the value of actions.

2.1.4 Depressive Realism

Depressive Realism is a theory introduced by Alloy and Abramson [22]. They explain the theory using a task that measures the contingency recognised by participants, which is the link between the action and its outcome. High contingency means that the outcome is highly dependent on the

action. In contrast, low contingency means the outcome is largely unrelated to the actions. In their experiments, students were asked to determine the contingency between pressing a button and the appearance of a light. Both a group of depressed subjects and a group of non-depressed subjects were able to report a contingency when it was there, but the depressed group was able to report the lack of contingency more accurately. The authors refer to the "Learned Helplessness" model of depression discussed above, and claim that this goes hand in hand with the idea that depressed people are better at recognising a lack of contingency compared to non-depressed people. Namely, they understand when their actions have no impact on the outcome. The authors suggest that non-depressed people have a positivity bias, while depressed people perceive reality more accurately. Framed in computational terms, this improvement in accuracy points to the exploration-exploitation balance, linking it back to learned helplessness. More information on the modern view of this theory can be found in [23].

Taken together, these symptoms and theories describe depression at the level of perception, motivation, belief, and action, the same components that define a learning agent. The rest of this chapter places this in a reinforcement learning framework. Section 2.2 introduces the framework, and Section 3 connects each of the symptoms and theories above to it.

2.2 Reinforcement Learning

In this section, we will discuss Reinforcement Learning as a computational framework. This includes the motivation for using this framework to investigate depression, followed by basic terminology that will facilitate a better understanding of the models presented later.

2.2.1 Motivating Reinforcement Learning as a Computational Framework

Nowadays, psychiatric disorders are investigated from multiple angles: biological, evolutionary perspectives, behavioural and neural observations. The perspectives (models and theory) and observations (data and measurements) are complimentary. Based on observations alone, it is difficult to draw a clear conclusion about underlying mechanisms, while theory without real-world hypothesis testing is purely speculative. When it comes to investigating depression, the origin of this impairment is heavily debated. The development of computational models that aim to replicate human behaviour can therefore shed new light on this phenomenon, and, more specifically, offer insight in the underlying mechanisms.

RL is the appropriate computational framework for this, because beliefs, perceptions of stimuli, and interactions with the environment influence action selection, and lead to the observed learning behaviour (in the form of learning curves) which provide information about action selection. In an RL model, these components can be easily expressed, manipulated, and isolated. This allows us to gain insight into a mechanism rather than a collection of symptoms. This can either support or weaken certain theories and give us new directions for further research.

In addition to Broekens [9, 11, 24], previous models have been developed using RL as a framework to investigate phenomena related to decision-making, affect, and, specifically, depression. In addition, RL has been applied as a framework for studying decision-making in emotional disorders, in which

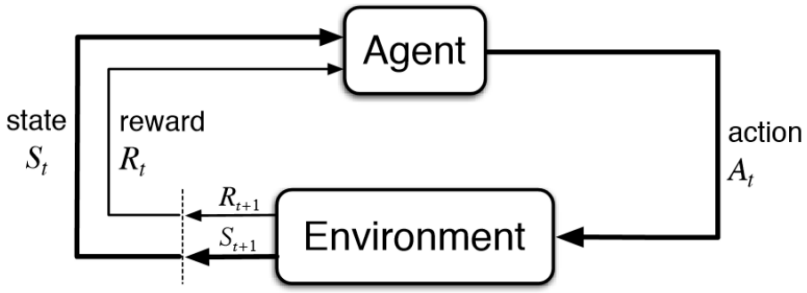

Figure 2: The agent-environment interaction, taken from [3]. An agent first selects an action A_t which affects the environment. From that it obtains a reward R_{t+1} and the environment evolves to state S_{t+1}.

prior beliefs and environmental interactions shape action selection [1].

Prior work in computational psychiatry links symptoms and theories of depression to a computational reinforcement learning framework. Two examples can be found in [25] and [13]. The former presents different aspects of anhedonia and their corresponding translation into RL terminology. The latter conducted a behavioural meta-analysis of a reinforcement learning task and concluded that anhedonia is expressed as a reduction in reward sensitivity (rather than in learning rate). This directly motivates the reduced-sensitivity model discussed later in Section 6.10.

2.2.2 What is Reinforcement Learning

Reinforcement Learning is a computational approach to learning from a goal-directed interaction. Among machine learning methods, RL is the closest to human and animal learning, and some aspects of this approach were inspired by biological learning mechanisms [3]. The idea of RL is that agents learn from interacting with their environment. Namely, they select an action, observe the outcome and update their beliefs based on the observation. The goal of the agent is to maximise the cumulative reward along the task.

2.2.3 Multi-Armed Bandit Tasks

To understand the idea of multi-armed bandits, we will open with an example. Imagine the following scenario: you go on a two-week holiday, and there are multiple restaurants in the city you are staying in. You and your partner are planning to have dinner every evening. You have never visited any of them before, but you have some recommendations from friends, and you read some reviews. Your goal is to have the tastiest meals on average during the holiday. The recommendations and reviews, which are your estimated knowledge about each restaurant (that will be updated according to your visits) is the $Q_t(a)$. Each evening is the time step t, and each dinner is the reward r. The idea is to pick the best restaurant every evening. But what if the cook is sick that day? What if the other restaurant that you haven't tried the food at yet is so much better? What should you do? The agent faces the same type of issues in a multi-armed bandit task, where each action is an arm

and with each performed action, the agent observes a reward. This thesis focuses on a two-armed bandit task, namely, with only two actions available.

2.2.4 Q-learning in Bandit Problems

In the bandit algorithm, the agent is in a certain state S_t at time step t. At each time step, the agent selects an action a (A_t in Figure 2) from a finite set of available actions in the environment. For a given action, the agent might obtain a reward with a given probability (in a stochastic environment). As mentioned, the agent's goal is to maximise the rewards obtained by selecting optimal actions, a process also called policy selection. The agent has an internal belief of the state-action value, $Q(a)_t$, which is updated based on its interactions with the environment. It performs this using an update rule. One such update rule that can be used for Q-learning is the following:

$$Q(a)_{t+1} = Q(a)_t + \alpha \cdot (r_t - Q(a)_t), \tag{1}$$

where $Q(a)_{t+1}$ is the new Q-action value, $Q(a)_t$ is the old stored one. α is the learning rate, which determines how much of the new reward is incorporated into the agent's new belief. The larger the α, the more the agent learns from new observations. $r_t - Q(a)$ is the prediction error, the difference between the actual reward obtained and the agent's belief about the expected reward.

2.2.5 Action Selection, Exploration and Exploitation

Different action selection mechanisms or policies exist, such as greedy (argmax), epsilon-greedy, and random. In this thesis, we focus on the softmax policy, which, according to literature [26], describes human choice behaviour well. The softmax policy, also known as the Boltzmann or Gibbs exploration, has been used in Q-learning since its early development [27], and is defined as:

$$P(a_i) = \frac{e^{Q(a_i)/\beta}}{\sum_{j=1}^{n} e^{Q(a_j)/\beta}} \tag{2}$$

In this thesis, the task we deal with has only two actions to choose from. The softmax policy for two actions, derived from Equation 2.2.5, is defined as follows:

$$P_t(a_1) = \frac{1}{1 + e^{\frac{Q_t(a_0) - Q_t(a_1)}{\beta}}}, \qquad P_t(a_0) = 1 - P_t(a_1) \tag{3}$$

where, in a two-action choice experiment, $P_t(a_1)$ is the probability of selecting action 1, $Q_t(a_0)$ is the Q-value of action 0. β is the parameter which controls the exploration-exploitation trade-off of the agent within its action selection mechanism. The higher the β, also called temperature, the more the agent tends to explore (to pick an action with a lower Q-action value, which mostly leads to picking new actions it has not tried before or tried less times). The lower the beta, the more exploiting the agent is. Namely, the more likely it is to pick the action that has a high Q-action-value, which usually leads to an action it has already tried before, and therefore uses its previous knowledge rather than exploring new options.

3 From Psychology to RL

In the background section, different symptoms of depression were discussed, as well as different psychological theories of depression. Here, we bridge the gap between theory and RL implications, laying the foundations for the models discussed later in Section 5.

Anhedonia Recall the consummatory, motivational and decisional aspects of anhedonia (Section 2.1.1). Each maps to a different component of the learning agent. In [25], the authors present different aspects of anhedonia and their corresponding translation into RL terminology.

For the consummatory aspect, the direct ability to "like" and perceive the stimulus as rewarding, the most straightforward translation is in reward perception mechanisms. Computationally, it means that when the agent perceives the reward, it scales it down. In addition, in [13], the authors conducted a behavioural meta-analysis of a reinforcement learning task and concluded that anhedonia is expressed as a reduction in reward sensitivity, rather than in learning rate. This directly motivates the reduced-sensitivity model discussed later in Section 6.10. It can also be expressed with a learning rate bias. For example, "depressed" agents are simulated with a lower learning rate when obtaining a positive reward, compared to a higher learning rate when obtaining a negative reward.

The motivational and decisional aspects of anhedonia, which are the aspects that are related to the ability of "wanting" a reward and action towards obtaining it, in RL terms, relate to the Q-value or Q-action value of a certain action. They are captured by manipulating the initialisation, as well as action selection mechanisms, biasing the agent away from exploitation, as the agent has less motivation to explore in order to learn new things.

Learned Helplessness The "giving up after repeated failure" (Section 2.1.2) maps onto the action selection mechanism (policy), but can be interpreted in two different ways. In the first, after obtaining a negative reward, the agent exploits more, having "understood" that its actions have no impact. In the second, giving up is seen as more of an exploration than an exploitation, as the agent selects its actions more randomly because it "sees" no value in continuing to explore, and this is its way of giving up. Both are realised through the exploration-exploitation balance.

Negative Beliefs and Mood Congruency Mood congruency (Section 2.1.3) can also be modelled with RL in different ways. First, the agent may perceive the reward differently (as explained in the anhedonia paragraph). Because the agent is "depressed", it perceives more negative reward and less positive reward. This can be expressed as a learning rate bias and can also result from a negative prediction error. In addition, negative beliefs can be expressed by manipulating Q-values and how the agent initialises them (as discussed in the context of motivational and decisional anhedonia).

Depressive Realism According to this theory (Section 2.1.4), depressed people have a more realistic view of the results of their actions. In RL terms, it can be translated into the exploration-exploitation trade-off. Those who are not depressed, according to this theory, tend to have a

positivity bias, which makes them believe their actions are connected to the outcome more than they actually are. That way, they tend to be more explorative. For the depressed patients it goes the other way around, leaving them more exploiting and holding on to their previous beliefs, as trying out new actions is pointless.

4 The Negativity Bias Model (Vandendriessche et al. 2022)

This thesis builds upon a paper written by Vandendriessche et al. [8], which introduces a negativity bias model to reproduce and explain human trial data. This paper presents a study involving a RL task performed by participants in 2 different groups: The first group contained 30 MDD patients (all under medical anti-depression treatment of SSRIs). The second group contained 26 participants without MDD, matched by age, gender, and socioeconomic status. Both groups performed a 2-armed bandit task, in which they repeatedly chose between two symbols (the arms). Each symbol corresponds to a probability function of the reward, depending on the context. In the "poor context", the probability of obtaining a positive reward was 10% and 40% for each arm, respectively. In the "rich" context, it was 60% and 90%. For each context, the task was performed 50 trials. The probability function was never explicit, and the participants had to learn it from the trials.

During the experiment, the participants were presented with the explicit reward. After each trial, the reward appeared: either a positive (+1, green smiley) or a negative (-1, red smiley). The participants were told to try to maximise the reward.

After analysing the results, the researchers observed the following phenomena: the non-depressed group showed no significant difference in learning between the "rich" and "poor" contexts. Depressed patients showed a worse learning curve in the "poor" context. This can be seen in Figure 3.

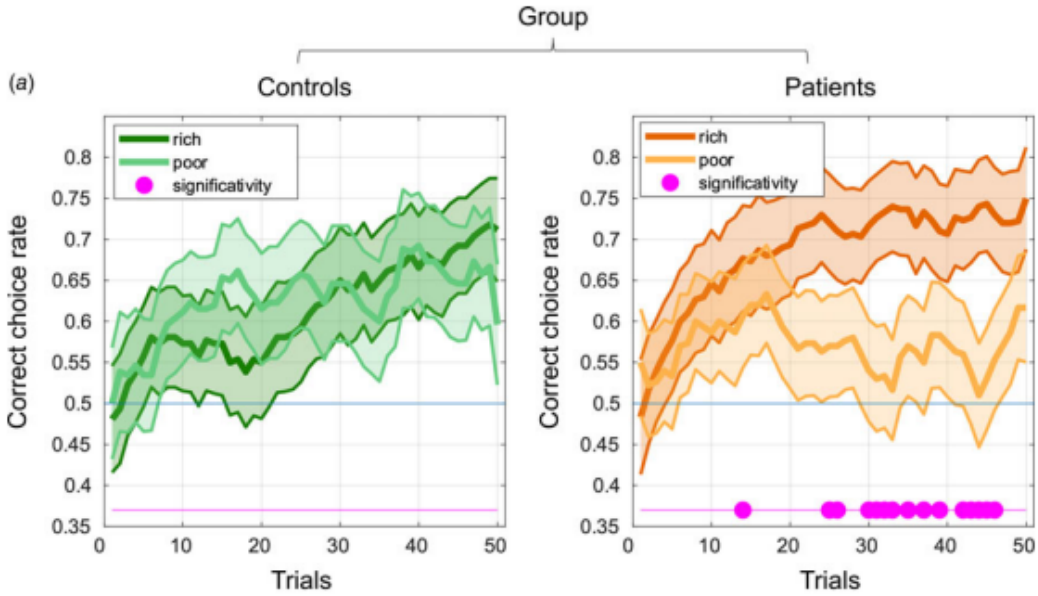

Figure 3: Data collected from the participant's repeated trials, taken from the paper [8]. The y-axis is the correct choice rate, which is the probability of choosing the correct arm. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The pink dots show the trials that significantly differ between the poor and the rich contexts ($p < 0.05$)

To replicate these results, the researchers developed a computational model, which will be presented

in more detail in the Section 6.1. The model, as shown in their paper, is depicted in Figure 4. The model shows that the "depressed" agents learn more from the negative reward they obtain than from the positive reward. For the "healthy" agent, the opposite is true. This, according to the authors, provides an explanation for the difference between the results of the healthy and depressed groups in the experiments.

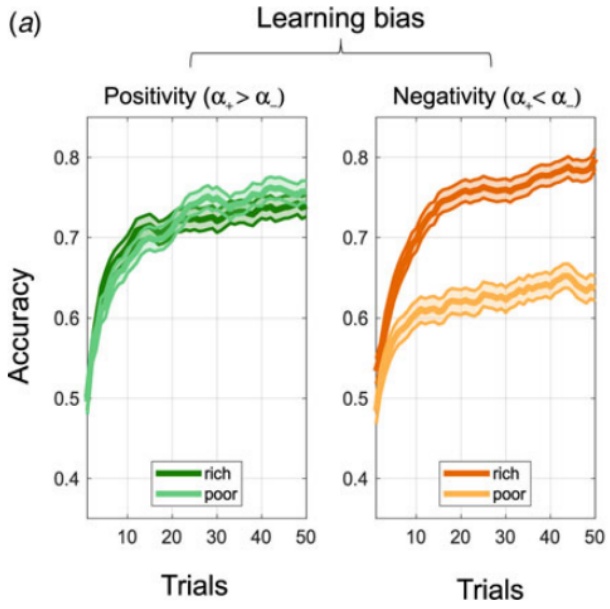

Figure 4: The negativity bias model, taken from the paper [8]. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the number of times the action has appeared. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. On the left are the agents simulated with a positive bias in learning rate (the "healthy" agents) and on the right with a negative bias (the "depressed" agents)

The problem this thesis aims to address is whether different RL mechanisms (other than the discussed negativity bias model) can replicate these results observed in human data. In the next chapter, the other computational models are introduced, along with corresponding motivations.

5 Methodology

This chapter presents the experimental methodology for evaluating the models. First, the task structure and environment will be introduced, which are shared across all models. The models are divided into three main categories: learning rate asymmetry (Sections 6.1 to 6.3), action selection asymmetry (Sections 6.5 to 6.8) and perception and prior beliefs models (Sections 6.9 and 6.10)

Each experiment contains a two-arm bandit. There are two contexts: a rich context, in which the probability of obtaining a positive reward (+1) is 60% for the first arm and 90% for the second arm. In the poor context, the probability of obtaining a reward is 10% for the first arm and 40% for the second. If no positive reward is obtained, a negative reward (-1) is received.

In each context, 1000 agents participated, each with 50 trials. Each agent participated in both contexts with the same hyperparameter sampled, but no knowledge was carried over from one context to the other.

The results of each experiment include the correct choice rate across trials (5-trial average). All experiments use the mean choice rate as the outcome measure. Choice rate is defined as "how many times a given option has been chosen, divided by the number of times a given option has been presented" [8].

An overview of all models can be found in Table 1.

Family	Model	Group	Learning rate distribution	Learning rule	Temperature distribution	Action selection
Baseline	Asymmetric LR on reward (§6.1)	Control	$\alpha^+ \sim \mathcal{U}(0,1)$ $\alpha^- \sim \mathcal{U}(0,\alpha^+)$	Asymmetric on reward sign (α^+ if $r_t > 0$; α^- if $r_t < 0$)	$\beta \sim \mathcal{U}(0,1)$	Softmax over Q with temperature β (Eq. 3)
		Depressed	$\alpha^- \sim \mathcal{U}(0,1)$ $\alpha^+ \sim \mathcal{U}(0,\alpha^-)$			
Learning rule *(asymmetry in α)*	Prediction-error asym. LR (§6.2)	Control	$\alpha^+ \sim \mathcal{U}(0,1)$ $\alpha^- \sim \mathcal{U}(0,\alpha^+)$	Asymmetric on PE sign (α^+ if $\delta_t > 0$; α^- if $\delta_t < 0$)	$\beta \sim \mathcal{U}(0,1)$	Standard softmax (Eq. 3)
		Depressed	$\alpha^- \sim \mathcal{U}(0,1)$ $\alpha^+ \sim \mathcal{U}(0,\alpha^-)$			
	Regret-dependent asym. LR (§6.3)	Control	$\alpha^+ \sim \mathcal{U}(0,1)$ $\alpha^- \sim \mathcal{U}(0,\alpha^+)$	Asymmetric on regret (α^+ if $\text{regret}_t = 0$; α^- if $\text{regret}_t > 0$)		
		Depressed	$\alpha^- \sim \mathcal{U}(0,1)$ $\alpha^+ \sim \mathcal{U}(0,\alpha^-)$			
Action selection *(asymmetry in β)*	Simple β manipulation (§6.5)	Control	$\alpha \sim \mathcal{U}(0,1)$ (symmetric)	Symmetric (Eq. 1)	Exp. 1: $\mathcal{U}(0,0.5)$ Exp. 2: $\mathcal{U}(0,0.125)$	Standard softmax (Eq. 3)
		Depressed			Exp. 1: $\mathcal{U}(0.5,1)$ Exp. 2: $\mathcal{U}(0.9,1)$	
	Reward-dependent asym. β (§6.6)	Control	$\alpha \sim \mathcal{U}(0,1)$ (symmetric)	Symmetric (Eq. 1)	$\beta^+ \sim \mathcal{U}(0,1)$ $\beta^- \sim \mathcal{U}(0,\beta^+)$	Softmax with $\beta = \beta^+$ if $r_t > 0$, else β^-
		Depressed			$\beta^- \sim \mathcal{U}(0,1)$ $\beta^+ \sim \mathcal{U}(0,\beta^-)$	
	Prediction-error asym. β (§6.7)	Control			$\beta^+ \sim \mathcal{U}(0,1)$ $\beta^- \sim \mathcal{U}(0,\beta^+)$	Softmax with $\beta = \beta^+$ if $\delta_t > 0$, else β^-
		Depressed			$\beta^- \sim \mathcal{U}(0,1)$ $\beta^+ \sim \mathcal{U}(0,\beta^-)$	
	Regret-dependent asym. β (§6.8)	Control			$\beta^+ \sim \mathcal{U}(0,1)$ $\beta^- \sim \mathcal{U}(0,\beta^+)$	Softmax with $\beta = \beta^+$ if $\text{regret}_t = 0$, else β^-
		Depressed			$\beta^- \sim \mathcal{U}(0,1)$ $\beta^+ \sim \mathcal{U}(0,\beta^-)$	
Perception & prior belief	Pessimistic prior (§6.9)	Control	$\alpha \sim \mathcal{U}(0,1)$ (symmetric)	Symmetric (Eq. 1); $Q_0 = 0$ vs $Q_0 < 0$	$\beta \sim \mathcal{U}(0,1)$	Standard softmax (Eq. 3)
		Depressed				
	Reduced sensitivity (all rewards, §6.10)	Control		Symmetric on $\tilde{r}_t = \rho\, r_t$; $\rho = 1$ vs $\rho \sim \mathcal{U}(0,1)$		
		Depressed				
	Reduced sensitivity (positive only, §6.10)	Control		Symmetric; $\tilde{r}_t = \rho\, r_t$ if $r_t = 1$, else r_t		
		Depressed				

Table 1: Overview of the baseline and alternative models. Models are grouped by the mechanism in which control and depressed agents differ: action selection (β^-/β^+ asymmetry), learning rule (α^-/α^+ asymmetry), and prior belief / reward perception. Distributions shown per group indicate the parameters that drive the control-vs-depressed asymmetry.

6 Experiments

Nine experiments, each for a different model, were performed. Each subsection of the experiment presents the motivation for developing the model, the numerical approach, and the simulated model. It is followed by a brief discussion of each model's results. The python code used to generate the results can be consulted at https://github.com/NoamDunsky/Simulating-Depression-with-RL–BSc-Thesis.

6.1 Experiment 0: Baseline Model

6.1.1 Model and Initialisation

The baseline model, the first experiment conducted, follows the computational framework described in the paper by [8]. This model provides the replication baseline against which the other models are compared. It consists of asymmetric learning rates, with "depressed" agents having a negative bias and "healthy" agents having a positive bias.

Learning Rule

There are 2 learning rates, each corresponding to a different case.

According to the model, the Q-values are updated, depending on whether the reward obtained is positive or negative. In the case of positive reward, the Q-value is updated using a "positive" learning rate, which is larger for the control group agents. If a negative reward is obtained, the Q-value is updated using a "negative" learning rate, which is larger than the "positive" learning rate in "depressed" agents, and vice versa for the control agents. This is described in the following formula:

$$Q_{t+1}(a) = \begin{cases} Q_t(a) + \alpha^+ \cdot (r_t - Q_t(a)) & \text{if } r_t > 0 \\ Q_t(a) + \alpha^- \cdot (r_t - Q_t(a)) & \text{if } r_t < 0 \end{cases} \tag{4}$$

where $Q_{t+1}(a)$ is the new Q-action-value, $Q_t(a)$ is the previous action value, r_t is the obtained reward, $r_t - Q_t(a)$ is the prediction error, α^+ is the learning rate for the positive obtained reward and α^- is the learning rate for the obtained negative reward.

Action Selection

The policy is a softmax function, calculated with Equation 3.

Hyper-Parameter Initialisation

The temperature β is sampled from a uniform distribution between 0 and 1 ($\beta \sim \mathcal{U}(0,1)$). For the positive bias ("control" agents) the learning rate for positive obtained reward (α^+) was drawn from a uniform distribution between 0 and 1 ($\alpha^+ \sim \mathcal{U}(0,1)$), while for the negative reward (α^-) it was drawn from a uniform distribution between 0 and α^+, meaning smaller than α^+ ($\alpha^- \sim \mathcal{U}(0,\alpha^+)$).

For the negative bias (the "depressed" agents), α^- was drawn from a uniform distribution between

0 and 1 ($\alpha^- \sim \mathcal{U}(0,1)$), while α^+ was drawn from a uniform distribution between 0 and α^- ($\alpha^+ \sim \mathcal{U}(0,\alpha^-)$).

6.1.2 Results

The first experiment conducted was to replicate the model discussed in [8], see Figure 4. In Figure 5, the results show that the agents (positive and negative bias) across both contexts start learning around 0.5 correct choice rate. For the "healthy" population of agents, most learning occurs during the first 10 trials. After that, in the rich context, the curve starts flattening, and the agents continue to choose the correct arm around 72.5% of the time. In the poor context, it finishes around 75%. In contrast, when we observe the "depressed", negative-biased agents, there is a large difference between the learning curve in the rich environment and the poor environment. Both start around 50% (at chance level), but in the poor environment, the learning curve is flatter and does not exceed 65%. However, the rich context curve rises sharply during the first 10 trials, and the slope remains positive but less steep. The correct choice rate at the final trial ends around 80%, which is higher than for the "healthy" population in both environments.

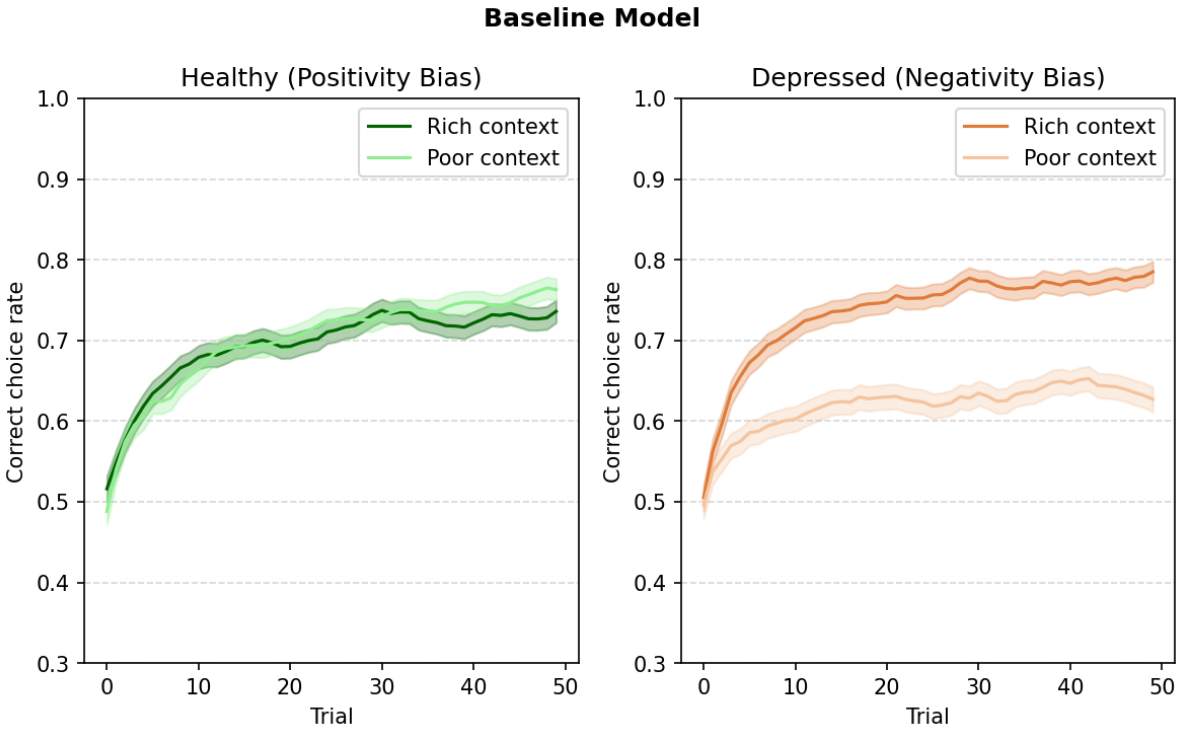

Figure 5: Baseline model: Learning rate asymmetry. For each group of 1000 simulated agents, if a positive reward is obtained, the agent learns using α^+. In case of obtaining a negative reward, it uses α^-. For the "healthy" population (on the left, in green): $\alpha^+ > \alpha^-$, and for the depressed (right, in orange) it is the other way around. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the number of agents.

6.1.3 Discussion

The model successfully replicates the negativity bias of [8] as discussed in Section 4. There is no gap in the learning curves of the "healthy" agents simulated in both contexts, whereas there is a large gap between contexts for the "depressed" agents, with those simulated in the "rich" context outperforming those in the "poor" context. However, when examining the human data from [8] (Figure 3), the negativity bias model presented by the authors doesn't accurately replicate the collected human data as well.

First, in the control participants of the human data of [8], they performed better in the rich context at the 50th trial, while in the model, the opposite is obtained, and learning seems to have converged, looking at the healthy human control data. Moreover, looking at the start and end of the curve of the depressed patients, until the 10th trial, both groups performed approximately the same, whereas in our and their models, there is already a difference in the learning curve at the 10th trial. Finally, it seems that towards the final trials, the patients' learning in both contexts starts to converge, a behaviour not reflected in the simulated and baseline models. It is important to note that it aims to replicate the model and not the human data.

The original model, as presented in [8], is based on the concept of learning rate asymmetry. There, the learning rate depends on the reward obtained. However, this idea can be expanded to different signals, and depending on which, the learning rate can differ. In the next section, these models will be presented.

6.2 Experiment 1: Prediction Error Asymmetric Learning Rate Model

6.2.1 Model and Initialisation

The baseline model (reward-dependent asymmetric alpha model) obtains a signal from the reward. For this model, the signal is the prediction error. It is defined as follows:

$$\delta_t = r_t - Q_t(a) \tag{5}$$

In this model, if the prediction error is negative, "depressed" agents are programmed to learn more. Like in the previous experiment, we use two learning rates corresponding to the two different cases, as in the baseline model. The idea of prediction error as a negative or positive affect (termed "joy and distress") can be seen in Broekens's work [24]. The model is inspired by "mood congruency". For a depressed person who carries negative views and thoughts (according to Beck's model), they will learn more from an experience like "distress" rather than "joy". For the "healthy" agents, it is the opposite.

According to the model, the Q-values are updated depending on whether the obtained prediction error is positive or negative. In the case of positive prediction error, the Q-value is updated using a "positive" learning rate, which is larger for the "healthy" agents. If a negative prediction error is obtained, the Q-value is updated using a "negative" learning rate, which is larger than the "positive" learning rate in "depressed" agents, and vice versa. In the case of a negative prediction error, the

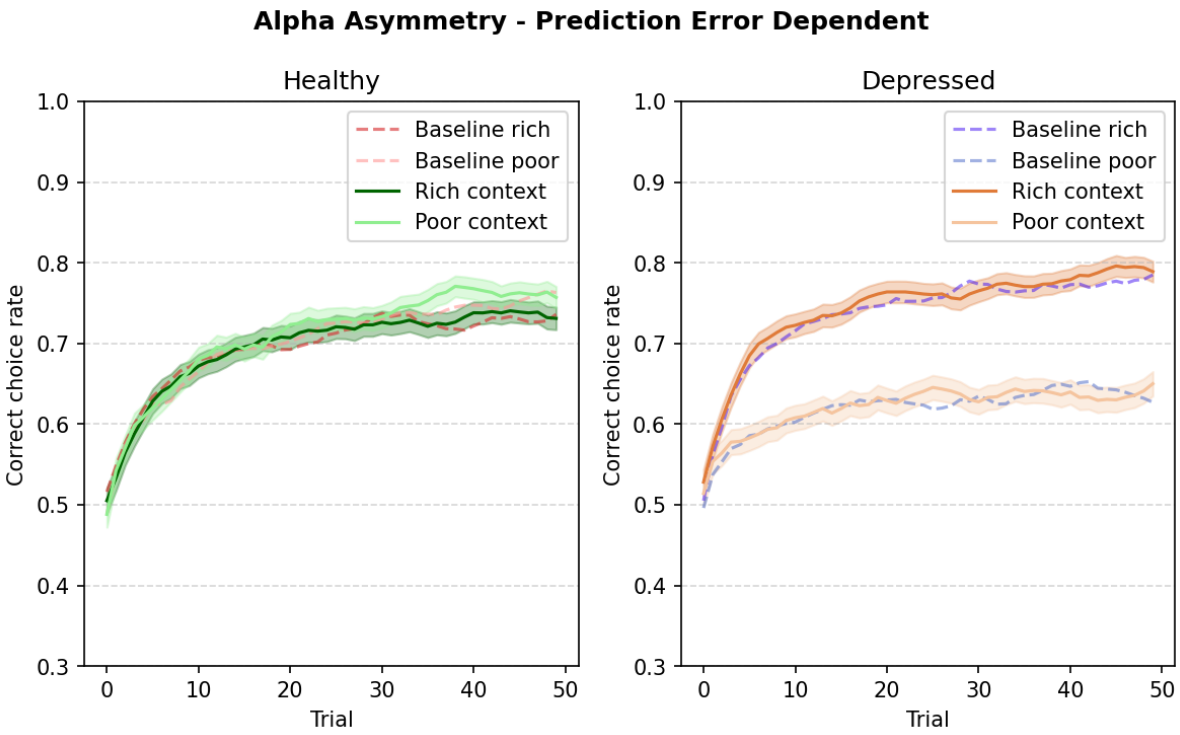

Figure 6: Learning rate asymmetry: Prediction Error Dependent. For each group of 1000 simulated agents, if a positive prediction error is obtained, the agent learns using α^+. In case of obtaining a negative prediction error, it uses α^-. For the "healthy" agents (on the left, in green): $\alpha^+ > \alpha^-$, and for the "depressed" (right, in orange) it is the other way around. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the amount of agents. The dotted darker curve in each plot represents the baseline model results in the rich environment, while the lighter represents the results in the poor context, for each population.

agent uses α^- in the update rule. For a negative prediction error, it uses α^+.

$$Q_{t+1}(a) = \begin{cases} Q_t(a) + \alpha^+ \cdot \delta_t & \text{if } \delta_t > 0 \\ Q_t(a) + \alpha^- \cdot \delta_t & \text{if } \delta_t < 0 \end{cases} \tag{6}$$

where δ_t is the prediction error of Equation 5.

6.2.2 Results

In this experiment, visible in Figure 6, the graphs behave similarly to the baseline model, and do not exceed the shadowed areas, which are the standard error of the mean.

6.2.3 Discussion

Based on the learning curves, this model successfully replicates the baseline model's results, seemingly offering a first alternative mechanism. But diving deeper into the mechanism, it becomes clear that, for this experimental setup, both models are mathematically equivalent. The reason is the binary nature of the reward (+1 or -1) and that the model relies on the sign of the prediction error, which is, in this case, always equal to the sign of the reward itself. The mathematical proof can be viewed in detail in Appendix A.

6.3 Experiment 2: Regret-Dependent Asymmetric Learning Rate

6.3.1 Model and Initialisation

After experimenting with the idea of a different signal to bias the learning rate, the "regret" signal came to mind, inspired by Broeken's paper, [11]. The rest of the model is the same as the baseline model and the prediction error alpha asymmetry model.

The idea of this model is that when depressed people feel regret, they tend to learn less compared to when they do not feel regret. This is based on the idea of "learned helplessness"- when the "depressed" agents experience a failure, they tend to give up.

Regret is defined as the difference between the best stored value and the value of the chosen arm:

$$\text{regret}_t = \max(Q_t(a_0), Q_t(a_1)) - Q_t(a) \tag{7}$$

where $\max(Q_t(a_0), Q_t(a_1))$ is the maximum stored Q-action-value, meaning the arm the agent should have chosen (if it is exploiting), and $Q_t(a)$ is the arm it eventually chose.

The learning rule of this model is defined as follows:

$$Q_{t+1}(a) = \begin{cases} Q_t(a) + \alpha^+ \cdot \delta_t & \text{if regret}_t > 0 \\ Q_t(a) + \alpha^- \cdot \delta_t & \text{if regret}_t = 0 \end{cases} \tag{8}$$

Again, for depressed agents $\alpha^- > \alpha^+$, and the opposite for the healthy.

6.3.2 Results

In this experiment, as shown in Figure 7, the plots for the "healthy" agents (the green on the left) behave similarly to the baseline model in the first 10 trials. After that, the agents simulated in the rich environment perform better than those in the poor environment, and they end up with a correct choice rate of 80%, compared to the baseline that ended around 75%. In the poor environments, both (this model and the baseline) end around 72% correct choice rate.

Among the "depressed" agents, the model's pattern is similar to that of agents simulated in the rich context environment. For the poor context, the learning rate of this model is slightly higher than the baseline. Around the 15th trial, the learning rate evolves a slightly higher, ending at almost 70% correct choice rate, while the baseline rate ends around 64%.

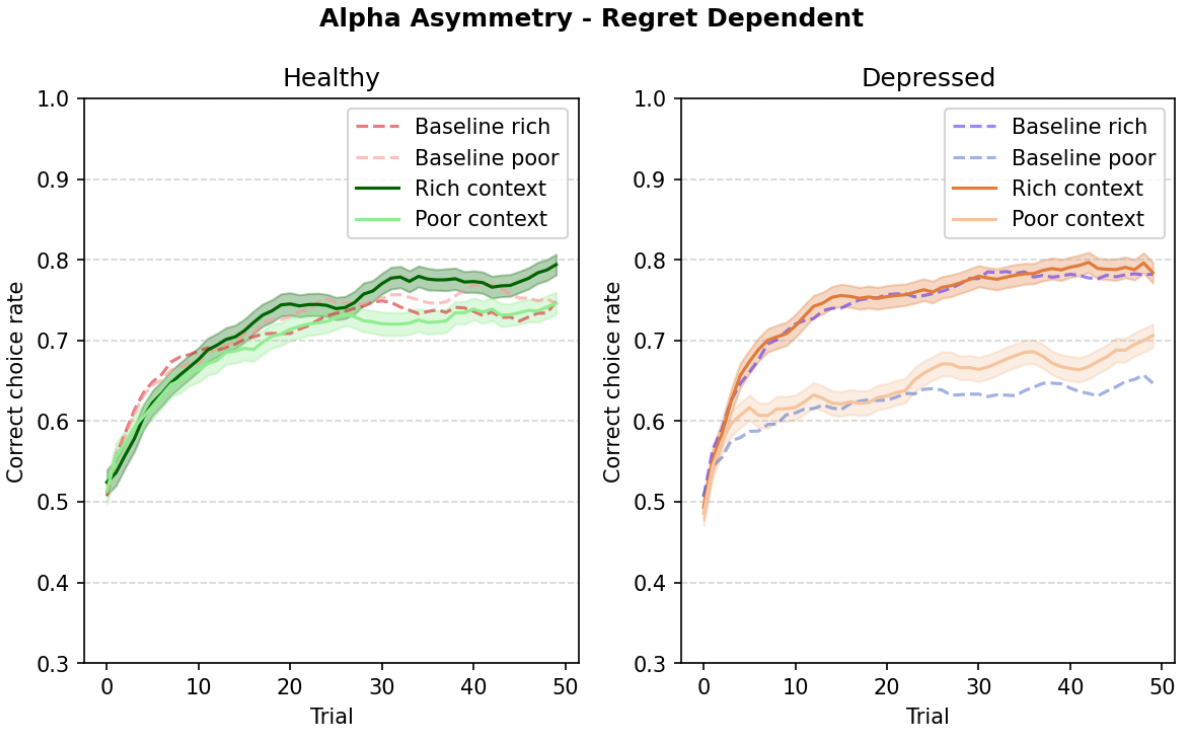

Figure 7: Learning rate asymmetry: Regret-dependent. For each group of 1000 simulated agents, if a positive "regret" is obtained, the agent learns using α^+. In case of obtaining a negative "regret", it uses α^-. For the "healthy" agents (on the left, in green): $\alpha^+ > \alpha^-$, and for the "depressed" (right, in orange) it is the other way around. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the amount of agents. The dotted darker curve in each plot represents the baseline model results in the rich environment, while the lighter represents the results in the poor context, for each population.

6.3.3 Discussion

The learning rate pattern is similar across both agent groups, especially in the "depressed" group, with only minor differences. The graph shows the same behaviour pattern as the baseline model, indicating that the gap exists. The "depressed" agent group plots for this model are slightly above those of the baseline model. In the "healthy" agent's plot, the agents in the "poor" context are slightly higher than in the "rich" context, but the difference is minor. Due to the small differences, it appears to be a good candidate for replicating the baseline model.

6.4 Action Selection Models

In experiments 3-6, instead of exploring variations of the learning rule signals, alternative action selection mechanisms are considered. These experiments involve different variations of a bias in temperature rather than in the learning rate. These present variations of different signals, as in the previous group of experiments.

In the softmax formula for action selection in the baseline model (Equation 3), the β parameter controls the balance between exploration and exploitation. Namely, the larger the β, the more the agent is exploring (the agent selects the action more randomly), and the lower the β, the more the agent is exploiting (chooses the actions based on previously stored knowledge of Q-action values). For this group of models, agents update their Q-values regardless of the reward they receive (i.e., with a symmetric learning rate). This is described with the simple learning rule in Equation 1.

In these experiments, the learning rate α is sampled from a uniform distribution between 0 and 1 ($\alpha \sim \mathcal{U}(0,1)$). For the positive bias ("control" agents), β^+, the temperature for positive obtained reward, was drawn from a uniform distribution between 0 and 1 ($\beta^+ \sim \mathcal{U}(0,1)$), while β^- was drawn from a uniform distribution between 0 and β^+ ($\beta^- \sim \mathcal{U}(0,\beta^+)$), meaning smaller than β^+.

For the negative bias (the "depressed" agents), β^- was drawn from a uniform distribution between 0 and 1 ($\beta^- \sim \mathcal{U}(0,1)$), while β^+ was drawn from a uniform distribution between 0 and β^- ($\beta^+ \sim \mathcal{U}(0,\beta^-)$). Each β is sampled per agent across both contexts.

6.5 Experiment 3: Simple Beta Manipulation

6.5.1 Model and Initialisation

This model aims to explain the difference by assuming that "depressed" agents have a lower temperature (beta) than "healthy" agents, rather than biased learning rates. This is based on the idea of "learned helplessness" (as can bee seen in Section 2.1.2), in which depressed patients feel like their actions have less effect on the environment, so they stop trying, and thus explore less. The idea is that lower exploration might lead to worse learning in the "poor" context, as the rewards are less frequent.

For each agent, α is sampled uniformly between 0 and 1, once per agent across both contexts. To investigate the effect of the different β's, multiple experiments have been conducted, with a range of β, to reflect different distances between temperatures of "healthy" and "depressed" agent groups. The list of combinations can be consulted in Table 2.

Combination	Healthy β	Depressed β
1	$\beta \sim \mathcal{U}(0.7, 1.0)$	$\beta \sim \mathcal{U}(0.0, 0.3)$
2	$\beta \sim \mathcal{U}(0.5, 1.0)$	$\beta \sim \mathcal{U}(0.0, 0.5)$
3	$\beta \sim \mathcal{U}(0.5, 0.8)$	$\beta \sim \mathcal{U}(0.3, 0.5)$
4	$\beta \sim \mathcal{U}(0.6, 0.9)$	$\beta \sim \mathcal{U}(0.4, 0.6)$
5	$\beta \sim \mathcal{U}(0.75, 1.0)$	$\beta \sim \mathcal{U}(0.5, 0.65)$

Table 2: Parameter combinations tested in the systematic search for the Simple Beta model. The healthy β range is strictly higher than the "depressed" β range in all combinations, reflecting the assumption that "depressed" agents explore less.

6.5.2 Results

In this experiment, as shown in Figure 8, the curves of the "healthy" agents have a gap. In the "rich" context, the results are close to the original baseline model, but the plot for the "poor" context is below it. From the 10th trial onwards, a difference in correct choice rate of approximately 8% is observed. This gap in the "healthy" agents group is inconsistent with the baseline model pattern, suggesting that this parameter combination does not fully capture the "healthy" agents' behaviour.

In the "depressed" agents' curves on the right, agents in both contexts exhibit behaviour similar to the curves of "depressed" agents of the baseline model, except for a minor difference in the "poor" context. There, the model's curve is slightly higher than that of the baseline model. The results of the other parameter combinations can be found in Appendix B, along with the parameter table in Table 2.

6.5.3 Discussion

This model appears to partially replicate the baseline model. The plots of the "depressed" agent group are similar. However, when examining the "healthy" plots, it is clear that agents in the "poor" context learn less than those in the "rich" context, creating a gap we aim to avoid in replicating the human data.

6.6 Experiment 4: Reward-Dependent Asymmetric Beta

Experiments 4-6 are based on the concept of temperature asymmetry and are conditioned on different signals, similar to the learning rule models presented in experiments 0-2. We start by defining the temperature asymmetry. For "healthy" agents, $\beta^+ > \beta^-$, meaning they tend to explore more when they receive a positive reward and exploit more when they receive a negative reward. For the "depressed" agents, it is the other way around. When "depressed" agents obtain a negative reward, they tend to explore more compared to the positive reward, which connects to the second interpretation of "learned helplessness": they give up and act more randomly after a "failure". This can be seen in more detail in Section 3.

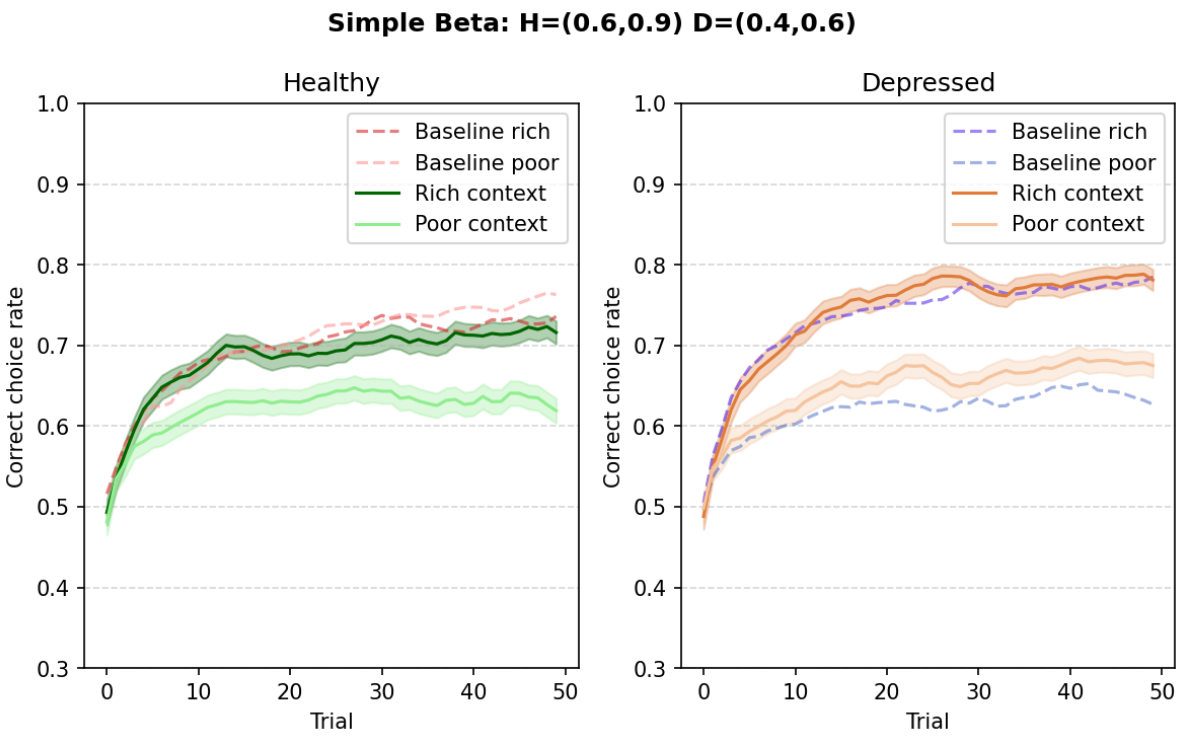

Figure 8: Static model of temperature (beta) asymmetry. For each group of 1000 simulated agents. The temperature of the "healthy" agents (on the left, green) is higher than that of the "depressed" agents (on the right, orange). The β (temperature) of the "healthy" agents is sampled from a range between 0.6 and 0.9, while the β of the depressed is between 0.4 and 0.6. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the amount of agents. The dotted darker curve in each plot represents the baseline model results in the rich environment, while the lighter represents the results in the poor context, for each population.

6.6.1 Model and Initialisation

For the first asymmetric beta model, the temperature is dependent on the previously obtained reward. In the case of a negative reward, the agent uses a different temperature than in the case of a positive reward:

$$P_t(a_1) = \frac{1}{1 + e^{\frac{Q_t(a_0) - Q_t(a_1)}{\beta_t}}} \tag{9}$$

$$\beta_t = \begin{cases} \beta^+ & \text{if } r_{t-1} > 0 \\ \beta^- & \text{if } r_{t-1} < 0 \end{cases} \tag{10}$$

where β_t is the temperature that is used in a specific timestep and r_{t-1} is the previously obtained reward.

6.6.2 Results

In this experiment, as shown in Figure 9, the "healthy" agents' behaviour is similar to the baseline model's in the "poor" context. However, in the "rich" context, the learning curve starts climbing steeply from the beginning, and stabilises after the 20th trial at around 80%.

For the "depressed" agents on the right, we can again observe a clear separation between those simulated in the "rich" context and those in the "poor" context. However, when comparing it to the baseline model curves, it is clear that they both perform better across contexts in terms of learning rate. The agents simulated in the "rich" environment eventually end the trials with around an 84% correct choice rate, and those simulated in the "poor" context with 70%.

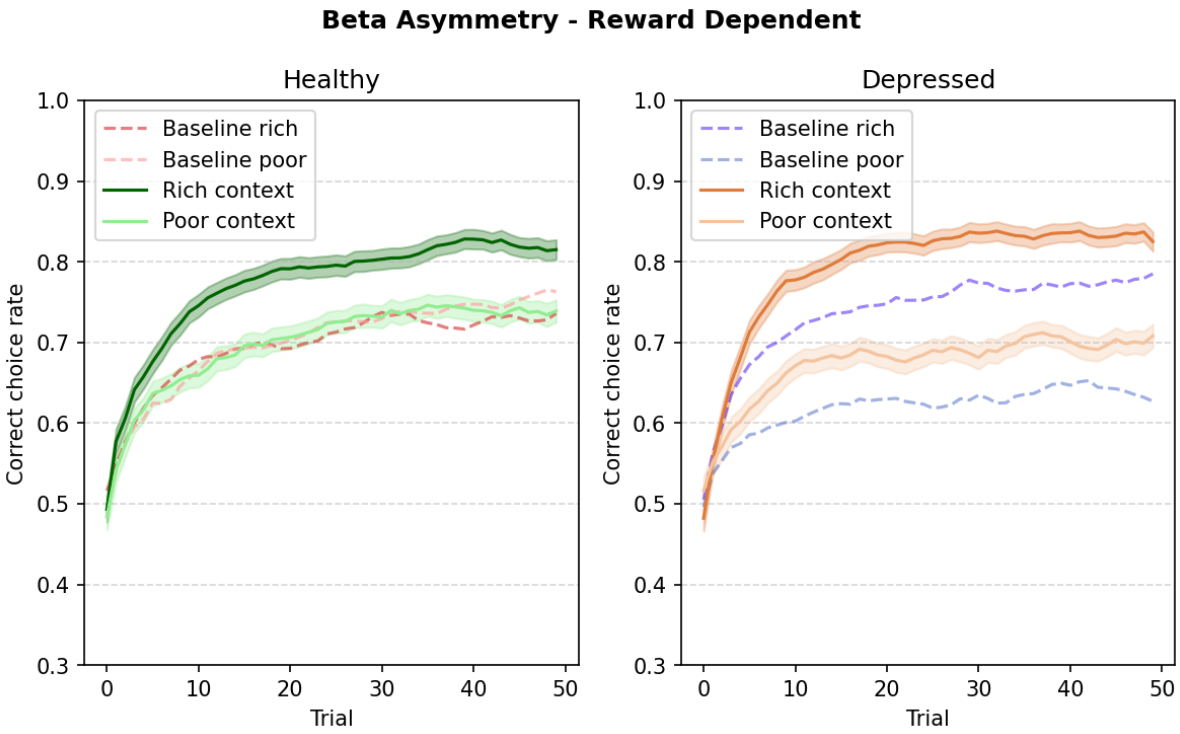

Figure 9: Reward-dependent temperature (beta) asymmetry. For each group of 1000 simulated agents, if a positive reward is obtained, the agent select actions using β^+. In case of obtaining a negative reward, it uses β^-. For the "healthy" agents (on the left, in green): $\beta^+ > \beta^-$, and for the "depressed" (right, in orange) it is the other way around. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the amount of agents. The dotted darker curve in each plot represents the baseline model results in the rich environment, while the lighter represents the results in the poor context, for each population.

6.6.3 Discussion

This model does not fully replicate the baseline model. From observing the plots, in the "depressed" agents group, there is a gap between the two contexts, but they don't overlap with the baseline model's results. In the "healthy" agents group, the plot for the agents in the "poor" context overlaps with the baseline model's plots, but the plot for the "rich" context is higher, creating a gap between the two contexts that should not exist according to the baseline model.

6.7 Experiment 5: Prediction Error Asymmetric Beta

6.7.1 Model and Initialisation

This model is analogous to the reward-dependent asymmetric beta model, except that the signal is the prediction error rather than the reward. The prediction error definition can be seen in Equation 5. If the prediction error is negative, "depressed" agents tend to explore more. This is based on the idea of the second interpretation of "learned helplessness" as in experiment 4 (reward-dependent beta asymmetry).

In this experiment, the temperature depends on the prediction error:

$$\beta_t = \begin{cases} \beta^+ & \text{if } \delta_{t-1} > 0 \\ \beta^- & \text{if } \delta_{t-1} < 0 \end{cases} \tag{11}$$

where δ_{t-1} is the previously obtained prediction error.

6.7.2 Results

In this experiment, as shown in Figure 10, the behaviour is similar to the model described under the model "beta asymmetry-reward dependent" (Section 6.6).

6.7.3 Discussion

The similarity between the models is due to their mathematical equivalence. This is the same logic that applies to the baseline model and the alpha asymmetry prediction-error-dependent model (experiment 1), and it can be found in Appendix A.

6.8 Experiment 6: Regret-Dependent Asymmetric Beta

6.8.1 Model and Initialisation

Similar to the previous beta-asymmetry models, but now the signal is "regret". More about the "regret" signal can be seen in Section 6.3.

Regret, as discussed before, is a feeling that can occur after making a wrong choice. In this sense, regret connects to the first interpretation of the "learned helplessness" theory of depression. When an agent "experiences" failure or regret, they stop trying. In this model, when an agent

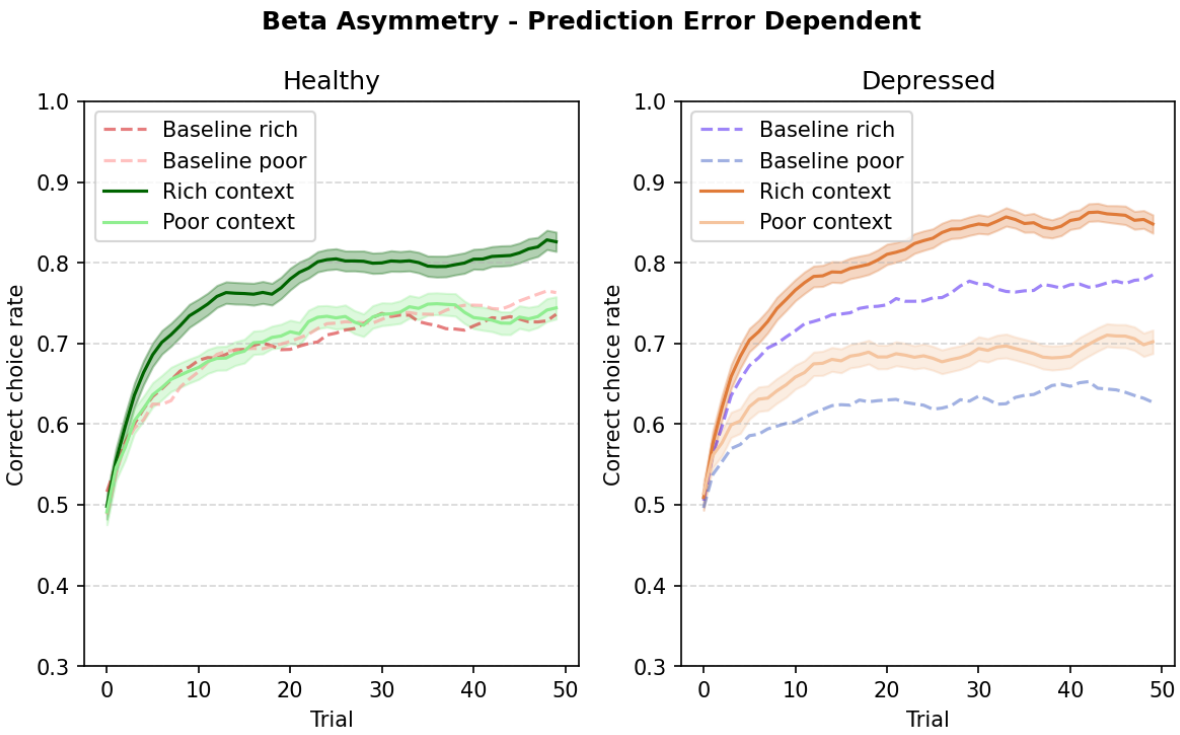

Figure 10: Temperature (beta) asymmetry: prediction error dependent. For each group of 1000 simulated agents, if a positive prediction error is obtained, the agent select actions using β^+. In case of obtaining a negative prediction error, it uses β^-. For the "healthy" agents (on the left, in green): $\beta^+ > \beta^-$, and for the "depressed" (right, in orange) it is the other way around. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the amount of agents. The dotted darker curve in each plot represents the baseline model results in the rich environment, while the lighter represents the results in the poor context, for each population.

"experiences" regret, they stop exploring, whereas the "healthy" group of agents explores more after experiencing regret, meaning they don't give up after a failure. Another theory that can explain it is "depressive realism", as can be seen in Section 2.1.4, in which "depressed" agents explore less after failure, because they are more aware that their actions aren't always connected to results so they stop trying new actions.

In this model, we define a regret-dependent temperature. In the case of no regret, the agent uses β^-. For a "positive regret" it uses β^+.

$$\beta_t = \begin{cases} \beta^+ & \text{if } \text{regret}_{t-1} > 0 \\ \beta^- & \text{if } \text{regret}_{t-1} = 0 \end{cases} \tag{12}$$

where regret_{t-1} is the previously obtained regret.

6.8.2 Results

In the experiment, as shown in Figure 11, the left plots ("healthy" agent group) show a large gap between the two contexts. The agents in the "rich" context outperform the agents in the "poor" context. The former ends in a correct choice rate of almost 90%, whereas the latter ends in 75%. Looking at the right plots of the "depressed" group, there is a smaller gap between the two contexts. The agents simulated in the "rich" contexts almost overlap with the baseline model, whereas those in the "poor" contexts outperform the baseline model (the former finished in 70% while the latter in 62%.

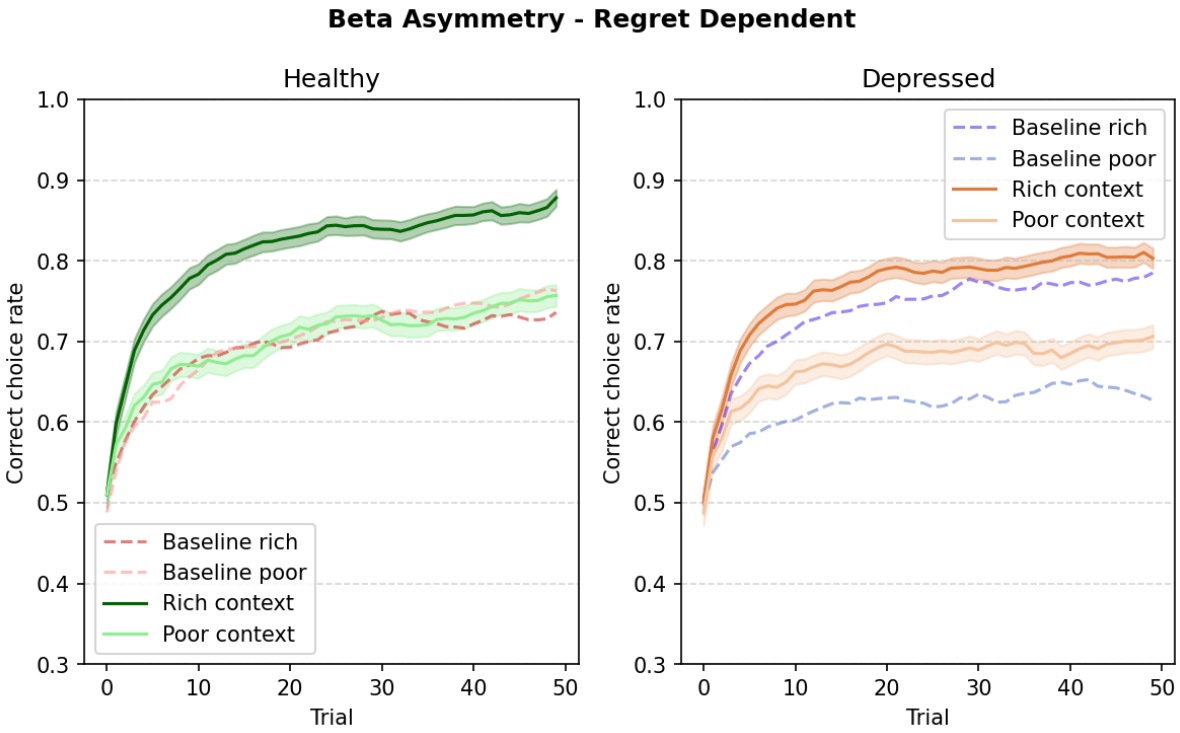

Figure 11: Temperature (beta) asymmetry: regret dependent. For each group of 1000 simulated agents, if a regret is obtained, the agent selects actions using β^+. In case of obtaining no regret, it uses β^-. For the "healthy" agents (on the left, in green): $\beta^+ > \beta^-$, and for the "depressed" (right, in orange) it is the other way around. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the number of agents. The darker dotted curve in each plot represents the baseline model results in the rich environment, while the lighter curve represents the results in the poor context for each population.

6.8.3 Discussion

There is a gap between the agents simulated in both contexts. In the "healthy" agent group, the gap is bigger than in the "depressed" agent group. This means that the model should have been

swapped between the two. In the "depressed" agent group, when "depressed" agents "feel" regret, they should explore more. In any of these cases, it is clear that the model fails to capture the behaviour of the baseline model plots, making it a poor candidate for replication.

6.9 Experiment 7: Pessimistic Prior

6.9.1 Model and Initialisation

This model is based on Beck's cognitive model of depression, introduced in Section 2.1.3. According to Beck, depressed people tend to have negative prior beliefs about themselves and about the world. In this model, the Q-values were initialised differently between "depressed" agents and "healthy" agents. For the "depressed" agents, the Q-action values are initialised to -1, while for the "healthy" agents they remain initialised to 0 as before.

In this experiment, the learning rule is the basic update rule, Equation 6. The hyperparameters are selected as follows: The learning rate α is sampled from a uniform distribution between 0 and 1 ($\alpha \sim \mathcal{U}(0,1)$) and the temperature β is sampled from a uniform distribution between 0 and 1 ($\beta \sim \mathcal{U}(0,1)$).

6.9.2 Results

In this experiment, as shown in Figure 12, the left graph for the "healthy" agent group shows that the agents simulated in both the "rich" and "poor" contexts differ from those simulated in the baseline model. In the "rich" context, the simulated agents in this model outperform those in the baseline model from the first trials, whereas in the "poor" context, they underperformed. In the "rich" context, they stabilise around 80% and then drop down a bit (to 77%), versus 69% in the "poor" context.

For the "depressed" agents, there is a gap between the two contexts, but it is smaller than the baseline model and the gap across contexts in the "healthy" agents graph (Figure 12). In the "rich" context, the plot overlaps with the baseline model. In the "poor" context, the "depressed" agents outperform the baseline model and the "healthy" agents, ending at around 71%, while the baseline plot around 62%.

6.9.3 Discussion

Similarly to the regret-dependent asymmetric beta model (Section 6.8), the pessimistic prior exhibits gaps in correct choice rate between the two contexts. Switching between the graphs, initialising the "healthy" agents with a Q-value of -1 and the "depressed" ones with 0, will produce better results, but still doesn't close the gap between the two plots and doesn't align with the model's theoretical grounding. Therefore, the model is a poor candidate for replicating the baseline model. Further experiments with different variations could be conducted to improve the model by using different initialisations for Q-values.

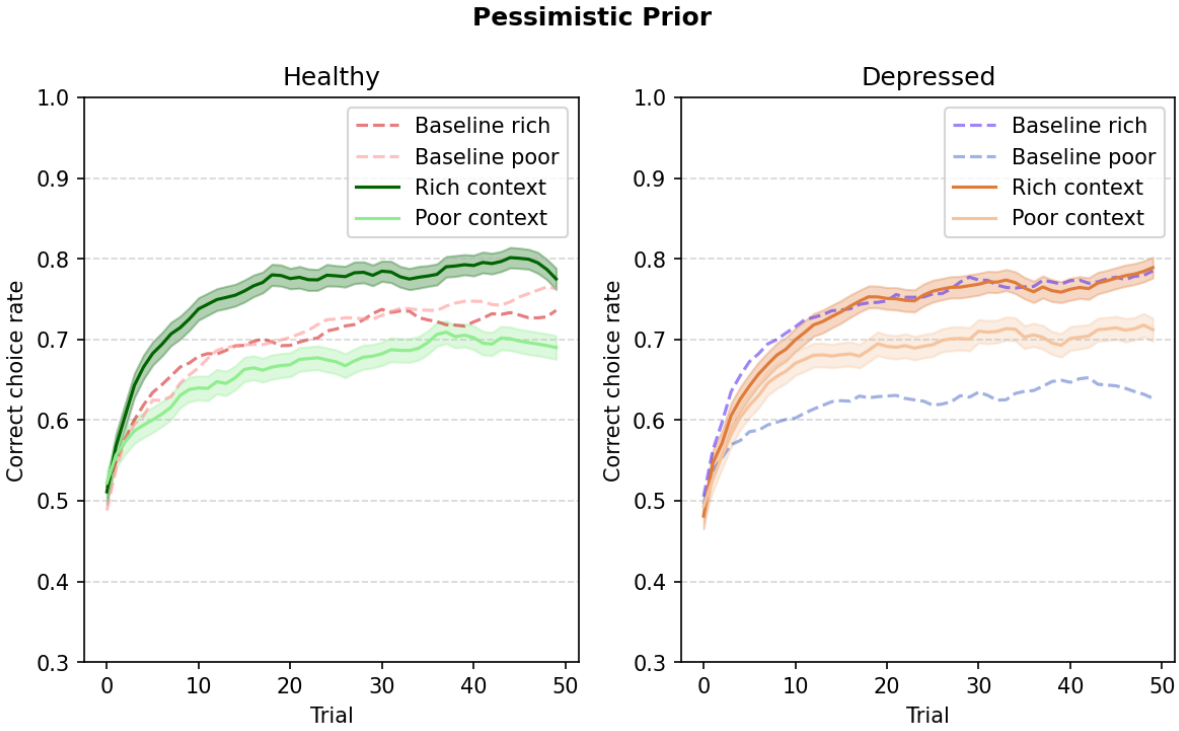

Figure 12: Pessimistic prior. For each group of 1000 simulated agents, in the "healthy" agents (on the left, in green), agents initialise their Q-values to 0, whereas in the "depressed" population, they initialise them to -1. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the number of agents. The darker dotted curve in each plot represents the baseline model results in the rich environment, while the lighter curve represents the results in the poor context for each population.

6.10 Experiment 8: Reduced Sensitivity

6.10.1 Model and Initialisation

In this section, two variants of the reduced-sensitivity model are discussed.

The models are based on the idea of anhedonia, as described in the Section 2.1.1. This translates to scaling down the reward value, as discussed in the paper [13]. There are two variations:

- A model in which the reward is being scaled down for both positive and negative rewards, only for "depressed" agents.
- A model in which rewards are scaled down in the case of obtaining a positive reward, only for "depressed agents".

The learning rule is symmetric, with one learning rate. For both variants, the Q-value is updated

using the scaled reward $\tilde{r}_t$ in place of the original reward:

$$Q_{t+1}(a) = Q_t(a) + \alpha \cdot (\tilde{r}_t - Q_t(a)) \tag{13}$$

where $\tilde{r}_t$ is the perceived reward. For variant 1, the reward is a simple scaling function:

$$\tilde{r}_{1,t} = r_t \cdot \rho \tag{14}$$

For variant 2, we only scale the positive rewards:

$$\tilde{r}_{2,t} = \begin{cases} r_t \cdot \rho & \text{if } r_t = 1 \\ r_t & \text{if } r_t = -1 \end{cases} \tag{15}$$

where $\rho \in [0, 1]$ is the sensitivity parameter, with $\rho = 1$ for "healthy" agents (no scaling), representing full reward perception. For "depressed" agents, $\rho \sim \mathcal{U}(0, 1)$, representing reduced reward sensitivity.

For each agent, $\alpha \sim \mathcal{U}(0, 1)$ and $\beta \sim \mathcal{U}(0, 1)$ are sampled once per agent in both contexts.

6.10.2 Results

The results of the second variant will be shown ("depressed" agents scale down positive rewards). The results of the first variant can be found in Appendix C.

In this experiment, as shown in Figure 13, in the left graph for the "healthy" agent group, the agents simulated in both the "rich" and "poor" contexts differ from those simulated in the baseline model. In the "rich" context, the simulated agents in this model outperform those in the baseline model from the first trials, whereas in the "poor" context, they underperformed. In the "rich" context, they end up at 80%, and 67% in the "poor" context.

The "depressed" agents present different picture that looks more similar to the baseline model. There is a clear distinction between the results in both contexts. In the "rich" context, the agents learn better compared to the "poor" context. They end up at around 77% and 65%, respectively, similar to the baseline model.

6.10.3 Discussion

Both variants fail to replicate the baseline model, creating a gap between the plots for the "rich" and "poor" contexts within the "healthy" agent group. More variations of the model can be tested as future work.

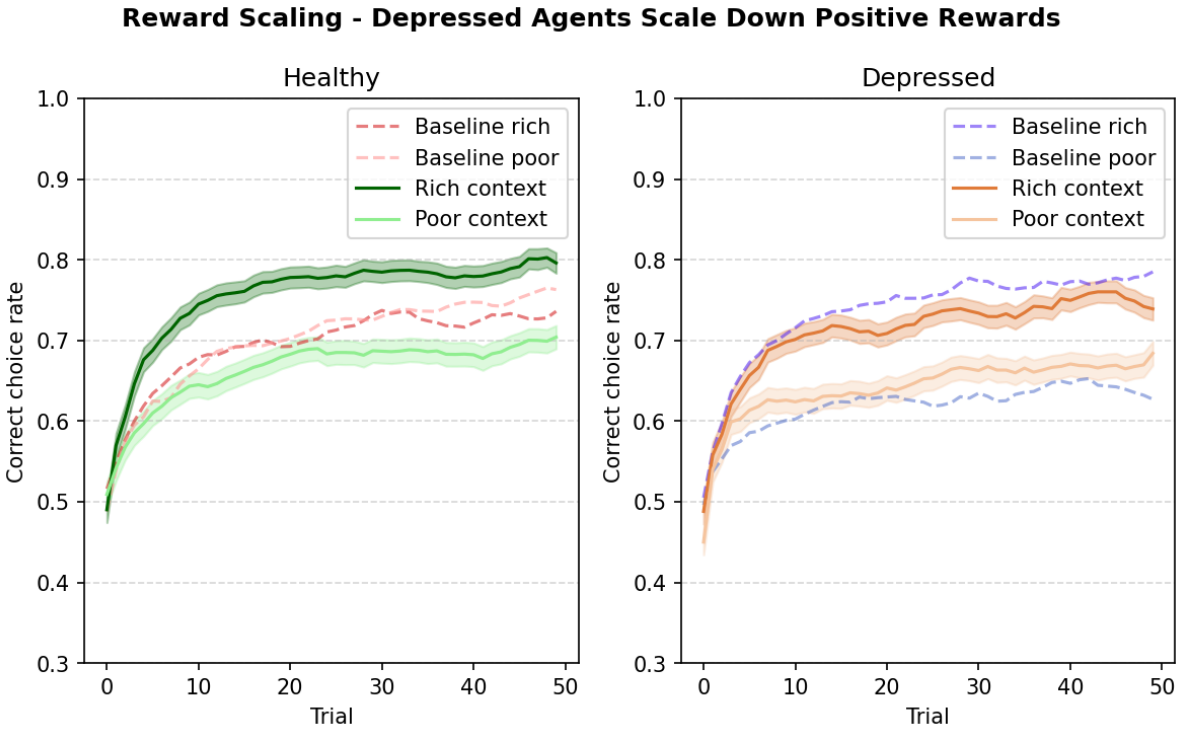

Figure 13: Variant 2 of Reward scaling model: "depressed" agents scale down positive rewards. For each group of 1000 simulated agents, for the "healthy" agents (on the left, in green), the agents update their Q-values with the obtained reward. For the "depressed" agents (on the right, in orange), if the reward is positive, it is being scaled down. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the amount of agents. The dotted darker curve in each plot represents the baseline model results in the rich environment, while the lighter represents the results in the poor context, for each population.

7 Discussion

In this thesis, we have examined the results of simulations of different models, each with distinct mechanisms, signals, or parameters. The aim was to investigate whether different mechanisms could be identified as equal candidates for replicating the human data from [8]. From the results, it is clear that some of these models were found to be candidates for replicating the human data, namely, the learning rate asymmetry prediction error dependent (Section 6.2) and the learning rate asymmetry regret dependent (Section 6.3), which share the same mechanism as the baseline model but differ in the signal they rely on. Note that the alignment of the candidates was assessed qualitatively through visual inspection.

The prediction error is the difference between the obtained reward and the Q-value. In other words, it is the difference between the reality as it is perceived by the agent and the previous belief it was "holding". Consequently, if the prediction error is positive, the results exceed expectations; otherwise, they fall short. Broekens [24] compared this to a "joy and distress" emotions, respectively. According to Feldman Barrett [28], the brain functions as a prediction machine, and emotions arise when there is a mismatch between expectations and reality. This also connects to the idea of regret, a failure in choosing correctly, which is the difference between what an agent should have chosen and what it actually chose. Hence, it may not come as a surprise that the mechanism related to depression, being an emotional disorder, is well explained by models that apply asymmetry in prediction error signals and regret and not only the obtained reward.

As it turns out, due to the structure and symmetry of the task in [8], there are mathematical equivalences between some of the models discussed. However, the research question is answered, as a different RL mechanism yielded the same result. That means there might be more complexity than the reward alone, and it is a direction worth further developing and investigating across different tasks. Furthermore, the experiments show that a learning bias mechanism is most effective at replicating the human data and is therefore the most promising candidate for explaining the difference in results between the "healthy" and "depressed" agents groups. This shows that the theories about anhedonia, negative beliefs, mood congruency, and learned helplessness are good potential explanations for depression.

7.1 Limitations and Future Work

There are multiple limitations to consider in [8]. First, it is mentioned that all participants are being treated with SSRIs. Therefore, it is worth controlling for medication treatment and conducting the same task on participants who are not receiving medical treatment, or with a different medication, to rule out that the difference between the results is a side effect of the medication.

Second, the setup of the experiment is relatively simple. As a result, several elements of the discussed results can be mathematically equivalent. To differentiate between the specific mechanisms, there needs to be a more complex task that enables a more precise comparison. This can be achieved in at least three ways:

1. A bandit task with three or more arms;

2. Non-binary rewards;

3. A multi-step environment, i.e. an MDP (Markov Decision Process), accounting for delayed rewards, which allows disentanglement between reward and temporal difference signal, in contrast to the k-armed bandit, which only considers immediate rewards.

Third, the simulation in both cases (this thesis and [8]) is the ideal within-subject experimental setup, which most probably could not occur in real life with human participants. It is difficult to know how much learning is being carried from one context to the other in humans, whereas in the simulated agents, no learning is carried over.

Future work should include developing a model that more accurately captures the shape of the observed learning curves in clinical trials. In addition, future work can be dedicated to create more variants to existing models in order to improve them.

8 Conclusion

This thesis succeeded in providing an alternative model to [8] that could replicate human results. Among all models, the best candidates are the learning rate asymmetry models, which share the same mechanism but differ in the signal they rely on. Specifically, the learning rate asymmetry prediction error dependent (as shown in experiment 1), produces the same results for both "healthy" and "depressed" agents and is shown to be mathematically equivalent for this specific task. Still, in terms of theory and interpretation, it suggests that a different mechanism can be an alternative explanation and reproduce the observed behaviour equally well. Depression may need to be modelled in a more complex way than a simple negativity bias based solely on perceived rewards.

9 A.I. Disclosure

During the preparation of this thesis, a large-language-model assistant (Anthropic Claude) was consulted for formatting and presentation tasks, specifically the generation of LaTeX code for the model overview table and formatting of equations, as well as Python plotting code for the learning curve result figures. All derivations, results and interpretations are the original work of the author, who takes full responsibility for the content of this thesis.

References

[1] Martin P. Paulus and Angela J. Yu. Emotion and decision-making: affect-driven belief systems in anxiety and depression. *Trends in Cognitive Sciences*, 16(9):476–483, 2012.

[2] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. American Psychiatric Association, Washington, DC, 5 edition, 2013.

[3] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2 edition, 2018.

[4] Djallel Bouneffouf, Irina Rish, and Charu C. Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.

[5] Nina Rouhani and Yael Niv. Depressive symptoms bias the prediction-error enhancement of memory towards negative events in reinforcement learning. *Psychopharmacology*, 236(8):2425–2435, 2019.

[6] J. Bakic, G. Pourtois, M. Jepma, R. Duprat, R. De Raedt, and C. Baeken. Spared internal but impaired external reward prediction error signals in major depressive disorder during reinforcement learning. *Depression and Anxiety*, 34(1):89–96, 2017.

[7] Yoshihiko Kunisato, Yasumasa Okamoto, Kazutaka Ueda, Keiichi Onoda, Go Okada, Shinpei Yoshimura, Shin-ichi Suzuki, Kazuyuki Samejima, and Shigeto Yamawaki. Effects of depression on reward-based decision making and variability of action in probabilistic learning. *Journal of Behavior Therapy and Experimental Psychiatry*, 43(4):1088–1094, 2012.

[8] Henri Vandendriessche, Amel Demmou, Sophie Bavard, Julien Yadak, Cédric Lemogne, Thomas Mauras, and Stefano Palminteri. Contextual influence of reinforcement learning performance of depression: evidence for a negativity bias? *Psychological Medicine*, 53(10):4696–4706, 2023.

[9] Joost Broekens. Modeling the experience of emotion. *International Journal of Synthetic Emotions*, 1(1):1–17, 2010.

[10] Elmer Jacobs, Joost Broekens, and Catholijn Jonker. Joy, distress, hope, and fear in reinforcement learning (extended abstract). In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, pages 1615–1616, Paris, France, 2014.

[11] Joost Broekens and Laduona Dai. A TDRL model for the emotion of regret. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 150–156, Cambridge, UK, 2019. IEEE.

[12] Elaine Fox. Determinants of emotional disorders. In *Emotion Science: Cognitive and Neuroscientific Approaches to Understanding Human Emotions*, pages 265–306. Palgrave Macmillan, Basingstoke, 2008.

[13] Quentin J. M. Huys, Diego A. Pizzagalli, Ryan Bogdan, and Peter Dayan. Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis. *Biology of Mood & Anxiety Disorders*, 3(1):12, 2013.

[14] Quentin J. M. Huys, Joshua T. Vogelstein, and Peter Dayan. Psychiatry: Insights into depression through normative decision-making models. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

[15] Kent C. Berridge and Terry E. Robinson. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, 28(3):309–369, 1998.

[16] Michael T. Treadway and David H. Zald. Reconsidering anhedonia in depression: Lessons from translational neuroscience. *Neuroscience & Biobehavioral Reviews*, 35(3):537–555, 2011.

[17] Xueqing Ma, Angad Sahni, and Ciara McCabe. Subcomponents of reward processing in adolescent anhedonia. *Nature Reviews Psychology*, 5:185–200, 2026.

[18] Martin E. P. Seligman. Learned helplessness. *Annual Review of Medicine*, 23:407–412, 1972.

[19] William R. Miller and Martin E. P. Seligman. Depression and learned helplessness in man. *Journal of Abnormal Psychology*, 84(3):228–238, 1975.

[20] Steven F. Maier and Martin E. P. Seligman. Learned helplessness at fifty: Insights from neuroscience. *Psychological review*, 123(4):349–367, 2016.

[21] Aaron T. Beck. Cognitive models of depression. *The Journal of Cognitive Psychotherapy: An International Quarterly*, 1(1):5–37, 1987.

[22] Lauren B. Alloy and Lyn Y. Abramson. Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of experimental psychology: General*, 108(4):441–485, 1979.

[23] Michael T. Moore and David M. Fresco. Depressive realism: A meta-analytic review. *Clinical Psychology Review*, 32(6):496–509, 2012.

[24] Joost Broekens, Elmer Jacobs, and Catholijn M. Jonker. A reinforcement learning model of joy, distress, hope and fear. *Connection Science*, 27(3):215–233, 2015.

[25] Quentin J. M. Huys and Michael Browning. A computational view on the nature of reward and value in anhedonia. In *Anhedonia: Preclinical, Translational, and Clinical Integration*, pages 421–441. Springer International Publishing, Cham, 2021.

[26] Lucy Lai and Samuel J. Gershman. Policy compression: An information bottleneck in action selection. In *Psychology of Learning and Motivation*, volume 74, pages 195–232. Academic Press, 2021.

[27] John S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer Berlin Heidelberg, 1990.

[28] Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 01 2017.

A Asymmetric Alpha on Reward - Asymmetric Alpha on Prediction Error Equivalence

Here, I will present the proof of the mathematical equivalence between the baseline model (negativity bias: alpha-asymmetry reward-dependent) and the alpha-asymmetry prediction-error model. The idea is that, in the specific experimental setting, the prediction error and the reward have the same sign.

Note the Rescorla-Wagner update rule:

$$Q_{t+1}(a) = Q_t(a) + \alpha \cdot \underbrace{(r_t - Q_t(a))}_{\delta_t} \tag{16}$$

In this experiment setup, the reward is binary, namely, $r \in \{-1, +1\}$. The way our update rule is defined in both models, it only depends on the sign of the reward. Therefore, if the sign of the reward is identical to the sign of the Q-action value, the models are mathematically equivalent.

For the case in which $r = +1$: We want to prove that: $\delta_t > 0$. For the other case ($r = -1 \Rightarrow \delta_t < 0$), the same reasoning applies. So, in order to prove that $\delta_t > 0$, we need to show that $Q_t(a) < r_t$, which means that $Q_t(a) < 1$. For the sake of symmetry (for the other case), I will show that $Q(a) \in [-1, 1]$. Rewriting Equation 16, we obtain:

$$\begin{aligned} Q_{t+1}(a) &= Q_t(a) + \alpha r_t - \alpha Q_t(a) \\ &= Q_t(a)(1 - \alpha) + \alpha r_t \end{aligned} \tag{17}$$

From how the hyperparameters are sampled in each model, we know that $\alpha \in [0, 1)$. Therefore, from Equation 17, $Q_{t+1}(a)$ is a weighted sum of r_t and $Q_t(a)$ (because $1 - \alpha + \alpha = 1$). We know that $r_t \in [0, 1]$, so if $Q_t(a) \in [0, 1]$ we can claim that $Q_{t+1} \in [0, 1]$ for every t, And therefore, that every $Q_t(a) \in [0, 1]$, and finally, as a result, δ_t is always positive.

To prove that $Q_{t+1} \in [-1, 1]$, we will use induction. Base case: When t=0: $Q_0 = 0$ (by definition of the initialisation of the Q-action values), therefore, $Q_0 \in [-1, 1]$.
Inductive Hypothesis: $Q_t \in [-1, 1]$
What we need to prove: $Q_{t+1} \in [-1, 1]$
From the weighted sum argument, we can see that based on the IH, $Q_{t+1} \in [-1, 1]$.

Therefore, we have proved that $r_t = +1 \Rightarrow \delta_t > 0$ holds.

B Simple Beta Manipulation Model: Experimenting with Different β ranges

The following figures present the results of the remaining experiments discussed in experiment 3, as shown in Table 2.

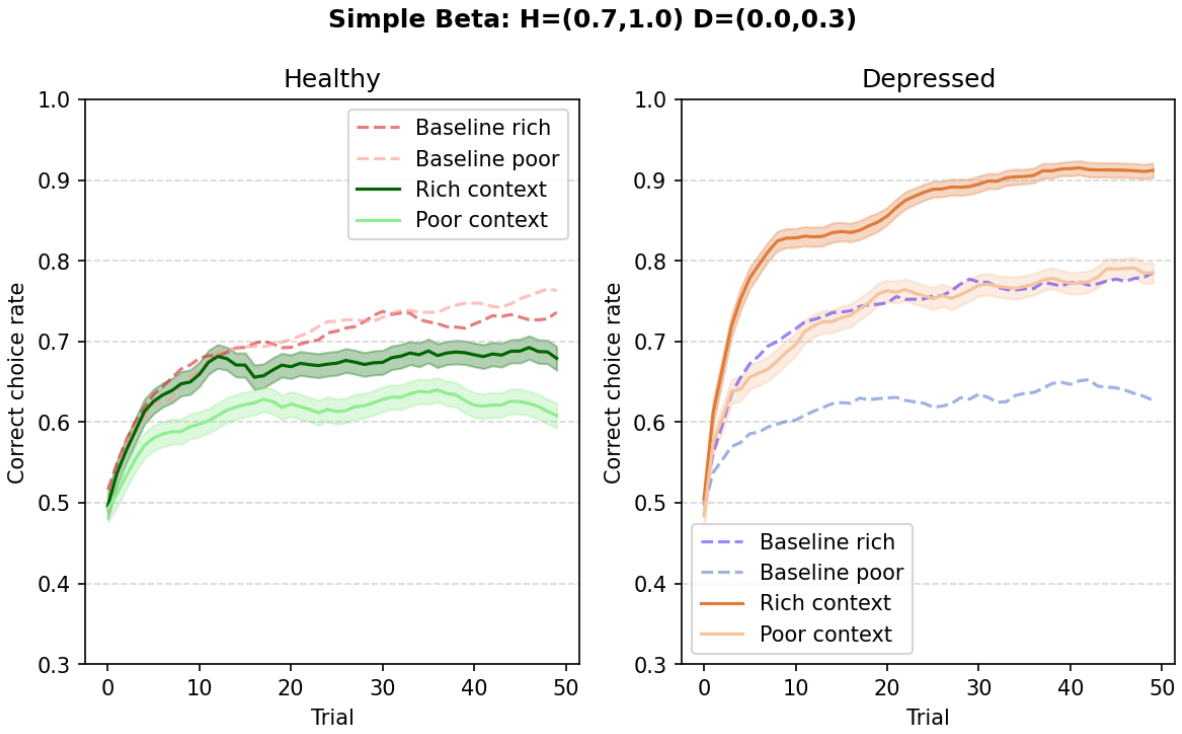

Figure 14: Static model of temperature (beta) asymmetry. For each group of 1000 simulated agents. The temperature of the "healthy" agents (on the left, green) is higher than that of the "depressed" agents (on the right, orange). The healthy β is sampled from a range between 0.7 and 1.0, while the β of the depressed is between 0.0 and 0.3. The thick lines represent a smoothed running average (window of 5 trials), and shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The darker dotted curve in each plot represents the baseline model results in the rich environment, while the lighter curve represents the results in the poor context for each population. The x-axis is the number of trials, the y-axis is the correct choice rate, which is the number of times the agent chose the "correct" arm- the one that maximises the reward, divided by the number of trials.

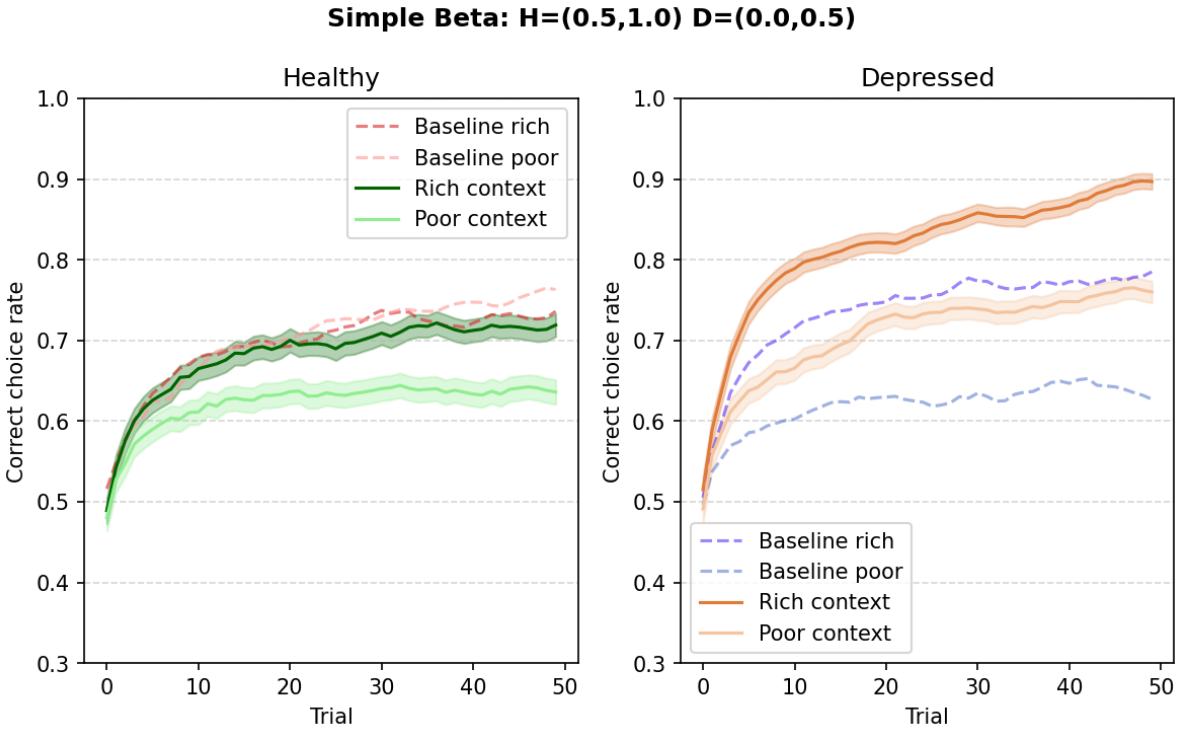

Figure 15: Static model of temperature (beta) asymmetry. For each group of 1000 simulated agents. The temperature of the "healthy" agents (on the left, green) is higher than that of the "depressed" agents (on the right, orange). The healthy β is sampled from a range between 0.5 and 1.0, while the β of the depressed is between 0.0 and 0.5. The thick lines represent a smoothed running average (window of 5 trials), and shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The darker dotted curve in each plot represents the baseline model results in the rich environment, while the lighter curve represents the results in the poor context for each population. The x-axis is the number of trials, the y-axis is the correct choice rate, which is the number of times the agent chose the "correct" arm- the one that maximises the reward, divided by the number of trials.

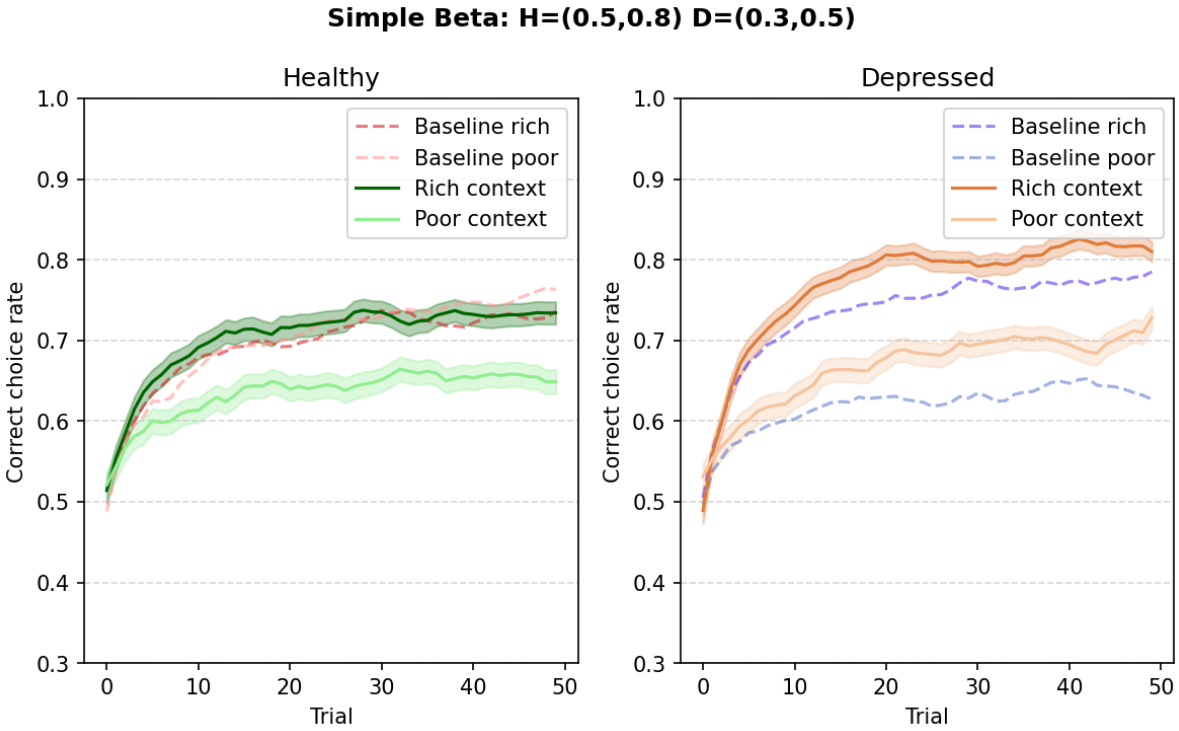

Figure 16: Static model of temperature (beta) asymmetry. For each group of 1000 simulated agents. The temperature of the "healthy" agents (on the left, green) is higher than that of the "depressed" agents (on the right, orange). The healthy β is sampled from a range between 0.5 and 0.8, while the β of the depressed is between 0.3 and 0.5. The thick lines represent a smoothed running average (window of 5 trials), and shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The darker dotted curve in each plot represents the baseline model results in the rich environment, while the lighter curve represents the results in the poor context for each population. The x-axis is the number of trials, the y-axis is the correct choice rate, which is the number of times the agent chose the "correct" arm- the one that maximises the reward, divided by the number of trials.

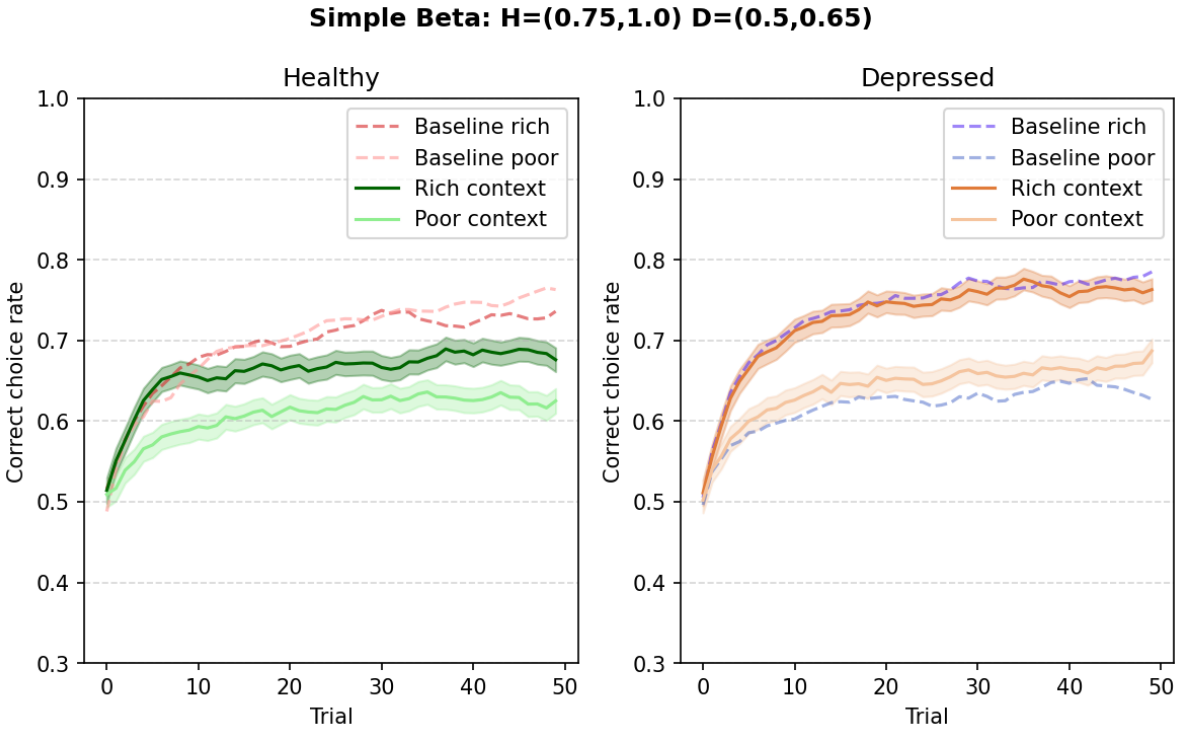

Figure 17: Static model of temperature (beta) asymmetry. For each group of 1000 simulated agents. The temperature of the "healthy" agents (on the left, green) is higher than that of the "depressed" agents (on the right, orange). The healthy β is sampled from a range between 0.75 and 1.0, while the β of the depressed is between 0.5 and 0.65. The thick lines represent a smoothed running average (window of 5 trials), and shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The darker dotted curve in each plot represents the baseline model results in the rich environment, while the lighter curve represents the results in the poor context for each population. The x-axis is the number of trials, the y-axis is the correct choice rate, which is the number of times the agent chose the "correct" arm- the one that maximises the reward, divided by the number of trials.

C Reduced Sensitivity Model: Depressed Agents Scale Down All Rewards

The following figure describes the results of the first variant of experiment 8 (Section 6.10).

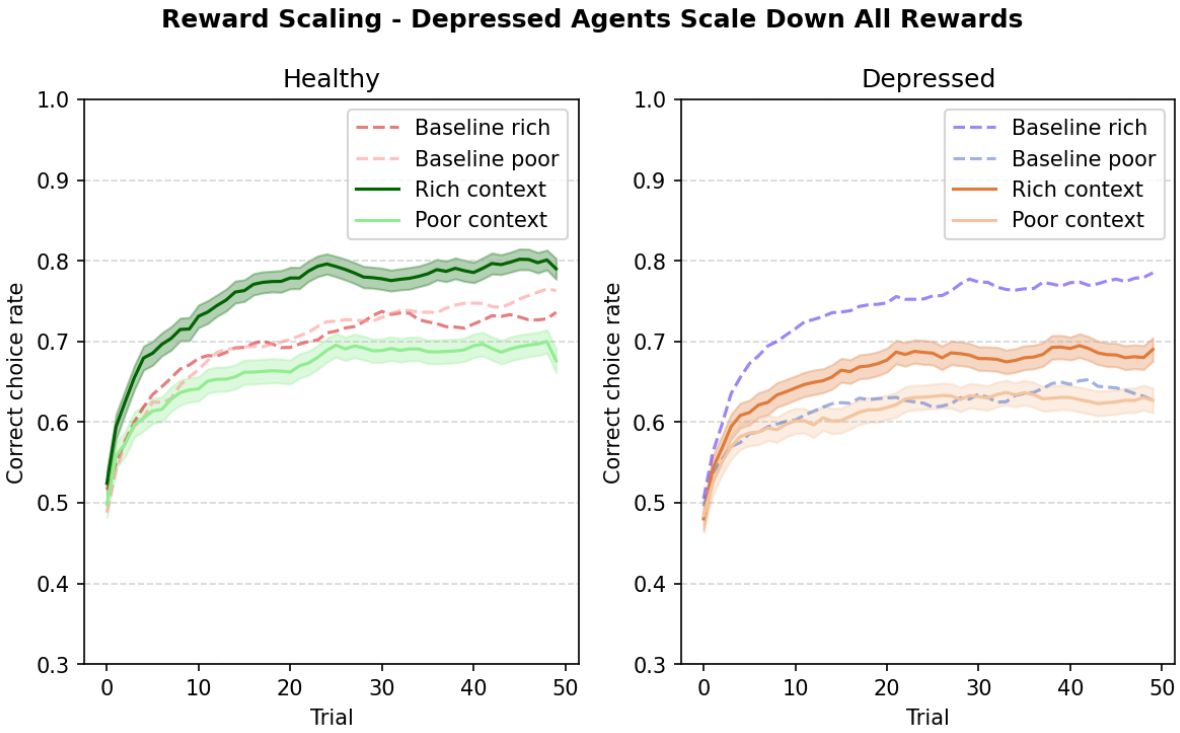

Figure 18: Variant 1 of Reward scaling model: "depressed" agents scale down all rewards. For each group of 1000 simulated agents, For the "healthy" population (on the left, in green), the agents update their Q-values with the obtained reward. For the "depressed" agents, all rewards are being scaled down. The thick lines are a smoothed 5-trial average and the shaded areas are the standard error of the mean. The darker colour represents the rich context, while the lighter one represents the poor context. The x-axis is the trial number, the y-axis is the correct choice rate, which is the number of times the agents chose the arm that maximises the reward, divided by the amount of agents. The dotted darker curve in each plot represents the baseline model results in the rich environment, while the lighter represents the results in the poor context, for each population.