Deanship of Graduate Studies

Al-Quds University

An Extended Actor-Critic Architecture with Phasic Behavioral Inhibition: The Case of Dopamine-Serotonin Interaction

Aya Hussein Ahmad Mousa

M.Sc. Thesis

Jerusalem-Palestine

1440-2018

An Extended Actor-Critic Architecture with Phasic Behavioral Inhibition: The Case of Dopamine-Serotonin Interaction

Prepared By:

Aya Hussein Ahmad Mousa

B.Sc. Computer System Engineering, Birzeit University, Palestine, 2015

Supervisor: Mohammad M. Herzallah, M.D., Ph.D.

A thesis submitted to Faculty of Engineering, Al-Quds University in Partial fulfilment of the requirements for the degree of Master of Electronic and Computer Engineering.

1440 - 2018

Al-Quds University

Deanship of Graduate Studies

Master of Electronic and Computer Engineering

Thesis Approval

An Extended Actor-Critic Architecture with Phasic Behavioral Inhibition: The Case of Dopamine-Serotonin Interaction

Prepared By: Aya Hussein Ahmad Mousa

Registration No. 21612407

Supervisor: Mohammad M. Herzallah, M.D., Ph.D.

Master thesis submitted and accepted. Date: 15/ 12/2018

The names and signatures of the examining committee members are as follows:

1- Head of Committee: Dr. Mohammad Herzallah Signature:.....

2- Internal Examiner: Dr. Radwan Qasrawi Signature:.....

3- External Examiner: Dr. Mahmoud Al-Saheb (PPU) Signature:.....

Jerusalem – Palestine

1440 – 2018

Dedication

To my parents, Hussein & Aisha for their endless love and support

To my aunties, Aysha and Ruqaya for their unconditional support

To my beloved brothers, Ahmad, Ayman, Mohammad, Mahmoud,

Osama for pushing me to success and achieving my goal

To my amazing sisters Samah, Eman, Arwa, Mariam and Bayan for

encouraging me to follow my dreams

To my crazy friends, for becoming my psychiatric therapist at each time I

felt depressed and frustrated

To people who believe in me

I dedicate this research

Declaration

I certify that this thesis submitted for the degree of Master, is the result of my own research, except where otherwise acknowledged, and that this sis study (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Signed.....

Name: Aya Hussein Ahmad Mousa

Date: 15/12/2018

Acknowledgments

I would like to express my sincere gratitude and appreciation to my supervisor **Dr. Mohammad Herzallah,** for his continuous support, patience, motivation, enthusiasm, and immense knowledge. Without his guidance and support; it would have never been possible for me to complete this thesis.

I would like to thank my supervisor at the PNI **Ashar Natsheh** for her invaluable help and support; her guidance helped me on developing my model, I am deeply thankful to her.

I sincerely appreciate the cooperation and help by all members at Palestinian Neuroscience Initiative, thank you all for all your support.

I would like to thank the academic staff of the Faculty of Engineering at Al-Quds University for their support.

Finally, I would like to thank my friends, colleagues, and every one helps this research to see the sunlight.

Abstract

The actor-critic architecture based on the temporal difference (TD) algorithms have been playing a critical role in reinforcement learning. The actor represents the policy structure and critic represents the value function. The TD prediction error signal is used as a teaching signal for both the actor and critic modules. Current models of the actor-critic architecture assume that only the unmodified TD signal can serve as a teaching signal for the actor and critic modules. In this thesis, we introduce an extended version of the actor-critic architecture that addresses the effect of two kinds of reinforcement signals; the TD signal and the behavioral inhibition signal. We argue that the role of the behavioral inhibition signal is to produce phasic opposition of the TD signal in order to ascertain the significance learning and fortify consolidation. Based on this logic, we construct a new neurocomputational model of the brain region the basal ganglia. This model addresses the effects of the neurotransmitters dopamine and serotonin in the reinforcement learning process. The dopamine function is represented by a TD prediction error signal, while serotonin is simulated as a behavioral inhibition signal whose role is to phasically inhibit the TD prediction error signal. We utilize major depressive disorder and selective serotonin reuptake inhibitor (SSRI) antidepressants as experimental representations of variable levels of dopamine and serotonin to study their interaction in reinforcement learning. We use three different modeling approaches to simulate experimental reinforcement learning data: (1) TD only model, (2) TD and risk prediction model, and (3) Our proposed TD and behavioral inhibition model. Simulation results show that our proposed model simulated experimental reinforcement learning data from MDD and SSRIs significantly better the other two modeling approaches. This extended actor-critic architecture can have a myriad of applications in robotics as well as neuroscience.

هيكلية متطورة لخوارزمية الفاعل والناقد بهدف التثبيط السلوكي والتفاعل بين الناقلين العصبيين السيرتونين والدوبامين كفرضية لدراسة الهيكلية

إعداد الطالبة: آية حسين أحمد موسى

إشراف: د. محمد حرزالله

الملخص

تحتل هيكلية الناقد والفاعل القائمة على خوازميات التنبؤ بالخطأ للفرق الزمني دورًا حاسمًا في تعلم بواسطة التحفيز الايجابي والسلبي. تمثل وحدة " الفاعل" الهيكلية التي يتم بواسطتها تحديد الفعل الذي سيقوم به الوكيل للحصول على أفضل مكافأة ممكنة، بينما وحدة " الناقد" تعمل على مراقبة تأثير الفعل الذي قام به الفاعل. تُستخدم إشارة التنبؤ بالخطأ للفرق الزمني كإشارة تعليم لكل من وحدات الناقد والفاعل. تفترض النماذج الحالية لهيكلية الناقد والفاعل أن إشارة التنبؤ بالخطأ للفرق الزمني هي إشارة التعلم الوحيدة التي يمكنها التأثير على وحدتي الناقد والفاعل. خلال أطروحة الماجستير هذه، سنقوم بتقديم نسخة موسعة من معمارية الناقد والفاعل تتناول تأثير نوعين من إشارات التعزيز؛ إشارة التنبؤ بالفرق الزمني وإشارة تثبيط السلوك. نحن نفترض أن دور إشارة تثبيط السلوكية هو معارضة اشارة التنبؤ بالفرق الزمني من أجل التأكد من أهمية التعلم وتحصين الدمج. بناء على هذا المنطق، قمنا ببناء نموذج محوسب جديد لمنطقة في الدماغ تدعى العقد القاعدية. يتناول هذا النموذج تأثيرات النواقل العصبية الدوبامين والسيروتونين في عملية التعلم بالتحفيز من ردود الفعل الايجابية والسلبية. في هذا النموذج يتم محاكاة وظيفة الدوبامين بإشارة التنبؤ بالخطأ للفرق الزمني، في حين يتم محاكاة السيروتونين كإشارة تثبيط سلوكي يتمثل دورها في تثبيط إشارات التنبؤ بالخطأ للفرق الزمني. في هذا البحث استخدمنا مرض الاكتئاب السريري ومضادات الاكتئاب كمثال تجريبي لدراسة تأثير مستويات الدوبامين والسيروتونين وكذلك أثار التفاعل بينهما في التعلم عن طريق التحفيز الايجابي والسلبي. قمنا بنمذجة آثار الدوبامين والسيروتونين والتفاعل بينهم في التعلم عن طريق التحفيز الايجابي والسلبي باستخدام ثلاثة نماذج مختلفة: (1) نموذج اشارة التنبؤ بالفرق الزمني فقط، (2) نموذج إشارة التنبؤ بالفرق الزمني وإشارة التنبؤ بالمخاطرة، (3) نموذج إشارة التنبؤ بالخطأ للفرق الزمني

ونموذجنا المقترح الممثل بإشارة التثبيط السلوكي. تظهر نتائج المحاكاة أن نموذجنا المقترح الذي يدرس كلاً من إشارتي التنبؤ بالفرق الزمني وإشارة التثبيط السلوكي أظهر النتائج أفضل مقارنة بالنموذجين الآخرين. بالإضافة الى ذلك فإن نموذجنا قابل للاستخدام في عدة مجالات مثل بناء الروبوتات وأيضا في مجال علوم الأعصاب المحوسبة.

List of Contents:

List of Figures

List of Tables

List of Abbreviations

BG	Basal Ganglia
MDD	Major Depressive Disorder
DA	Dopamine
5-HT	Serotonin
CS	Conditional Stimulus
UCS	Unconditional Stimulus
GPe	Globus Pallidus Externa
GPi	Globus Pallidus Interna
STN	Subthalamic Nucleus
SNc	Substantia nigra pars compacta
VTA	Ventral tegmental area
DRN	Dorsal raphe nucleus
TD	Temporal difference
WTA	Winner Talks All
PFC	Prefrontal Cortex
RL	Reinforcement Learning
TD	Temporal Difference

Chapter One

Introduction

Section 1: Introduction

Section 2: Overview of Artificial Intelligence

Section 3: Problem Statement

Section 4: Motivation

Section 5: Thesis Objectives

Section 6: Research Questions

Section 7: Thesis Contribution

Section 8: Thesis Organization

Chapter One:

Introduction

1.1 Introduction

The core ideas of artificial intelligence algorithms are constructed based on human learning and neural mechanisms. Mathematicians and engineers copy the mechanisms of human neural circuits to implement the neural network theory (Figure 1.1) Furthermore, different algorithms in engineering and computer science are used to construct different kinds of neuro-computational models.

The starting point to link between neuroscience and computational reinforcement learning is linking the signals in the brain with signals playing prominent roles in reinforcement learning theory and algorithm. Scientist find that any problem in behavior learning models can be reduced to the three signals representing (1) action, (2) state and (3) reward. In addition to reward signals, scientists argue that there are other reinforcement signals such as value signals and signals conveying prediction errors. Each type of signals is labeled based on their functions in the reinforcement learning algorithm. On the other hand, signals in the brain refer to physiological events such as a burst of action potential or the secretion of a neurotransmitter. For example calling the phasic activity of a dopamine neuron a reward prediction error signal, means that the neural signal behaves like, and is conjectured to function as, the corresponding theoretical signal.[1] Although there are different types of reinforcement learning signals, there is no reinforcement signal that can play the inhibition role to other types of reinforcement learning signals given its phasic nature. However, in human brain there is a behavioral inhibition system that modulate activity in others systems as represented by the serotonergic pathway.

Computational neuroscience is an interdisciplinary field that links different fields of neuroscience, computer science, mathematics, and physics. It aims to construct computational models to investigate brain cognitive functions and explaining the unknown relationship between unrelated behaviors by implementing neural mechanism theory and artificial intelligence algorithms such as actor-critic architecture and temporal difference algorithm, [2]. These models are discussed in Chapter 2. In this thesis, we merge our knowledge of reinforcement learning and the actor-critic architecture with neuroscience and psychiatry to construct extended actor-critic model. This model addresses the effects

of behavioral inhibition modules by simulating behavioral inhibition signal that works to inhibit other reinforcement signals. In this model, this proposed signal inhibits the TD prediction error signal.

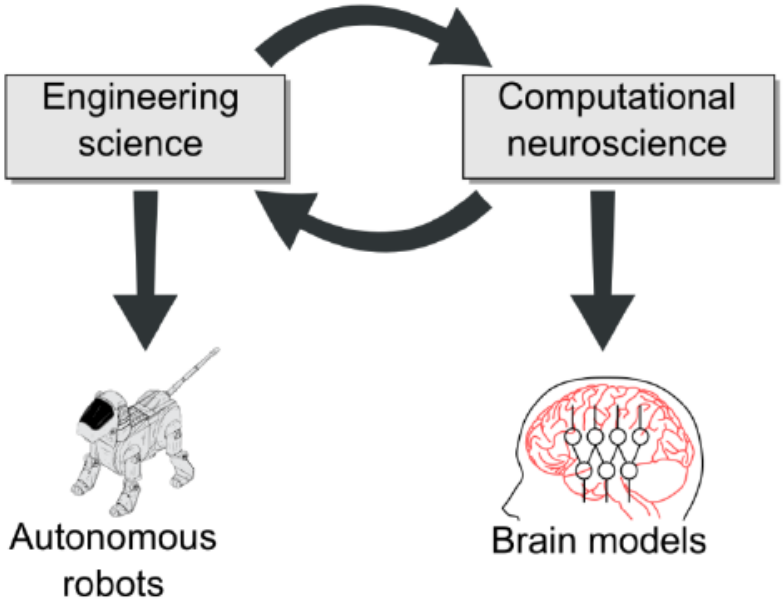

Figure 01.1 The relation between engineering and computational neuroscience showing how mathematicians and engineers copy the mechanisms of human neural circuits to implement the neural network theory and how they utilize these algorithms to construct brain models.

The majority of neurocomputational models use the actor-critic architecture to simulate the reinforcement learning process. The actor represents the policy structure while the critic represents the value function. The role of the critic is to strengthen or weaken the policy action. The temporal difference (TD) signal represents the output of critic which evaluates the quality of action selected. This signal is used to teach the actor and critic modules. In computational neuroscience, the TD signal represents the dopamine (DA) neurotransmitter signal that teaches the system modules[3]. In this thesis, we use an extended version of actor-critic architecture to simulate the effects of a behavioral inhibition brain signal, represented by serotonin (5HT), to construct a neurocomputational model of specified region in brain is called basal ganglia. Unlike earlier models, our model simulates the effects of both DA and 5HT neurotransmitter signals and their interaction. DA is represented as the TD signal and serotonin is represented as the behavioral inhibition signal.

A few studies try to study the contributions of other neurotransmitters signal such as 5HT signal in cognitive and reinforcement task. However, 5HT plays critical roles in controlling DA release, thus modulating impulsivity behavior and in behavioral inhibition. [4] The interaction between DA signal and 5HT signal plays a key role in normal and abnormal human behaviors. Therefore, understanding this interaction may help researchers reveal remarkable insights into the pathogenesis of various neuropsychiatric diseases such as major depressive. [5]

We employ three modeling approaches to study the cognitive correlates in major depressive disorder and the effects of antidepressants on cognitive function. The first model, proposed by Moustafa et al., 2010, used an actor-critic architecture based on the TD signal to simulate the DA signal. The second model, proposed by Balasubramani et al., 2014, represented the 5HT signal as a risk factor while DA was TD signal. Finally our proposed model which simulates 5HT as a behavioral inhibition signal and DA as a TD signal in the context of an extended actor-critic architecture.

1.2 Overview of Artificial Intelligence

In equation form, intelligence is the sum of knowledge and the ability to learn, perceive, feel, judge and understand. There are different definitions for artificial intelligence. One of them is a branch of computer science that simulates human behaviors to implement computational framework to reason like human. “Learning” is defined as the process that helps agents (learners) to improve their response in same task when it repeated at different times. Machine learning is a branch of artificial intelligence.[6] Its role is to adapt with new environments to detect and extrapolate patterns. In Chapter 2 of this thesis, we illustrate the different models and algorithms of machine learning that we used in this thesis.

1.3 Problem Statement

The problem with the current form of actor-critic architecture is that it assumes that no other reinforcement signals affect the TD signal. This assumption is unrealistic as there is electrophysiological evidence supporting that other reinforcement signals can modulate the TD signal. For instance, the behavioral inhibition signal can oppose the TD signal to increase the ultimate significance and applications of learning.

1.4 Motivation

The actor-critic architecture in its current form handles TD prediction error signal as a teaching signal for the system modules. There are different types of reinforcement or control signals may affect in learning modules at actor-critic models. For example, in feedback control systems, disturbance signals represent unfavorable inputs that affect the output of the control system and increase the system error. Moreover, in communication systems, noise signals are represented as high-frequency inputs which may cancel the output signal in transmission line. In reinforcement learning, there are different types of reinforcement signals that affect the learning process, but none of them exert an inhibitory role in the reinforcement learning process. Converging physiological evidences supports the existence of such an inhibitory signal in the reinforcement learning process. Here, our proposed model actor-critic addressed this critical gap in the actor-critic architecture by the implementation of a behavioral inhibition signal in reinforcement learning process. This is accomplished by constructing an additional module to carry out behavioral inhibition and phasically inhibit the TD signal.

In computational neuroscience few models studied the effects of the interaction between the DA and 5HT signals in the reinforcement learning process. Using our proposed model, we implemented a neurocomputational model which simulates the effects both DA and 5HT and their interaction in the generation of cognition. We represent DA as the TD signal and 5HT as the behavioral inhibition signal.

1.5 Thesis Objective

- Main Objective

  - Build extended version of actor-critic architecture, to handle the effects of the interaction between two reinforcement learning signals; TD signal and the behavioral inhibition signal.
  - Build computational model to study the effects of the interaction between dopamine and serotonin in major depressive disorder on reinforcement learning.

- Sub-objective

- Simulate the effects of selective serotonin reuptake inhibitor (SSRI) antidepressants on the interaction between dopamine and serotonin computationally.
- Simulate other learning paradigms in major depressive disorder to test the generalizability of the model.

1.6 Research Questions

At this thesis, we try to answer the following questions:

- What are the effects of the TD signal and behavioral inhibition signal in reinforcement learning process?
- What are the effects of the interaction between dopamine and serotonin in learning from reward and punishment feedback in patients with MDD?
- What are the effects of antidepressants SSRI in reinforcement learning process?

1.7 Thesis contribution

In this thesis, we proposed an extended actor-critic architecture; this model addresses the effects of other reinforcement signals such as behavioral inhibition signal in simulating TD prediction error signal. We assume that the behavioral inhibition signal opposes the TD prediction error signal and the resultant signal of both TD signal and behavioral inhibition signal works to learn the model modules. Our model unlike other theoretical models which are used only TD signals to simulate the reinforcement learning process.

Our extended version of actor-critic architecture will have significant potential applications in artificial intelligence systems, such as simulating inhibition systems and simulating other reinforcement learning signals.

In computational neuroscience modeling, the majority of models use the classical actor-critic architecture with TD signal which represents DA. Our extended actor-critic adds to the current literature the effects of 5HT signal in reinforcement learning process, which is represented as behavioral inhibition signal. We used our computational model of DA-5HT interaction to study cognitive function in patients with major depressive disorder given the relevance of both DA and 5HT dysfunction.

1.8 Thesis organization

Thesis chapters are organized as follows:

- **Chapter 1** includes an introduction about the topic of the thesis; an overview about artificial intelligence, neuroscience and neuro-computational modeling also it includes problem statement, motivation, thesis contribution and thesis organization
- **Chapter 2** includes background about different concepts of machine learning algorithms.
- **Chapter 3** includes background about different concepts of neuroscience that are related to our research
- **Chapter 4** presents review of some related studies of computational model for the basal ganglia, interaction between DA and 5HT, and major depressive disorder.
- **Chapter 5** focuses on the methodology we used for building the models.
- **Chapter 6** illustrates the results of model simulations.
- **Chapter 7** discusses modeling results and explains the limitations of the models.
- **Chapter 8** includes the conclusion of the thesis, discuss the main contributions of the thesis, and finally describe future work.

Chapter Two

Machine Learning Models

Section 1: Introduction

Section 2: Reinforcement Learning

Section 3: The TD Algorithm

Section 4: Actor-Critic Architecture

Section 5: Winners-Take-All Networks

Section 6: The Rescorla-Wagner Model

Section7: The Dynamics of Dopamine Neuronal Firing in Reinforcement Learning

Chapter Two:

Machine Learning Models:

In this chapter, we illustrate the main theoretical background of machine learning models which are used in this thesis.

2.1 Introduction

Machine learning is interdisciplinary field that use statistical algorithms to develop computerized frameworks that have the ability to "learn" by using different types of datasets.[7] Mainly, there are three different forms of learning according to the presentation of feedback in the learning system: supervised learning, unsupervised learning, and reinforcement learning.

In supervised learning, the role of the agent role is to find the function that matches examples from a sample set where each sample includes inputs and correct outputs. In unsupervised learning the agent tries to learn from patterns. But in reinforcement learning the agent doesn't have "knowledge" about the exact output for a particular input. Rather, it receives feedback signals which give indication about the quality of its behavior. [8] In this thesis, we focus on reinforcement learning models.

2.2 Reinforcement Learning (RL)

RL is a framework in which an agent (or controller) optimizes its behavior by interacting with its environment. The agent does not have any knowledge about the exact output for an input but it receives scalar feedback from the environment to indicate the quality of the action is takes.[6] Figure 2.1 describes the RL framework.

A good way for understanding the concept of RL is to consider real applications that utilized RL such as mobile robot. A mobile robot takes a decision to enter new room in order to collect a trash or return back to its station to recharge its battery. The decision is taken based on the current charge level of its battery and its past experience about how quickly and easily it can recharge its battery at this state [1]

The main goal of the agent is to find a policy that maximizes the total accumulated reward. [3]

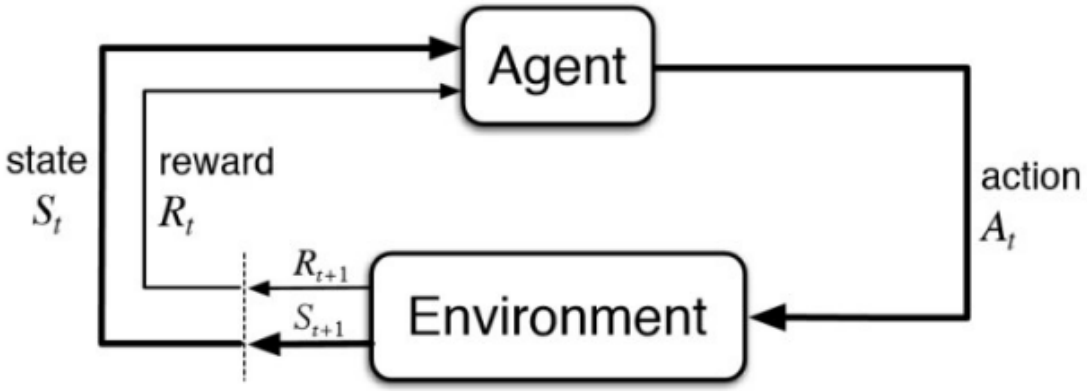

Figure 0.1 Reinforcement learning architecture, showing the interaction between the agent and the environment, with the ultimate goal being reward maximization.

In addition to the agent and the environment, a reinforcement learning system has four elements: (1) policy, (2) value function, (3) reward signal and (4) environment model. The policy is the core component of a reinforcement learning system. It defines the action learning agent takes in a given state. The reward signal is a feedback signal that indicates how good the immediate action was. The value function is the state-action pair used to estimate the goodness of action on the long run and update the policy structure. [9]

Knowing whether a current action gives reward or not when the action takes a long time is one of the challenges that face the implementation of reinforcement learning [3]. One approach to overcome this challenge is to use the temporal difference algorithm.

2.3 The Temporal Difference Algorithm (TD)

The TD algorithm is a bootstrapping approach that is used for prediction problems. [3] It is a combination of Monte Carlo and Dynamic programming ideas.[1] Dynamic programming refers to a collection of algorithms that can be used to estimate optimal policies given the perfect model of the environment, the key idea of DP and reinforcement learning is the use of value function to find the good polices. On the other hand, Monte Carlo methods solve the RL problem by estimating the value function using environmental sample sequences such as rewards, actions and states.

Both the Monte Carlo and TD methods can learn directly from new experiences without a model of the environment's dynamics. On other hand, TD methods like dynamic programming methods as both update estimates based in part on other learned estimates,

without waiting for a final outcome. The integration of the TD algorithm, Monte Carlo methods and dynamic programming represents the core idea of reinforcement learning.[10]

In reinforcement learning computations, the TD algorithm is used to estimate the value function, which is the action-state pair. This estimation aims to choose the action at a given state that maximizes the total reward.

2.4 The Actor-Critic Architecture

The majority of reinforcement learning and dynamic programming methods were categorized into critic only methods and actor only methods. Critic-only methods depend exclusively on value function approximation and its aim is learning the value function. On the other hand, actor-only approaches aim to find the optimal policy. The actor-critic approach combines both the critic-only and actor-only approaches in one architecture to estimate value function and maintain the actor policy. [11]

The actor-critic method can be viewed within the framework of control systems. The actor represents the policy structure. Its role is selecting the action that maximize the total system reward. Whereas the critic is represented as the value function. It plays a role in monitoring the quality of the action that is selected by the actor to strengthen or weaken the policy structure. The TD signal gives a feedback indication about the quality of the selected action. This signal is used to adapt the value function and update the actor policy[12]. Figure 2.2 shows the architecture of actor-critic method.

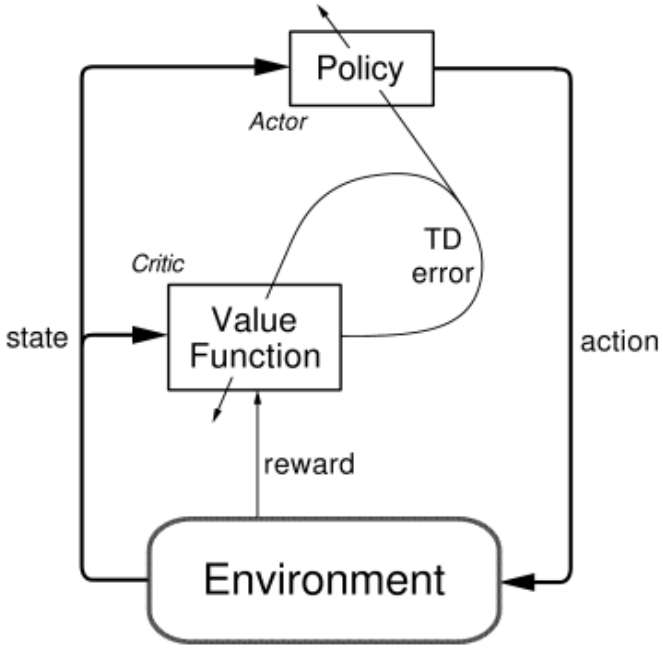

Figure 2.2 The actor-critic architecture, showing the interaction between the system modules and the environment. The TD signal is used to estimate value function and update the policy structure.

The actor-critic architecture has two significant apparent advantages:

1. It requires minimal computation cost in order to select actions.
2. It can learn an explicitly stochastic policy; that is, it can learn the optimal probabilities of selecting various actions

These distinctive properties of the actor-critic architecture make it very preferable in reinforcement-learning systems in real life, such as robotics and biological models. [11]

2.5 Winner-Take-All Networks

The approach that the brain utilizes for selective processing of a large number of inputs to maintains a unified perception it remains a mystery. In neuronal networks, a network in which all neurons respond the same to all stimuli would not transmit any information about the stimulus. In order to be useful, neurons must come to respond differentially to a variety of incoming signals. Different neural models and theories have been proposed to account for such ability. The winner-take-all network is one of the proposed mechanisms for developing action selection process through competition in simple recurrent networks.[13] Neural networks adopt a winner-take-all strategy from multiple layers of neurons: input

layer, hidden layers and output layer. In a winner-take-all network, the output layer nodes are in competition with the input layer nodes. Accordingly, the activation function uses the input signal of the input nodes to estimate the weight of the output node. The node that has the highest value of the activation function is declared the winner node.[14] This winner output node is moved closer to the input layer which are otherwise unchanged. Furthermore, the output nodes have lateral inhibitory connections. Therefore, the winner output node can inhibit other output nodes by an amount which is proportional to its activation level. As a result, one action can be selected at each stimulus [15] In the RL model, the input to the input nodes is a TD signal. This signal is used to estimate the weight of the output nodes.[14]

2.6 The Rescorla-Wagner Model

Classical conditioning is learning through association, in other words, two stimuli are linked together to produce a new learned response in a person or animal. The Rescorla-Wagner model describes the changes in associative strength (V) between a conditioned stimulus (CS) and the subsequent stimulus (unconditioned stimulus, US) as a result of a conditioning trial. [16] The concepts below are incorporated to formulate the mathematical basis of the Rescorla-Wagner model as follows:

- A change in the associative strength of a stimulus depends on the existing associative strength of that stimulus and all others present stimuli.
- If the existing associative strength is low, then the potential change is high. If the existing associative strength is high, then very little change occurs.
- The speed and asymptotic level of learning is determined by the strength of the CS and US.

The mathematical formula of Rescorla-Wagner is:

$$\Delta V_{cs} = c\,(V_{max} - V_{all)} \qquad (2.1)$$

V = associative strength

Δ = change (the amount of change)

c = learning rate parameter

Vmax = the maximum amount of associative strength that the UCS can support

Vall = total amount of associative strength for all stimuli present

Vcs = associative strength to the CS

Although the Rescorla-Wagner model is the best theory of classical conditioning, it fails to handle configurable learning without a tweaking and the implementation of latent inhibition.[10]

2.7 The Dynamics of DA Neuronal Firing in RL

Dopaminergic neurons behave similarly to the reinforcement learning model based on the TD algorithm. They encode the differences between the received and expected reward. This difference represents dopamine signal. We can define three different dynamics of prediction error signaling with reward[3] as shown in figure 2.3:

I. If there is no reward expectation then DA neurons fire in response to the reward which generate a positive prediction error signal. (Expected reward = 0 ➔ prediction error is positive).
II. If there is a CS related to the reward, then neurons don't fire for reward itself but instead fire in response to the CS which doesn't generate a reward signal. In other words, the prediction error signal equals 0 (Expected reward = Obtained reward).
III. If the reward is omitted after CS, DA neuronal firing dips down. This generates a negative prediction error signal. (Expected reward is positive and obtained reward = 0).

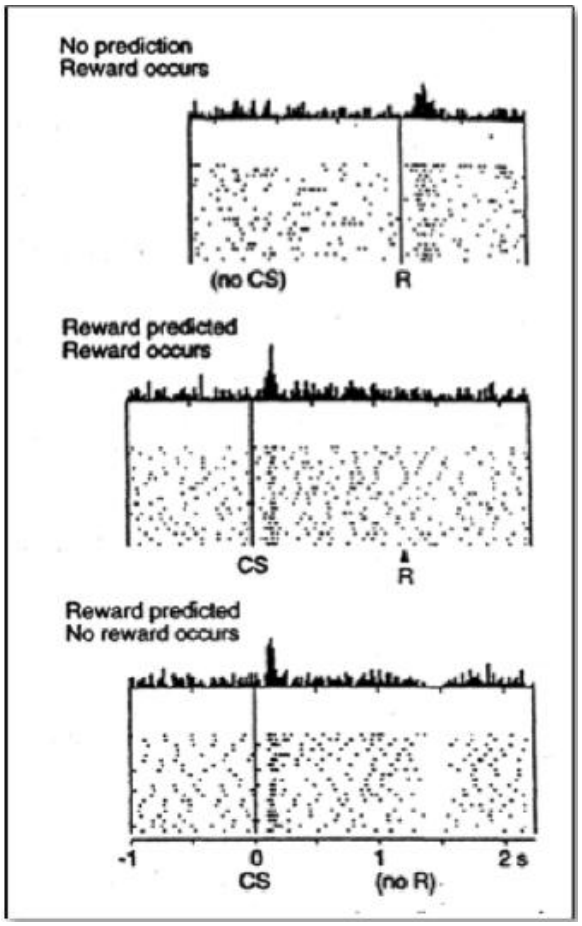

Figure 02.3 The Dynamics of DA neurons firing. The first part describes the DA firing when unpredicted stimuli is occurred. In this case, DA neurons are firing a positive prediction error signal. The second part describes the DA firing when the predicted stimulus occurred. Here, DA neurons don't respond to the reward and prediction error signal equal to zero. The third part describes the DA firing when reward is predicted but doesn't occurs. DA neurons produce a negative prediction error signal. [1]

Chapter Three

Neuroscience Background

Section 1: Introduction

Section 2: Neurons, Synapses, and neurotransmitters

Section 3: DA system

Section 4: 5HT system

Section 5: The Interaction between DA and 5HT

Section 6: The Basal Ganglia and Its Neurochemical Pathways

Section 7: Major Depressive Disorder and Selective Serotonin Reuptake Inhibitor Antidepressants

Chapter Three:

Neuroscience Background:

Here, we describe the definition of neuroscience. Also, we illustrate the main brain regions which are related to our research. Further, we provide background about neurons, neurotransmitters and major depressive disorder.

3.1 Introduction

Neuroscience is the multi-disciplinary study of the nervous systems. It investigates how neurons regulate human functions; control behavior; learning, and aging; and how cellular and molecular mechanisms collaborate to perform these functions.[17] For machine learning, one of the most exciting aspects of neuroscience lies in all the supporting evidence showing that humans and animals implement RL approaches to maximize learning and outcomes.

3.2 Neurons, Synapses, and Neurotransmitters

The cells in nerve system are called neurons. These cells are specialized to transmit information using electrical and chemical signals. A synapse transmits information from the presynaptic neuron's axon to a dendrite or cell body of the postsynaptic neuron.

Neurotransmitters are chemical messengers that carry information signal across synapses. They are released from presynaptic neurons. Each neurotransmitter has its own receptors located on the receiving neurons (postsynaptic). Neurotransmitters transmit a chemical signal upon the firing of the presynaptic neuron to activate or inhibit the postsynaptic neuron. The most common neurotransmitters in the brain include: glutamate, GABA, dopamine, and serotonin. [18] In following section, we will discuss in detail the dopaminergic system and serotonergic system.

3.3 The DA System

DA is a neurotransmitter which is represented as a training signal in computational model modules that correspond to specific brain areas. DA cells are mainly located in the midbrain, in the ventral tegmental area (VTA) and the substantia nigra pars compacta

(SNpc). The mechanism for regulating DA release into subcortical regions occurrs via two independent mechanisms:

1. Transient or phasic DA release caused by DA neuron firing.
2. Sustained, “background” tonic DA release regulated by prefrontal cortical afferents. [19]

DA neurons have two major classes of receptors: D1 receptors and D2 receptors. The D1 receptor family is primary excitatory; it activates postsynaptic neurons. Conversely, the D2 family is inhibitory; it suppresses postsynaptic neurons.

Different studies suggested that DA neuronal firing represents a prediction error signal for unexpected rewards. Further, the decay of DA neurons produces a selective deficit in learning from reward. [6] DA dysfunction has been implicated in various neurological and psychiatric disorders such as Parkinson’s disease, Huntington’s disease, and major depressive disorder. [20]

3.4 The 5HT System

5HT neurons are located in near midline of the brain stem. Ascending nuclei projecting to the forebrain mainly comprise the median raphe nucleus (MRN) and dorsal raphe nucleus (DRN).[21] 5HT neurons have at least 17 subtypes of receptors, thus giving 5HT neurons a highly interactive and complex range of effects [21]. 5HT plays a critical role in modulating DA release. 5HT is implicated in impulsive behavior as well as in behavioral inhibition.[22]

3.5 The Interaction between DA and 5HT

The interaction between DA and 5HT plays a key role in normal and abnormal human behaviors. Hence, understanding this interaction can reveal remarkable insights into the pathogenesis of various neuropsychiatric diseases such as major depressive disorder, or neurological conditions such as Parkinson’s disease.[21] Although multiple studies attempted to understand this interaction, it remains unclear how DA and 5HT interact to produce cognitive function for many reasons: [22]

1. The rich and widespread 5HT and DA innervations of in the brain.
2. The large number of 5HT receptors which has about 17 subtypes.
3. The release of co-transmitters by 5HT and DA neurons.

There are many lines of evidence to support the bidirectional relationship between 5HT (from the DRN) and DA (from the SNpc). Neuroanatomical data clearly elucidate that DA neurons in the SNpc receive prominent innervations from DRN 5HT neurons. Moreover, this interaction seems to be reciprocal where DRN 5HT neurons also receive innervations from SNpc DA neurons. [21] Electrophysiological studies support this bidirectional relation where the stimulation of DRN 5HT neurons induced a significant decay in the phasic firing of the SNc DA neurons. Also, by using retrograde tracing techniques, researchers observed a dense direct projection from SNpc DA neurons to DRN 5HT neurons. [23]

DRN 5HT neurons exert both excitatory and inhibitory control on ascending DA pathways. This can be further regulated by the type of 5HT receptors on DA neurons. [24] Based on this, and for the purposes of producing a parsimonious understanding the DA-5HT interaction, we will categorize 5HT receptors into two crude categories: excitatory receptors and inhibitory receptors. Table1 shows the classification of receptors families. Matias and colleagues found that 5HT neurons were sensitive to both positive and negative DA prediction error and they fire [surprise signal] to oppose the DA signal. Accordingly, we can conclude that the role of 5HT is to oppose DA firing. In other words, 5HT functions to "inhibit" the behaviors which are encoded for by DA neurons.[25] With the complexity of the phasic DA signal, any opposition by 5HT should match the generation of the phasic signal in a precise manner to counter the DA effect. In our research, we used this finding to simulate the interaction between the DA and 5HT, where the 5HT signal represents a "behavioral inhibition" module to oppose the critic policy structure. This extends the current actor-critic architecture in a novel way to account for a phasic inhibitory signal that opposes learning and increases stochasticity.

Table 1: The categories of 5-HT receptors

Family	**Potential**
5-HT1	Inhibitory
5-HT2	Excitatory
5-HT3	Excitatory
5-HT4	Excitatory
5-HT5	Inhibitory

5-HT6	Excitatory
5-HT7	Excitatory

3.6 The Basal Ganglia (BG) and Its Neurochemical Pathways

The BG are a set of subcortical nuclei that represent one of the brain's fundamental processing units for motor control system and cognitive function. They control different functions in brain such as motivation, decision taking and working memory. The BG consist of different nuclei: the striatum, the subthalamic nucleus, the external globus pallidus (GPe), the substantial nigra pars reticulata (SNPr) and the internal globus pallidus (GPi). [26]

The striatum receives input from the cortex and dopaminergic projection from the SNpc. These projections control the value associated with the reward signal. The striatum sends projections to the GPe, STN, GPi and SNpr via the direct and indirect pathways. In the striatum, medium spiny neurons with the D1 DA receptor project to the GPi via the direct pathway, while medium spiny neurons carrying the D2 DA receptor project to the GPe via the indirect pathway. Figure 3.1 shows direct and indirect pathways of the BG

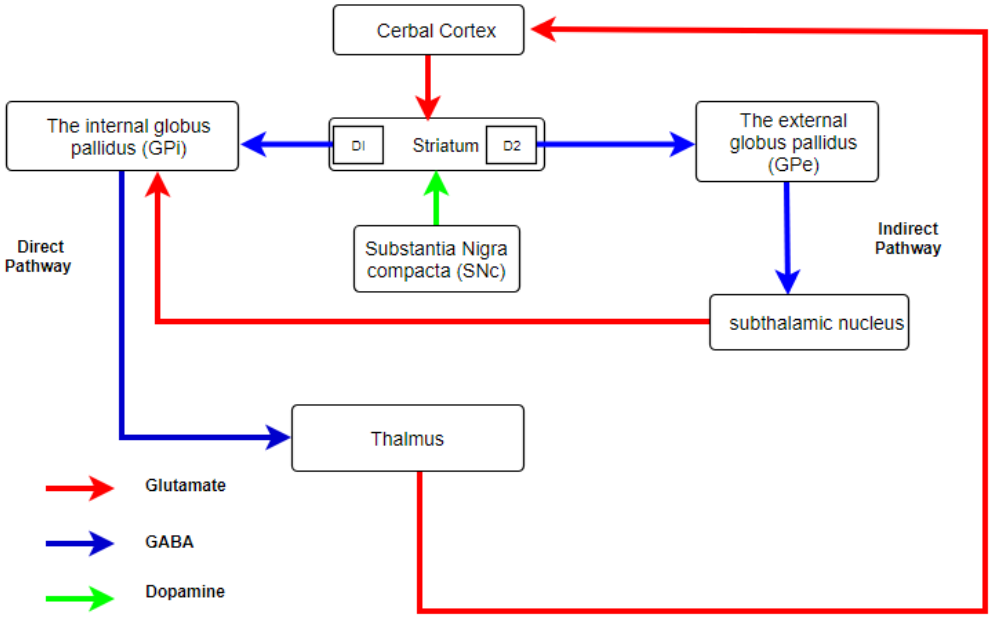

Figure 03.1 The direct and Indirect pathways in the BG. The striatum receives DA projections from the SNpc. The D1 and D2 receptors in the striatum control the direct and indirect pathway, respectively.

Neurons with the D1 receptor project to the GPi via the direct pathway, while neurons with the D2 receptor project to GPe via the indirect pathway.

3.7 Major Depressive Disorder (MDD) and Selective Serotonin Reuptake Inhibitor (SSRI) Antidepressants

MDD is a debilitating psychiatric disorder; it causes set of psychophysiological changes (in appetite, sleep, loss of ability to experience pleasure, and suicidal thoughts), which affect the patient's daily life activities. [27] Many studies suggest that the dysfunction of the 5HT and DA systems can lead to the development of various neuropsychiatric disorders such as MDD; patients with MDD have low concentration in both dopamine and serotonin, cognitive studies show that patients with MDD are anti-learned from positive feedback and are learned from negative feedback. [28]

SSRI antidepressants represent the first line treatment for MDD. These pharmacological agents block the reuptake of 5HT from the synapses, thus leading to an increase in the concentration of 5HT in the brain. [29]

Chapter Four

Literature Review

Section 1: Neurocomputational Models of the BG

Section 2: Neurocomputational Models of DA and 5HT

Section 3: Neurocomputational Models of MDD

Chapter Four:

Literature Review:

Here, we review multiple neuro-computational models of the BG to discuss different computational approaches of the interaction between DA and 5HT. Further, we review some computational studies of MDD.

4.1 Neurocomputational Models of the BG

Different neuro-computational models have studied the BG DA and its contributions in several cognitive functions. The majority of these models simulate the DA signal by using the TD algorithm and RL models. [2]

Gurney et al., presented a computational model of BG based on its anatomy. They suggest that the main role of BG is action selection. They encode the selection signal in scalar variable. This model used a neural network mechanism to examine the action selection process by mapping each node in the neural network to BG anatomy. [30]

Moustafa and Gluck proposed a neuro-computational model of DA and prefrontal–striatal interaction during feedback-based category learning in Parkinson's disease. In this model, they used the actor critic-architecture where the critic plays an important role in feedback-based learning and the actor is essential for action-selection learning. The critic sends a teaching signal (DA) to the actor to strengthen or weaken action selection learning. The TD algorithm is used to train the model.[31] In a subsequent paper, Moustafa et al. proposed a neuro-computational model to study the cognitive effects of levodopa and DA agonists in patients with Parkinson's disease. In this model, they used the actor-critic architecture with four modules: input, DA, motor, and cognitive. The TD signal was used to train the model. This model assumes that the BG DA signal is key for motor and reinforcement learning, where the PFC DA signal essential for stimulus selection learning. Furthermore, they used this model to explain the effects of DA agonists and levodopa on

motor and cognitive process in patients with Parkinson's disease. This model find levodopa enhances stimulus-response learning, while DA agonists impair this type of learning.[32]

Natsheh proposed a neurocomputational model to study the effect of TD predication error signal variations on reinforcement learning. She studied the interaction between the action selection and the action execution modules. She used the BG as the action execution module and prefrontal cortex as the action selection module. [33] to dissociate the effects of naturally-occurring genetic polymorphisms of the DA system on cognitive function. This model simulates the contributions of variations in the parameters of DAT1 (9R,10R) and COMT (Val, Met) genes in learning from reward and punishment feedback. Natsheh proposed that the variations of DAT1 gene affect the BG DA while the variations ofCOMT gene affect the prefrontal cortex DA. This model found that learning from reward feedback is governed by an inverted U-shaped function according to prefrontal cortex DA availability.

4.2 Neurocomputational Models of DA and 5H

Daw et al., 2002 suggested a new approach in which 5HT might act as a motivational opponent to DA system in cognitive function. They used a theoretical framework of average-case RL where they represented the 5HT signal as a long-run average reward rate to create a tonic opponent to the phasic DA signal. They hypothesized that DA, in turn, might report the long-run average punishment rate as a tonic opponent to a phasic 5HT signal.[34]

Doya presented a computational theory on the role of the ascending neuromodulatory inputs, such as the 5HT and DA systems, in the global teaching signal that controls learning mechanism in brain. As per Doya's theory, DA firing represents a prediction error signal while the 5HT signal is controls the time scale of prediction error. In the his RL model, Doya simulated DA as the TD signal and 5HT as a discount factor which determined the time in the future when the agent should consider reward prediction and action selection. Therefore, by controlling the discount factor one can manage the time scale of the reward prediction. [35]

Read and his colleagues suggested that the role of the 5HT projection from the DRN to the striatum is to control the balance between the direct and indirect pathways. [36]

Balasubramani et al., 2015 used a computational model to simulate the contributions of 5HT and DA in risk-based decision making, reward prediction, and punishment learning. They used a modified RL framework where DA represented the TD error as in most extant literature of DA signaling and RL, and 5HT controlled an additional risk prediction error. In this model they linked between the risk sensitivity and 5HT function to construct extended RL framework, they used the utility function in policy execution instead of value function. The utility function combined the value function and risk function and produced the weight value α to represent 5HT cognitive functions in BG.[37]

Another BG model is the one introduced by Balaraman et al., 2015. They developed an RL model of the BG to understand impulse control disorder in Parkinson's disease patients, where ICD is a multi-factorial problem that implies the tendency to act prematurely. This proposed model of the BG uses to mimic the impulsivity behavior in patients with Parkinson; it includes the anatomical modules of BG such as the striatum, GPe, GPi and STN. In addition to these modules, the model also addresses the role of neurotransmitters DA and 5HT in impulsivity; DA is represented as TD signal whereas 5HT is represented as risk prediction factor as reviewed above. This model used the utility function to model the action selection process and the associated reaction time. While neurons with the D1 receptor computed the value function, neurons co-expressing the D1 and D2 receptors computed the risk function. The neurons of striatum module project through direct and indirect pathways to the output nuclei at BG system, GPi. Then the GPi relays the signal to thalamus. The wining thalamic neuron represented the selected action. [38]

4.3 Neurocomputational Models of MDD

Although still in its infancy, different computational approaches have contributed virtually to our understanding of the pathophysiology of MDD. The mathematical analysis of RL allowed for clear testable predictions about behavior and learning, thus helping researchers to link behavior and learning in order to understand the processes that are affected by MDD. Moreover, the computational approaches of RL identify different parameters such as prediction error, learning rate and reward sensitivity to understand the effects of MDD.[39]

Different studies use mathematical analysis of RL to understand cognitive function in MDD. Herzallah et al. 2013 studied the effects of MDD and 5HT antidepressants on cognitive function using a reinforcement learning task that dissociates positive and negative feedback. They found that unmediated patients with MDD learned from punishment feedback, but not from the reward feedback. Conversely, medicated MDD patients with receiving SSRI antidepressants did not learn from either reward or punishment feedback. In this thesis, we used the same cognitive data set in this paper to build our model. [28] Gradian and his colleagues studied the effects of MDD and schizophrenia on prediction error and expected values by using RL model and functional MRI. Both studies found that there is an abnormality in prediction error in MDD.[40]

Most studies used RL mathematical analysis to study MDD, but there were limited modeling attempt to understand disease mechanisms in the context of the interaction between DA and 5HT within the BG.[39] Unlike previously reviewed models, we study the effects of the interaction between DA signal and 5HT signal in a BG model and its role in learning reward and punishment feedback. Our novel model used an extend actor-critic architecture, where the 5HT signal is represented using a behavioral inhibition module. This model can simulate the role of 5HT as a behavioral inhibition signal while actively interacting with the DA critic module to produce the TD signal.

Chapter Five

Methodology

Section 1: Proposed Methodology

Section 2: The Computer-Based RL Task

Section 3: Model Architectures

Section 4: MDD Modeling

Section 5: Tool Used

Section 6: Human Experimental Dataset

Section 7: System Requirement

Section 8: System Testing

Chapter Five:

Methodology:

Here, we illustrate the different aspects of our models. First, we introduce the modeling approaches we used. Then, we describe the computer-based cognitive task. Finally, we explain the model architectures and their different modules.

5.1 Methodology: An Overview

In this thesis, we proposed an extended version of actor-critic architecture. We added a new behavioral inhibition module to generate a phasic signal to oppose the TD signal. In fact, the idea of this model is based on biological evidences that support the role of 5HT signal in behavioral inhibition and its opponency with DA.

To simplify our modeling approach, let us consider an example. Suppose you received a gift from an unexpected person and without any occasion. This event will make you surprised and happy, do you know what would happen in your brain at this moment? The DA neurons are firing, thus the direct pathway in BG is activated and the cognitive functions such as feeling happy are at work. But what is the role of 5HT in this story? It inhibits this feeling of happiness. When DA neurons fire, they send input to 5HT neurons. This is followed by a behavioral inhibition signal generated by the 5HT neurons that oppose the DA phasic signal. The resultant TD signal is the sum of the DA and 5HT signals which ultimately acts on the direct and/or indirect pathways to control the inhibition and the excitation of the system.

Our proposed model is based on the actor-critic architecture, where the role of the critic is action selection, and role of the actor is action execution. The TD algorithm is used for model training by simulating different characteristics of DA firing and its role in reinforcement learning process. The behavioral inhibition module is proposed to simulate the 5HT firing and its role in inhibiting the TD signal.

In this thesis, we implement three modeling approaches of BG to study the cognitive correlates in patients with MDD:

- An actor-critic architecture and TD algorithm approach proposed by Moustafa et al 2010. [32]

- An actor-critic architecture, TD algorithm, and risk prediction approach proposed by Balasubramani at el 2015. [38]
- A novel extended actor-critic, TD algorithm, with behavioral inhibition module approach.

5.2 The Computer-Based RL Task

Imaging and animal studies suggested that different brain structures such as BG are involved in category learning. Thus this task was used to examine feedback-based category learning as it relates to BG function. [28]

This computer-based cognitive task utilizes category learning. In addition, it also tests if the subject learns either from positive or negative feedback. Subjects were asked to predict if a stimulus belonged to a rain or sun category. On each trial, one of four stimuli appeared on the screen and the subject chose if this stimulus belongs to the rain or sun category (figure 5.1). The four stimuli in this task, two were rewarded and two were punished. If the stimulus was rewarded and the subject response was correct then the subject would win 25 points (Figure 5.1, C). If the stimulus is rewarded and the subject response was incorrect, the subject will not get any feedback. Conversely, if the card was punished and subject response was correct, the subject will not get any feedback. But if the response is incorrect, the subject will lose 25 points (Figure 5.1, D). Each block has 40 trials with a total of 160 trials across the four blocks. Subjects learned to categorize the stimuli into the rain or the sun categories. This task is simulated by our model to represent the input.

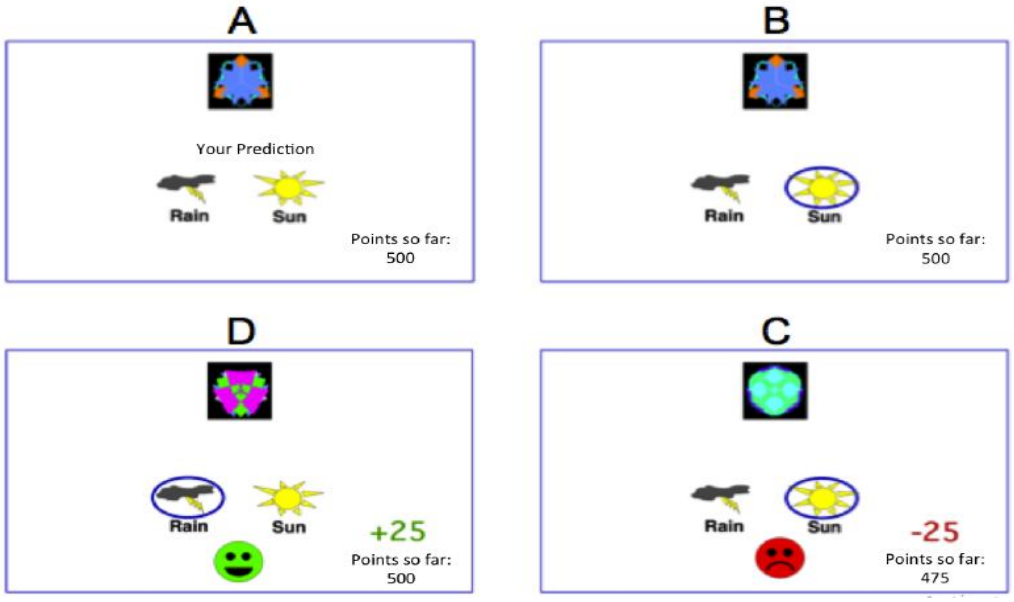

Figure 5.1 The RL probabilistic classification task. (A) on each trial the participant saw one of four cards and was asked whether this card predicts rain or sun. (B) if the card is rewarding and subject response is incorrect also if the card is punishing and subject response is correct, in both cases the subject is given no feedback. (D) For rewarding card, correct responses get rewarded with 25 points winning. (C) For punishing card, incorrect responses get punished with 25 points losing

5.3 Model Architectures

To examine the effectiveness of our proposed extended actor-critic architecture in building a neuro-computational model of the interaction of DA and 5HT in the BG, we tested three different modeling approaches. The first model used an actor-critic architecture and the TD algorithm as proposed by Mustafa et al., 2010. This model studied the effects of the DA signal in the BG to modulate cognitive function. The second model was also based on an actor-critic architecture, the TD algorithm and an accompanying TD for risk prediction. It was proposed by Balasubramani at el., 2015. This model investigates the effects of both DA and 5HT in BG, where DA represented the TD signal and 5HT is represented the risk prediction TD. The third and last approach was our proposed model where we utilize an extended actor-critic architecture, the TD algorithm, and behavioral inhibition module that

represents the 5HT signal and generates an opposing phasic signal to that produced by the DA critic.

In all modeling approaches, the actor represents the action selection network and critic represents feedback learning. The critic sends the TD signal to the actor in order to monitor the quality of the action selected by the actor. The core model is based on the RL framework and TD algorithm.

Approach#1: Actor-critic architecture and TD algorithm

In this approach, the model architecture is shown in figure 5.2; the DA signal is represented as the TD signal. This BG model has four modules: (1) the action selection network, (2) the action execution network, (3) the TD prediction error (DA) signal, and (4) the input module.

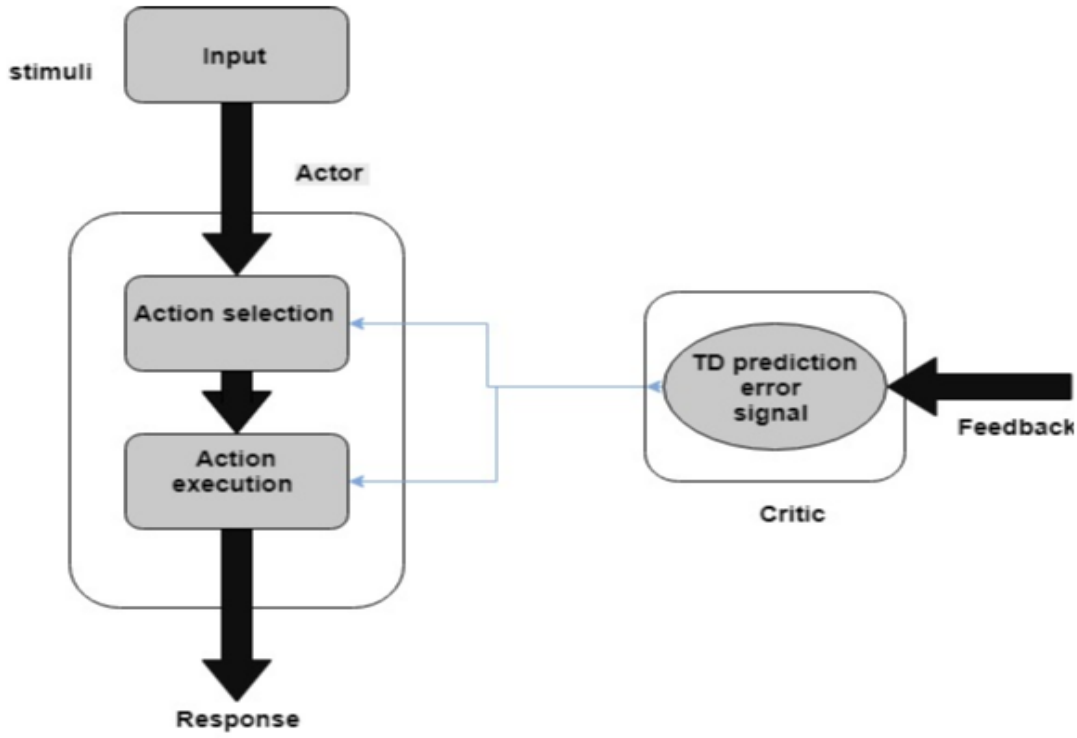

Figure 0.2 The actor-critic model architecture. The model has four modules: input, action selection, action execution and TD prediction error signal. The actor consists from action selection and action execution. Action selection represents the striatum while action execution represents the other anatomical components of BG (GPi, GPe, STN). The TD prediction error signal represents the critic. The critic sends TD signal to learn action selection and action execution modules. The input module sends projection to action selection module. The action selection module is connected to the action execution module. The activation node at input layer represents one of four stimuli in reinforcement task. The stimuli are presented in the action selection module. The activation node in action execution module represents the selected action.

The TD signal is driven as results of the action execution network from rewards and punishments; the TD signal represents the DA firing. The action selection module is connected to the action execution module. Also, both the input and the action selection modules have the same number of nodes. Each node in input module represents a stimulus that is presented in the action selection network. Input patterns are presented to activate the network in the input module. The input module sends projections to the action selection module. A winner-take-all network is used to simulate the connectivity within the action selection module. The action execution module learns to map the input stimuli to their responses. The winner-take-all algorithm is used to enhance connections within the action execution network as the wining node represents the action response. The value function estimates the expected sum of future rewards obtained by executing actions.[31]

The Modules

As mentioned above, all modeling approaches used four modules an action selection network, an action execution network, a TD prediction error signal (DA) signal, and an input module. The input to the model is the reward/punishment signal from the aforementioned cognitive task. The input is sent to the action selection module to perform the action selection process. Subsequently, it passes the selected action to the action execution network in order to select the response via the direct or indirect pathway. Based on the selected action if its reward or punishment, the TD module will generate a TD signal.

Reinforcement learning computations

Here, the mathematical formulas of models are discussed.

Action Selection Module (Striatum module)

The output of medium spiny neurons (D1R, D2R) which represents the DA receptors was represented by the variables y_{D1} and y_{D2} as follows:

$$\boldsymbol{y_{D1,t}(s_t, a_t) = w_{D1}(s_t, a_t)x(s_t)} \quad (5.1)$$

$$\boldsymbol{y_{D2,t}(s_t, a_t) = w_{D2}(s_t, a_t)x(s_t)} \quad (5.2)$$

Where x is modeled equal 1 at current state, and t denoted trial. Value function was used to estimate the expected future reward:

$$\boldsymbol{Q_t(s_t, a_t) = y_{D1,t}(s_t, a_t)} \tag{5.3}$$

After the action is selected, the model received feedback to update the weights of category. The updated weight equations for D1R and D2R nodes can be computed as follow:

$$\boldsymbol{\Delta w_{D1}(s_t, a_t) = \eta_{D1}\lambda_{D1}(\delta(t))X(s_t)} \tag{5.4}$$

$$\boldsymbol{\Delta w_{D2}(s_t, a_t) = \eta_{D2}\lambda_{D2}(\delta(t))X(s_t)} \tag{5.5}$$

Where $\boldsymbol{\eta}$ is the learning rate for each neuron, and $\boldsymbol{\lambda}$ is activation function for D1R and D2R, the values of activation function were computed as follows:

$$\boldsymbol{\lambda_{D1}(\delta) = \frac{2c1}{1 + \exp(c2(\delta + c3))} - 1} \tag{5.6}$$

$$\boldsymbol{\lambda_{D2}(\delta) = \frac{2c1}{1 + \exp(c2(\delta + c3))} - 1} \tag{5.7}$$

The $\boldsymbol{\delta}$s in the weight updates equation represents the classical TD error which simulates the immediate reward for activity update. It was calculated as in the equation below:

$$\boldsymbol{\delta(t) = r - Q_t(s_t, a_t)} \tag{5.8}$$

For the action selection purpose, the TD signal was calculated as follows:

$$\boldsymbol{\delta_{Q(t)} = Q_t(s_t, a_t) - Q_{t-1}(s_t, a_{t-1})} \tag{5.9}$$

The Action Execution Module

The main role of action execution module is to control the motor response in the selection process. It includes different anatomical components to simulate the projection of the signal through the BG direct and indirect pathways.

The direct and indirect pathway send projections to the GPi. The network of GPe-STN includes the same number of nodes for STN and GPe, where each node in the STN has bidirectional connections to other nodes in the GPe. The computations between STN and GPe followed the equations below: [38]

$$\tau_s \frac{dx_i^{STN}}{d_t} = -X_i^{STN} + W^{STN} y_i^{STN} - X_i^{GPe} \tag{5.10}$$

$$y_i^{STN} = \tanh(\lambda_i^{STN} X_i^{STN}) \tag{5.11}$$

$$\tau_g \frac{dx_i^{GPe}}{d_t} = -X_i^{GPe} + W^{GPe} x_i^{GPe} + y_i^{STN} - X_i^{IP} \tag{5.12}$$

Where X_i^{STN} and X_i^{GPe} are the internal state representation of the ith node in STN and GPe respectively.

$\mathbf{W^{GPe}}$: Lateral connections with GPe. It is set to ϵ_g node, which is equal to a negative number for all connections of GPe neuron.

$\mathbf{W^{STN}}$: Lateral connections with STN. It is set to $\epsilon_e$number, which is equal to a positive number for all connections of STN nodes. Constants ϵ_g and ϵ_e denote connection strength in each lateral connection for GPe and STN nodes respectively.

We set ϵ_g = - ϵ_e and ϵ_e . With learning rates $\frac{1}{\tau_s}$= 0.1 and $\frac{1}{\tau_j}$ =0.03, the slope of λ_{STN}= 0.7.

The transition of D1R neurons output via the direct pathway was computed as follows:

$$\boldsymbol{X_t^{DP}} = \boldsymbol{\lambda_{D1}}(\boldsymbol{\delta_Q(t)}\boldsymbol{y_{D1,t}}(\boldsymbol{s_t}, \boldsymbol{Q_t}) \quad (5.13)$$

While the transition of D2R nodes output via the indirect pathway was:

$$\boldsymbol{X_t^{IP}} = \boldsymbol{\lambda_{D2}}(\boldsymbol{\delta_Q(t)}\boldsymbol{y_{D2,t}}(\boldsymbol{s_t}, \boldsymbol{Q_t}) \quad (5.14)$$

Action selection at the GPi was computed as a combination of the direct and indirect pathways, the equation that follows represents the response of GPi:

$$\boldsymbol{x_t^{GPi}} = \boldsymbol{X_i^{DP}} + \boldsymbol{w_i^{STN-GPi}}\boldsymbol{y_i^{STN}} \quad (5.15)$$

While the action selection response in the thalamus can be computed as follows:

$$\boldsymbol{x_t^{Th}} = \boldsymbol{X_i^{DP}} + \boldsymbol{w_i^{STN-GPi}}\boldsymbol{y_i^{STN}} \quad (5.16)$$

We compute the activation of thalamus neurons as follows;

$$\frac{dy_i^{Th}}{dt} = -y_i^{Th} + x_i^{Th} \quad (5.17)$$

Where $\boldsymbol{w_i^{STN-GPi}}$ simulates the relative weight of projections from the STN to the GPi. It was set to 1 for all nodes in the simulation, y_i^{Th} is the state of i_{th} thalamus neuron, the action with maximumy_i^{Th} at that time is selected

Approach#2: Actor-critic architecture, TD algorithm and risk prediction

This modeling approach is proposed by Balasubramani et al., 2015. They suggested that the 5HT signal is represented by a risk prediction error. This model was extended from Approach #1.

RL computation

Here, we describe the modifications on the modules from Approch#1.

The Action Selection Module (Striatum module)

In Approach #1, D1R and D2R neurons received only DA projections. But in this model, both D1R and D2R received 5HT projections as well. The output of co-expressing D1R–D2R neurons was computed as follows: Figure 5.3 shows the striatum architecture

$$\boldsymbol{y_{D1D2,t}(s_t, a_t) = w_{D1D2}(s_t, a_t)x(s_t)} \tag{5.18}$$

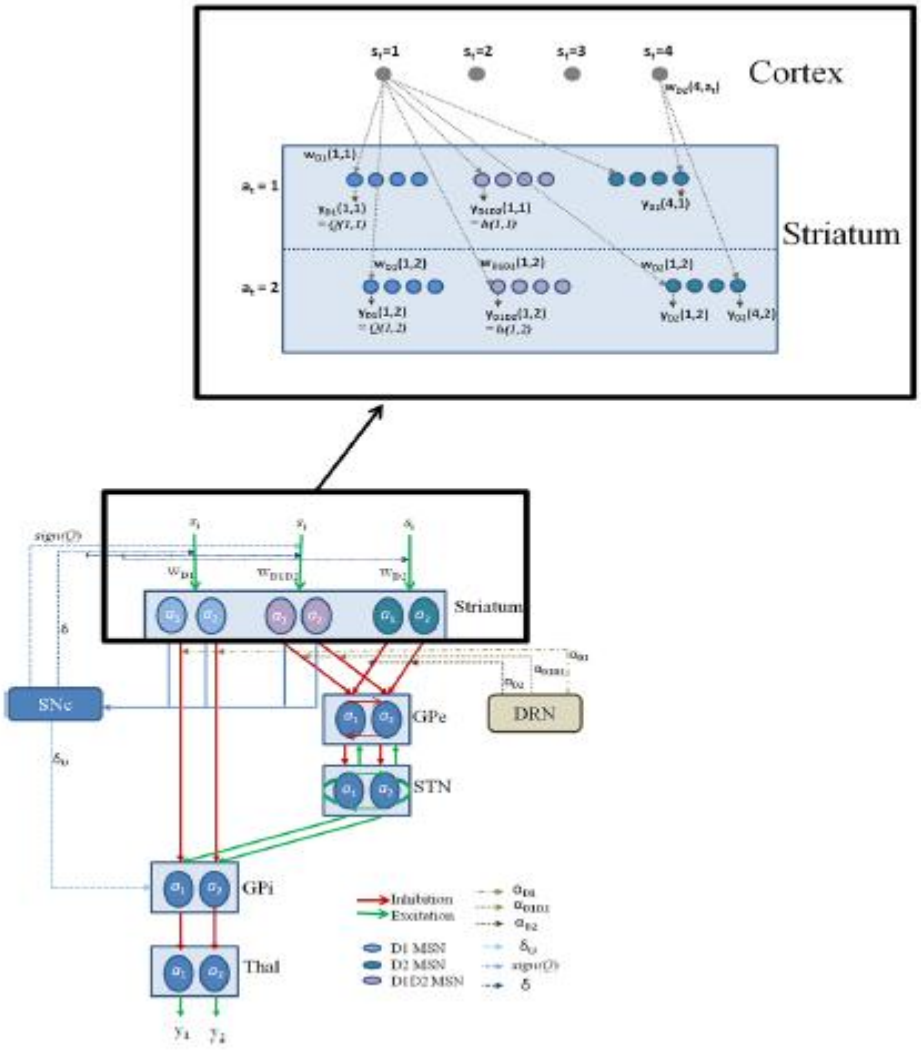

Figure 5.3 A complete schematic model of BG. The BG model components are the striatum, GPe, GPi, and STN along with SNpc, DRN, and thalamus. This model studies the contributions of DA and 5HT in the direct and indirect pathways. It proposes that the striatum has three kinds of the striatal neurons; D1R, D2R and D1R-D2R. D1R represents D1 receptor and it plays critical role in learning from reward while D2R represents D2 receptor and it controls learning from punishment. Finally, D1R-D2R is used to simulate the role of 5HT in risk prediction. Adapted from Balasubramani et al., 2015[38]

The utility function is combined from both the value function and the risk sensitivity function as follows:

$$\boldsymbol{U_t\,(s_t\,,a_t)} = \boldsymbol{Q_t(s_t\,,a_t)} - \boldsymbol{\alpha_{D1D2}\; sign\big(Q_t(s_t\,,a_t)\big)\sqrt{h_t}(s_t\,,a_t)} \tag{5.19}$$

Where

$$\boldsymbol{h_t(s_t\,,a_t)} = \boldsymbol{y_{D1D2,t}\,(s_t\,,a_t)} \tag{5.20}$$

To update the weight of D1R–D2R MSN neurons, the following equation was used:

$$\boldsymbol{\Delta w_{D1D2}\,(s_t\,,a_t) = \eta_{D1D2}\lambda_{D1D2}\big(\delta(t)\big)X(s_t)} \tag{5.21}$$

The activation function of D1R–D2R MSN neurons $\boldsymbol{\lambda_{D1D2}}$ was computed as follows:

$$\boldsymbol{\lambda_{D1D2}(\delta) = \lambda_{h-D1}(\delta) + \lambda_{h-D2}(\delta)} \tag{5.22}$$

Where:

$$\boldsymbol{\lambda_{h-D1}(\delta) = \frac{2c1}{1 + \exp\big(c2(\delta + c3)\big)}} \tag{5.23}$$

$$\boldsymbol{\lambda_{h-D2}(\delta) = \frac{2c1}{1 + \exp\big(c2(\delta + c3)\big)}} \tag{5.24}$$

For action selection purpose, TD signal was calculated as the difference in utility function:

$$\boldsymbol{\delta_U(t) = U_t(s_t, a_t) - U_{t-1}(s_t, a_{t-1})} \quad (5.25)$$

The Action Execution Module

The output of D2R and D1R-D2R transmitted to the GPe via the indirect pathway were computed as follows:

$$\boldsymbol{X_t^{IP} = \alpha_{D2}\lambda_{D2}\left(\delta_U(t)y_{D2,t}(s_t, Q_t)\right.} \quad (5.26)$$
$$\boldsymbol{\left. + \alpha_{D1D2}\, sign\big(Q_t(s_t, a_t)\big)\sqrt{h_t}(s_t, a_t)\right)}$$

The transition of D1R neurons output and D1R-D2R via the direct pathway was computed as follows:

$$\boldsymbol{X_t^{DP} = \alpha_{D1}\lambda_{D1}\left(\delta_U(t)y_{D1,t}(s_t, Q_t)\right)} \quad (5.27)$$

Approach#3: Extended actor-critic architecture, TD algorithm and behavioral inhibition module

In this approach, our proposed model also followed the actor-critic architecture. It had five modules: action selection, action execution, input, prediction error signal (DA) and behavioral inhibition signal (5HT). Figure 5.4 describes our proposed model architecture.

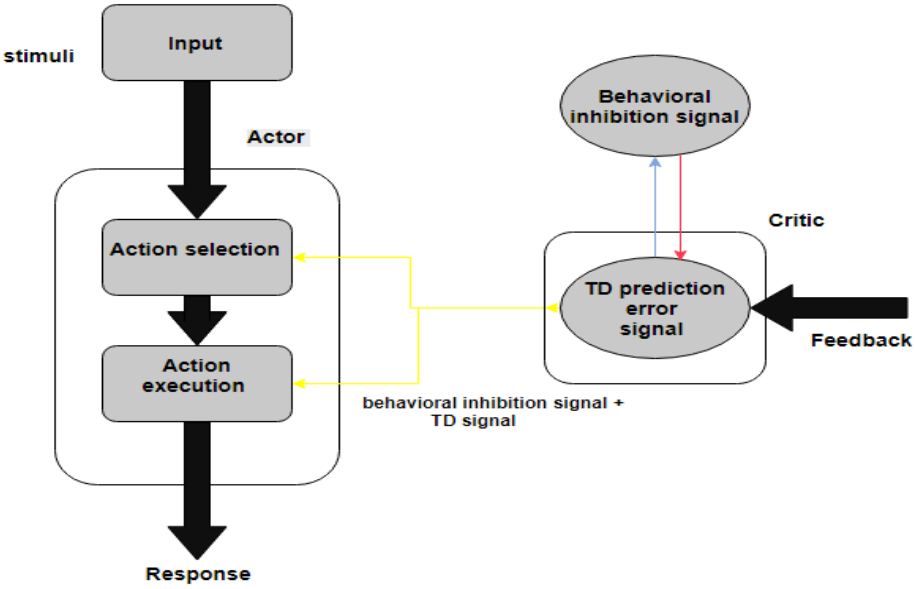

Figure 5.4 The proposed model actor-critic architecture. The model has five modules: input, action selection, action execution, TD prediction error signal and behavioral inhibition module. The critic corresponds to DA neurons whereas the actor represents both action selection and action execution modules and the behavioral inhibition module (5HT). Learning is modulated by phasic response from the critic module. This signal is inhibited by behavioral inhibition signal generated by the behavioral inhibition module. The input module sends projections to the action selection module which is connected to the action execution module. The activation node at the input layer represents one of four stimuli in the reinforcement task. This stimulus is presented in the action selection module where the activation node in the action execution module represents the selected action.

Different computational models represented the 5HT signal as punishment prediction error to oppose the DA reward prediction error signal and together they compute the balance between rewards and punishments to control the inhibition and excitation of system. [34] But in our model, the proposed function of 5HT can be summarized as a reality check system that opposes the formation of associations regardless of feedback type by antagonizing the phasic signals that result in the consolidation of these associations. For the DA system, the 5HT system opposes learning from positive and negative feedback to ascertain the significance of these associations. Figure 5.5 present the flowchart of the behavioral inhibition module role in opposing the TD signal.

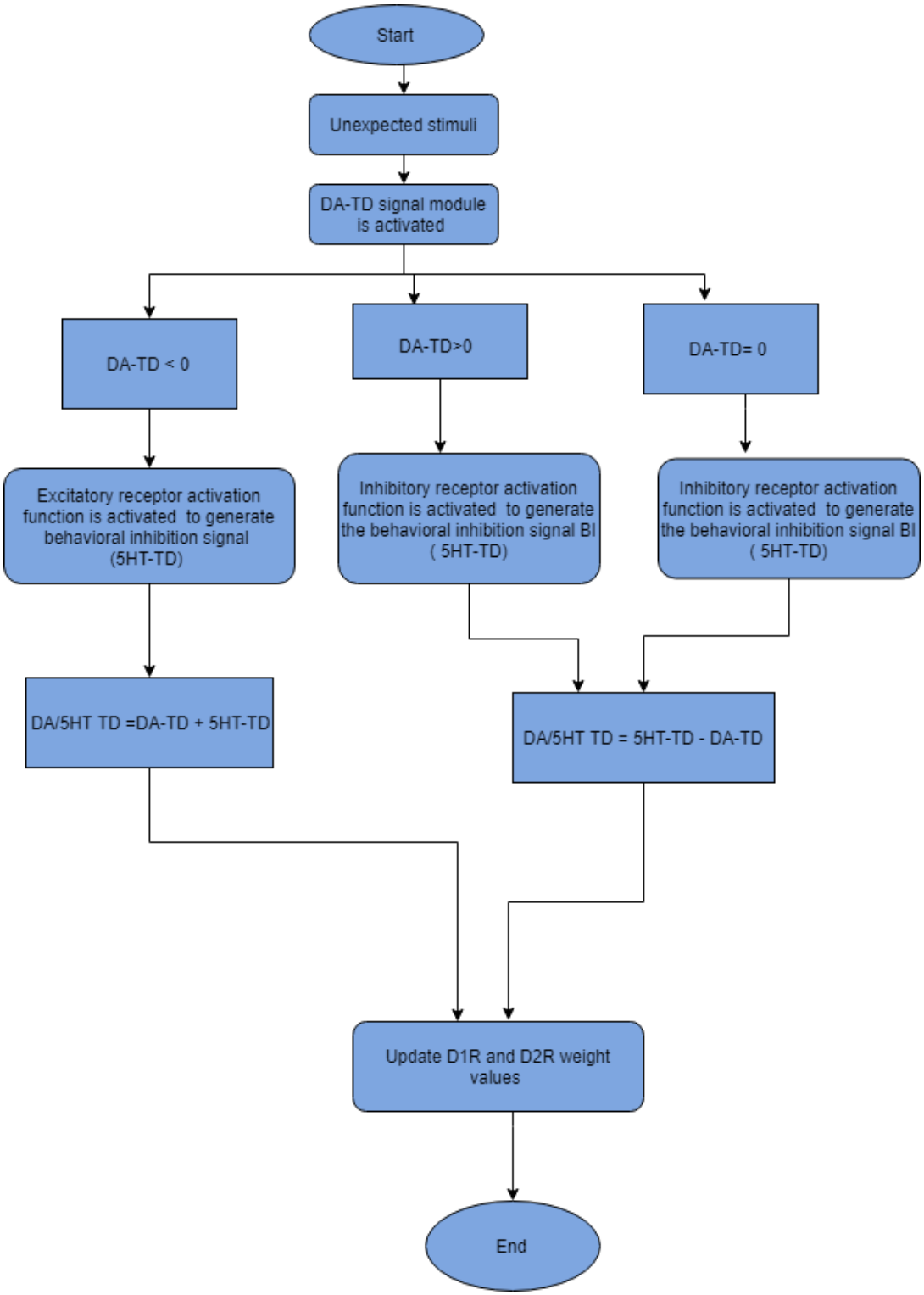

Figure 05.5 A Flowchart of the proposed behavioral inhibition signal module. It describes the role of the behavioral inhibition module in phasically inhibiting TD learning. When the DA-TD module is activated, the generated TD signal is used to activate the behavioral inhibition module. The behavioral inhibition module generates a behavioral inhibition signal (5HT-TD) based on the sign of DA-TD signal. Then the behavioral inhibition signal is added to the DA-TD signal to generate the resultant TD signal which, in turn, engages the D1 and D2 activation functions.

Our proposed module of the behavioral inhibition signal operates following these steps:

1. TD signal module is activated by unexpected stimulus.
2. The generated TD signal is conveyed to behavioral inhibition module
3. The behavioral inhibition module generates a behavioral inhibition signal to oppose the TD signal regardless if it is positive or negative via two types of excitatory and inhibitory receptor that can handle TD signal based on its sign.
4. The resultant TD signal is computed as a sum of both the TD and the behavioral inhibition signals.
5. The resultant TD signal activates the D1 and D2 activation function. Figure 5.6 shows the proposed model of the direct and indirect pathways.

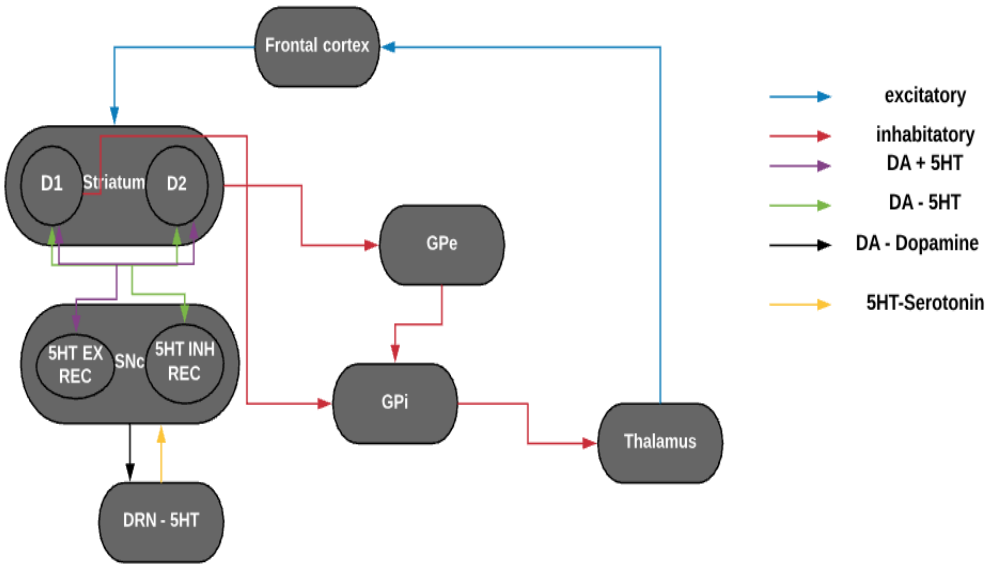

Figure 5.6 Our proposed model in direct and indirect pathways. DRN-5HT neurons are activated by the phasic DA signal, then they project 5HT phasic signal to the excitatory and inhibitory 5HT receptors which are located in the SNpc. The 5HT signal inhibits the phasic DA signal then the resultant signal of DA and 5HT signals activate the DA receptors in striatum whereas D1R activate direct pathway and D2R activate the indirect pathway.

RL Computations

Our proposed model is an extended version from approach #1. In this model we implement a new behavioral inhibition module to oppose the TD signal.

The Behavioral Inhibition Signal Module

We propose that there are two 5HT receptor categories; excitatory and inhibitory receptors. Each type of receptor has its own activation function as follows:

$$\boldsymbol{\alpha_{5HT\_INH\_REC}(\delta) = C1 * (1 + \exp(c2(\delta + c3)))} \quad (5.28)$$

$$\boldsymbol{\alpha_{5HT\_EX\_REC}(\delta) = C1 * (1 + \exp(c2(\delta - c3)))} \quad (5.29)$$

Where

$\boldsymbol{\alpha_{5HT\_INH\_REC}}$ And $\boldsymbol{\alpha_{5HT\_EX\_REC}}$ represent the value of phasic serotonin signal.

The resultant TD signal was computed as follows:

For positive reward signal:

$$\boldsymbol{\delta_{Resultant} = \delta - \alpha_{5HT\_INH\_REC}} \quad (5.30)$$

But for negative reward signal:

$$\boldsymbol{\delta_{Resultant} = \delta + \alpha_{5HT\_EX\_REC}(\delta)} \quad (5.31)$$

Then $\boldsymbol{\delta_{Resultant}}$ is used to activate D1R and D2R neurons. We used the same modules for action selection and action executions which were used in approach #1 model. Figure 5.7 shows the interacted modules in the proposed model.

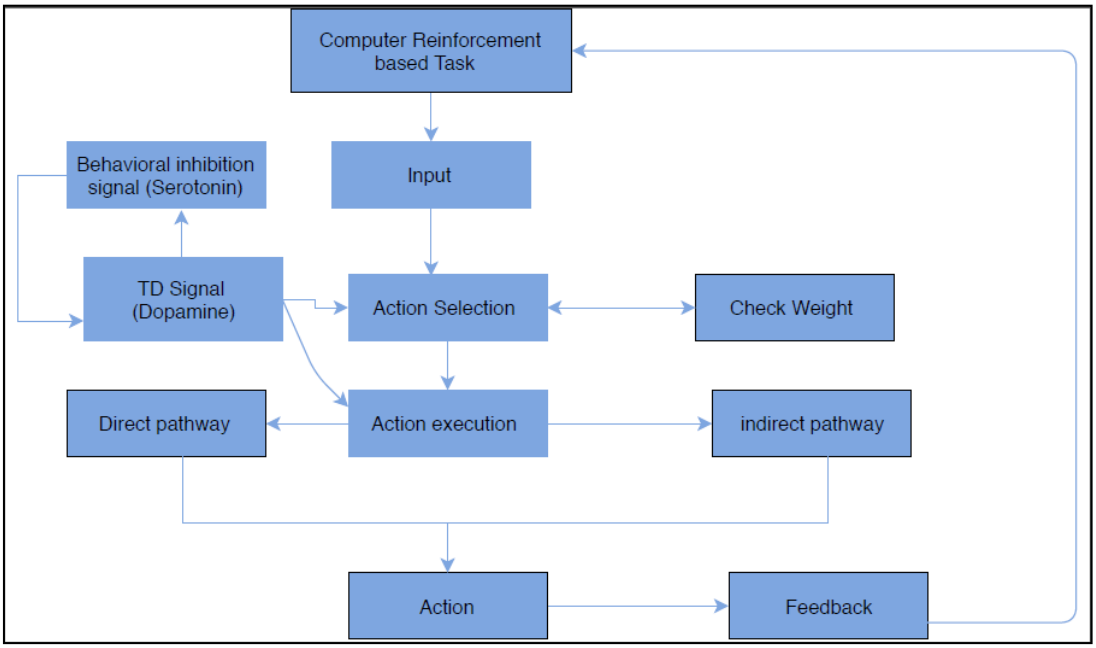

Figure 5.7 The modules in our proposed model. This figure describes the interaction between our behavioral inhibition module and the rest of the modules in the system. The computer-based reinforcement learning task sends the input to the input module. The input module sends the stimulus to the action selection module. The action selection module checks the weight of the selected action based on the DA-TD value. However, DA-TD signal is inhibited by the signal that is generated by the behavioral inhibition module. Based on the resultant TD signal (DA-TD and 5HT behavioral inhibition) the action execution module selects from two actions, direct and indirect pathway. Its selection depends on the weight estimation and the value function.

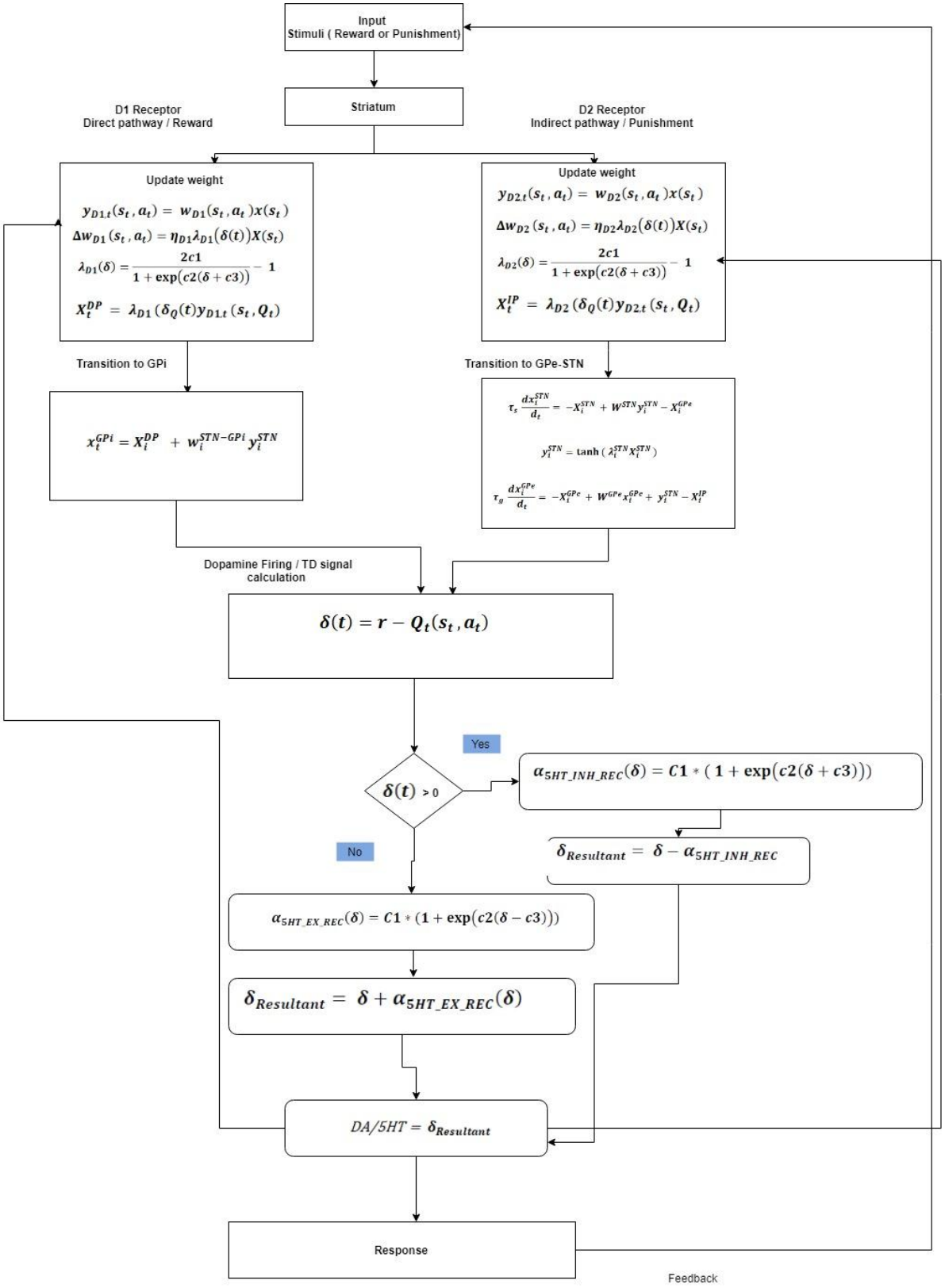

Figure 05.8: The mathematical representation of the anatomical modules in our model. This figure describes the interaction between our behavioral inhibition module and the rest of the anatomical modules in the

system. The computer-based reinforcement learning task sends the input to the striatum. The striatum checks the weight of the selected action based on the dopamine value, to select direct and indirect pathways. If D1R is activated, the direct pathways is activated Via GPi, on the other hand, if D2R receptor is activated then indirect pathway is activated Via GPe-STN. However, TD signal is inhibited by the signal that is generated by the 5HT module.

5.4 MDD Modeling

In this section, for testing the application of our models, we simulated the effects of DA and 5HT signals within BG models in patients with MDD. Also, we modeled the effects of SSRI antidepressants. The simulation results will be presented in Chapter 6.

MDD Modeling Rules:

1. Patients with MDD have a deficit in producing DA. Computationally using all modeling approaches, we can simulate this deficit by limiting the value of the TD signal or decreasing the activation function for D1 or D2 receptors.
2. Patients with MDD have a deficit in producing 5HT. computationally; we can simulate this deficit by manipulating the value of risk prediction in approach#2 and by decreasing the value of behavioral inhibition signal in approach#3.
3. SSRI antidepressants increase the concentration of both DA and 5HT.
4. SSRI antidepressants decrease the activity for D2 receptors [29].
5. All parameters are manipulated using trial by trial computational analysis. To find the best fitting between experimental results and simulation results.

MDD Modeling using Approach#1

In Approach #1, we simulated the effects of DA in patients with MDD on learning from positive and negative feedback in two ways:

1. Limiting the DA concentration by limiting the value of the TD signal which causes a deficit in learning from positive feedback.

$$\boldsymbol{\delta} > \boldsymbol{\delta}_{\mathbf{lim}};\ \boldsymbol{\delta} = \boldsymbol{\delta}_{\mathbf{lim}} \qquad 5.32$$

2. Decreasing the weight of the D1 receptor.

In Approach #1, we simulated the deficit in learning from positive feedback by decreasing the activation function of the D1 receptor. We changed the constant value C1 from 1.0 to 0.3 to decreasing the response of the D1 receptor.

To simulate the effects of SSRI antidepressants in patients with MDD, Herzallah et al., suggested that the medicated patients with SSRI have deficit in learning from both positive and negative feedback. We simulated the effects of SSRI in two ways:

1. Limiting the DA concentration through limiting the value of the TD signal which causes a deficit in learning from positive feedback. Also, we decreased the weight of D2 receptors to create a deficit in learning from negative feedback. We limited the TD value to 0.015 and we decreased the constant C1 in the activation function of D2 receptor from 1.0 to 0.25.
2. Decreasing the weight for both D1 and D2 receptors. We decreased the constant value C1 in the activation function of D1 from 1.0 to 0.3 also we decreased the constant C1 in the activation function of D2 from 1.0 to 0.3.

MDD Modeling using Approach #2

In this model, we studied the effects of both DA and 5HT in MDD patients, where DA is represented as TD signal and 5HT is represented by the risk prediction (α). We simulated MDD as follows as a decrease in the risk prediction value (5HT) from 0.006 at HC to 0.003 while also limiting the TD signal to 0.008.

To simulate the effects of SSRI antidepressants in patients with MDD, was increased the risk prediction from 0.003 in patients with MDD to 0.004. We also increased the TD signal from 0.008 in patients with MDD to 0.01. Finally, we decreased the D2 activation function by decreasing the D2 learning rate.

MDD Modeling using Approach#3

In our proposed model, we simulated the effects of DA and 5HT in patients with MDD using the TD signal and the behavioral inhibition signal. The concentration of DA was limited by limiting the TD signal, thus, leading a decrease in 5HT concentration. The 5HT activation function is a function of TD signal, hence decreasing of TD signal will lead to decreasing the 5HT concentration which ultimately leads to a deficit in learning from reward feedback.

To simulate the effects of SSRI antidepressants, we increased the activation function of 5HT receptors to inhibit learning from both reward and punishment feedback.

5.5 Tool Used

The modeling was conducted on a Dell laptop with Windows 10 pro, and processor 2.4 GHz Intel core i5. We used MATLAB R2015a environment for modeling implementation. For documentation and drawing we used the Microsoft Office package.

5.6 Human Experimental Data Set

We used data obtained from the aforementioned cognitive task by recording the response of patients with MDD on 160 trials. These data were collected at the Palestinian Neuroscience Initiative at Al Quds University, Abu Dis. It was published in 2013 [28]

5.7 System Requirements

1. Fitting experimental results for patients with MDD which were published in Herzallah et al., 2013. It implies that healthy subject learned from both reward and punishment feedback, but untreated patients with MDD learned from punishment feedback and but not from reward feedback. After receiving SSRI antidepressants, medicated MDD patients did not learn from reward or negative feedback. Experimental results are shown in figure 5.8 (A, B).
2. D1 receptor function should control reward learning while D2 receptor function control punishment learning.
3. The 5HT receptor functions should inhibit the DA signal (TD signal)
4. All learning curve peaks should exceed 50% accuracy (higher than chance (50%)).

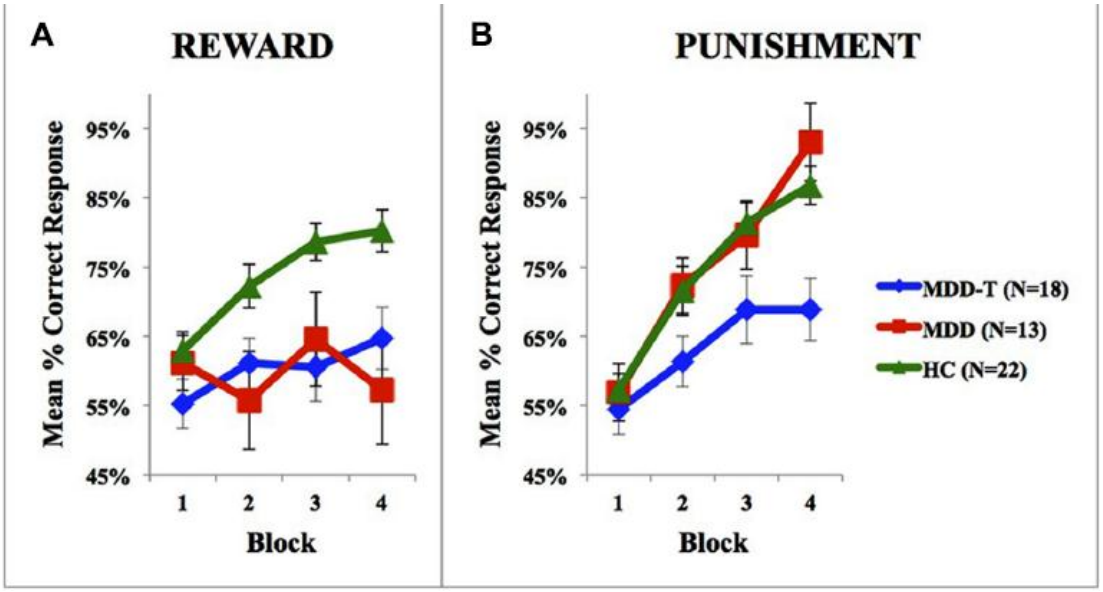

Figure 05.9 The Experimental Results. (A) The mean number of correct responses in the four phases or the reward stimuli (±SEM). (B) The mean number of correct responses in the four phases or the punishment stimuli (±SEM). MDD is medication naïve, MDD-T is on medication MDD patients, and HC is healthy controls. Adapted from Herzallah et al., 2013.

5.8 Testing Phase

First, we tested the Balasubramani at el 2015 model. We used the online dataset from the published paper for patients with Parkinson's disease. Our aim was to check the functionality and validity of the published model code.

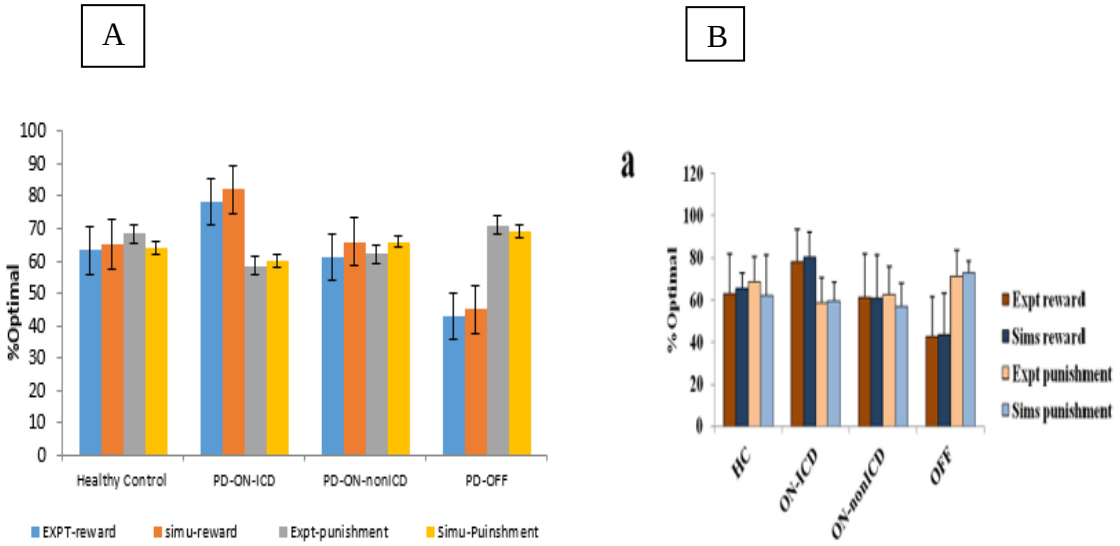

Figure 5.10 (A) Replication of the simulation results for data from patients with Parkinson's disease using Approch#2. (B) Simulation results reported in Balasubramani at el 2015 using Approach#2.

From figure 5.10 (A, B), we can see that our simulation results for patients with Parkinson are replicated the simulation results reported by Balasubramani at el 2015 model. Accordingly, our model in Approach#2 passed the verification and validation process by reproducing the simulation results reported in Balasubramani at el 2015. Therefore, we used the Approach#2 model to simulate the dataset from patients with MDD. [41]

We calculated the normalized error factor between experimental data and simulation results to evaluate the performance of our models. Chapter 6 presents normalized error curves for different modeling approaches.

Chapter Six

Results

Section 1: Simulation of DA-Only TD Signal in RL in MDD

Section 2: Simulation of DA/5HT TD and Risk Prediction in RL in MDD

Section 3: Simulation of DA/5HT TD and Behavioral Inhibition in RL in MDD

Section 4: Model Accuracy

.

Chapter Six:

Results:

In this chapter, we present the simulation results of the different modeling approaches. First, we present our simulation results of the DA signal to find the effects of D1, D2 receptor activation functions and TD signal in learning from positive and negative feedback on patients with MDD. We also describe the effects of SSRI antidepressants on the TD signal. Then, we present the results of simulating the TD signal and the risk prediction to study the effects of both DA and 5HT interaction on patients with MDD with and without SSRI antidepressants. Finally, we present the simulation results of the effects of the TD signal and the behavioral inhibition signal on learning from positive and negative feedback in patients with MDD before and after SSRI antidepressants based on our extended actor-critic architecture.

Simulation results of the models which are presented here have an average of 30 runs, where each run includes 100 instances and the results represent the average ($\pm$SEM) of these 100 instances.

6.1 Simulation of DA-Only TD Signal in RL in MDD (Approach#1)

In this approach, healthy people learned from both reward and punishment feedback, but unmedicated patients with MDD learned from punishment trials feedback but not from reward feedback. On other hand medicated MDD patients did not learn from reward or punishment feedback. Figure 6.1 & 6.2 show simulation results.

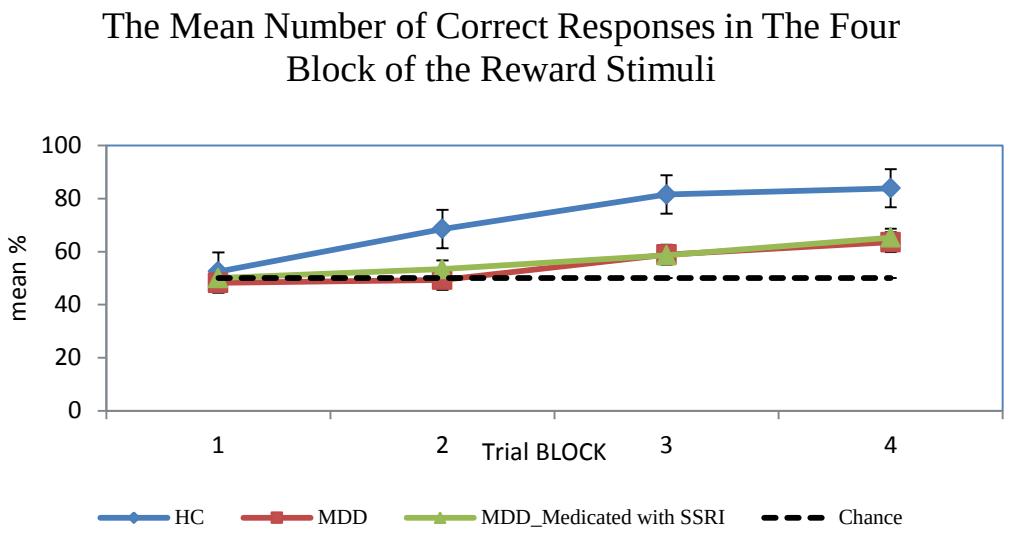

Figure 6.1 Simulation results for learning from reward using Approach#1. We modeled MDD by decreasing the value DA-TD signal. Also, we modeled the effects of SSRI antidepressants by increasing the DA-TD signal and decreasing the weight of D2 receptors. The lines represent the average learning accuracy (±SEM) from reward feedback per group. The black dashed line represents chance level (50%). Results show healthy subjects learned from reward feedback while unmedicated patients with MDD and medicated patients with MDD did not.

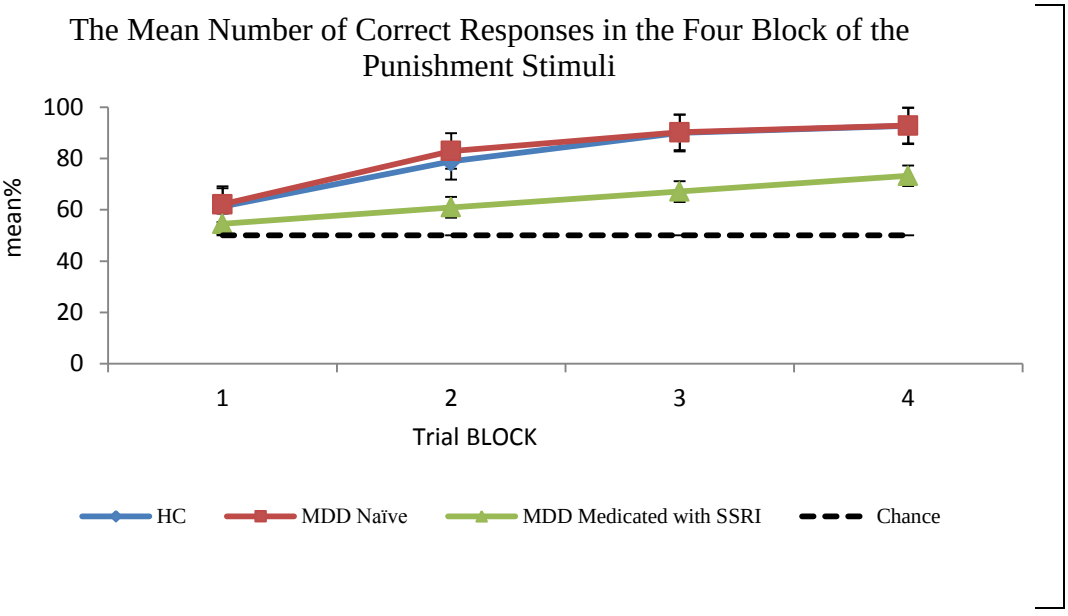

Figure 6.2 Simulation results for learning from punishment using Approach#1. We modeled MDD by decreasing the value DA-TD signal. Also, we modeled the effects of SSRI antidepressants by increasing the DA-TD signal and decreasing the weight of D2 receptors. The lines represent the average learning accuracy (±SEM) from punishment feedback per group. The black dashed line represents chance level (50%). Results show that healthy subjects and unmedicated patients with MDD learned from punishments feedback while medicated patients with MDD did not.

We simulated unmedicated patients with MDD in two ways:

1. Limiting the amount of dopamine concentration by limiting the value of TD signal. (Figure 6.3)
2. Decreasing the weight of D1 receptor activation function. (Figure 6.4)

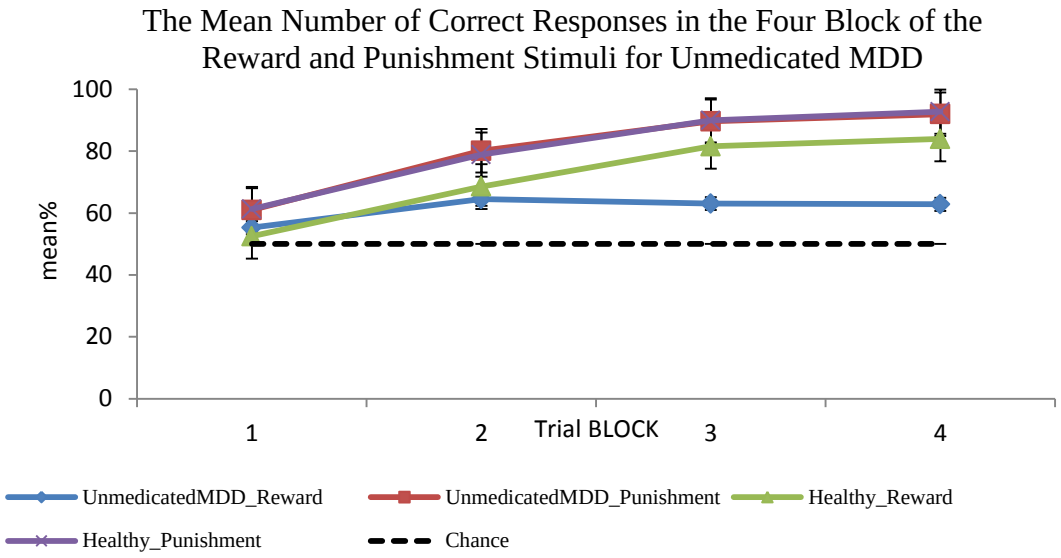

Figure 6.3 Simulation results for modeling MDD using Approach#1. We modeled MDD by limiting the TD signal. The TD signal represents DA. Hence, low DA concentration impaired learning from reward feedback and doesn't affect learning from punishment feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

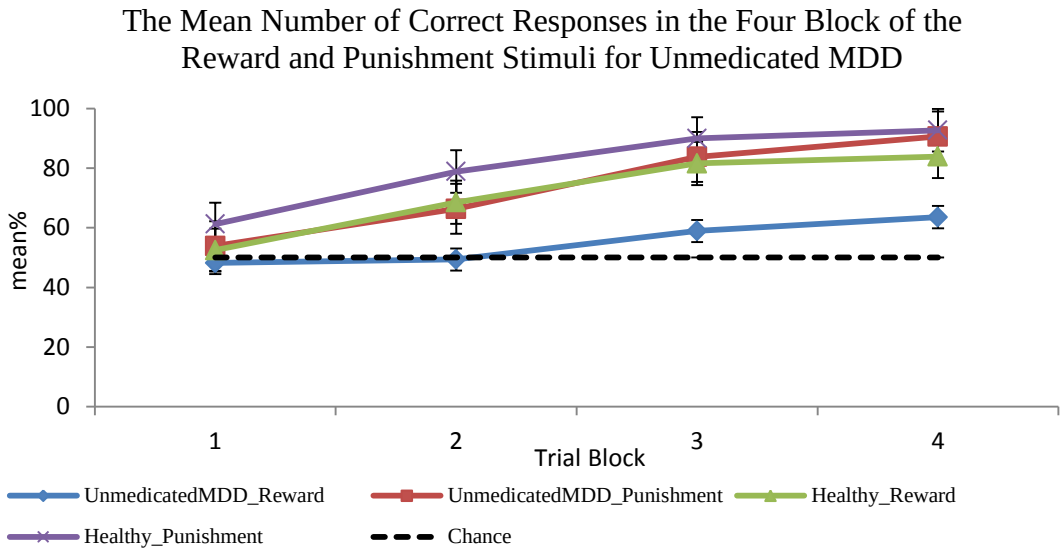

Figure 6.4 Simulation results for modeling MDD using Approach#1. We modeled MDD by decreasing the learning rate of D1R. This results in an impairment in learning from reward feedback and no change in learning from punishment feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%)

We simulated the effect of SSRI antidepressants in medicated patients with MDD using two approaches:

1. Limiting the DA concentration by limiting the value of the TD signal, and decreasing the weight of D2 receptor activation function (Figure 6.5).
2. Decreasing the weight of both D1 and D2 receptors activation functions. (Figure 6.6).

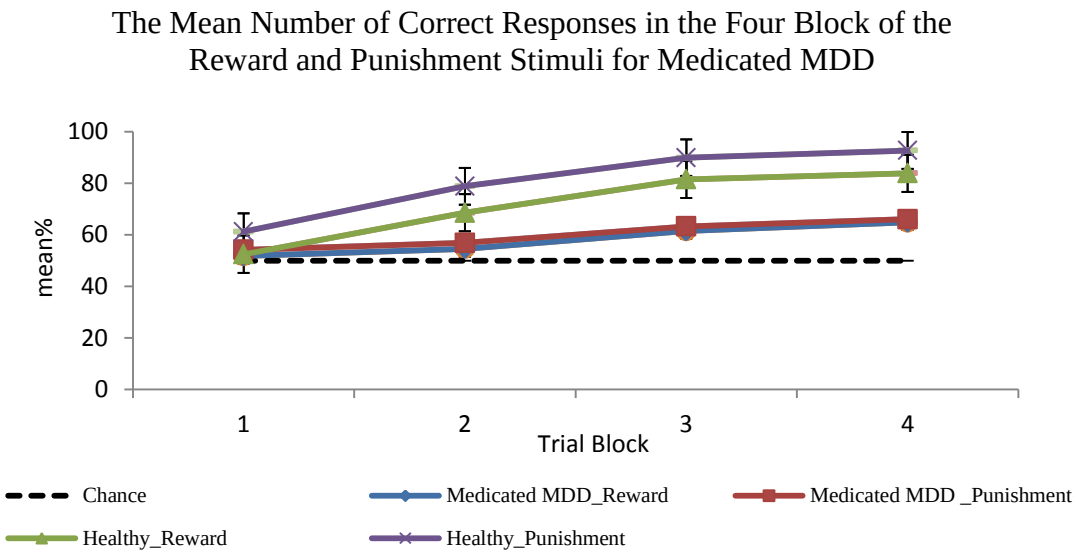

Figure 6.5 Simulation results for medicated MDD using Approach#1. We modeled MDD by limiting the value of TD signal. We modeled the effects of SSRIs by decreasing the weight of D2R. The figure shows an impairment in learning from both reward and punishment feedback, where limiting the TD value (decreasing DA concentration) impairs learning from reward feedback and decreasing the weight of D2 receptor impairs learning from punishment feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

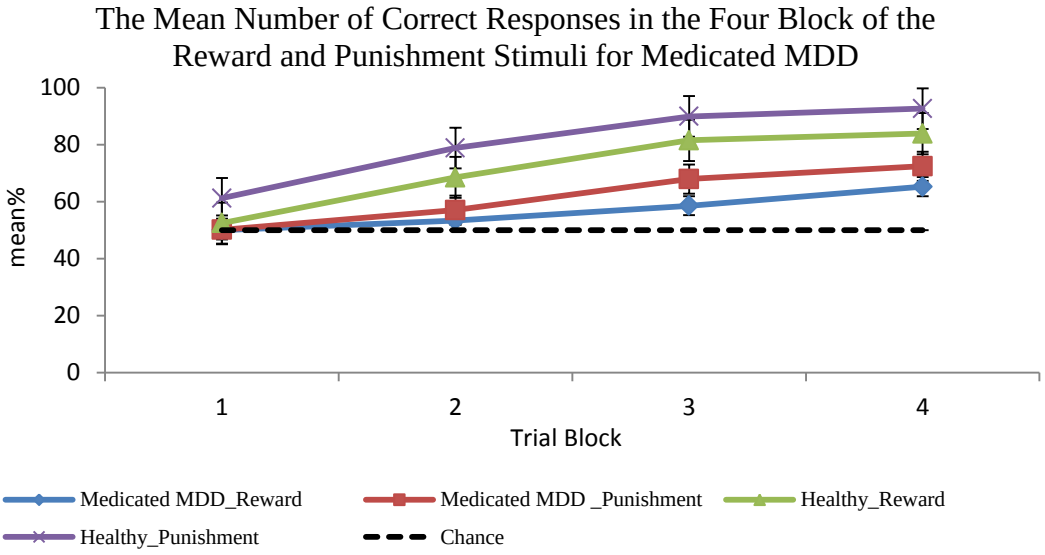

Figure 6.6 Simulation results for modeling medicated MDD using Approach#1. We modeled MDD by decreasing the weight for D1R. We modeled the effects of SSRIs by decreasing the weight for D2R. The figure shows an impairment in learning from reward and punishment, where decreasing the weight of D1R impairs learning from reward feedback and decreasing the weight of D2R impairs learning

from punishment feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

6.2 Simulation of DA/5HT TD and Risk Prediction in RL in MDD (Approach#2)

In this section, we present different simulation for healthy people, unmedicated patients with MDD and medicated patients with MDD. Simulation result showed that healthy controls learned from reward and punishment feedback. Figure (6.7)

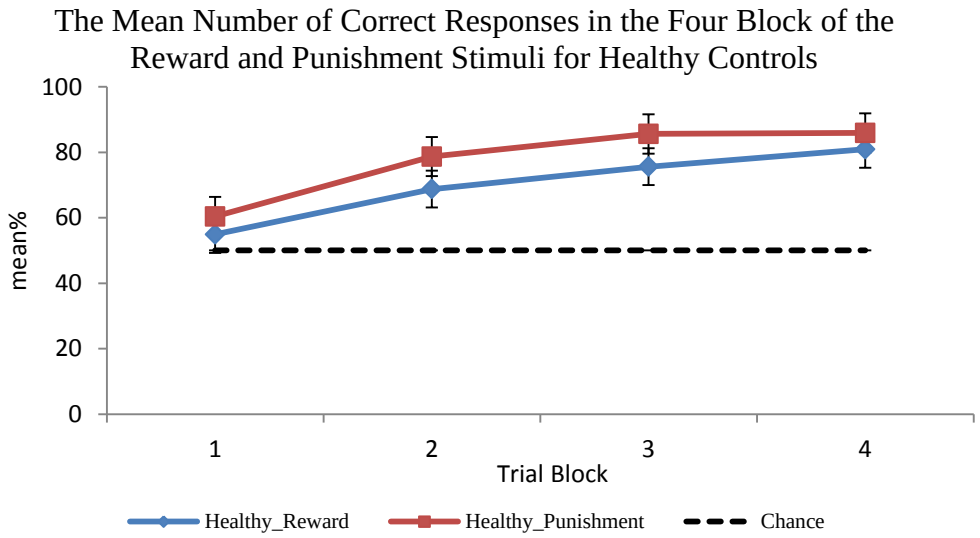

Figure 6.7 Simulation results for healthy control using Approach#2; learned from punishment and reward feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback. The black dashed line represents chance level (50%).

Here, we present different simulations to study the effects of TD signal and risk prediction error in patients with MDD.

1. Decreasing the risk prediction value. (Figure 6.8).
2. Limiting the value of the TD signal. (Figure 6.9).

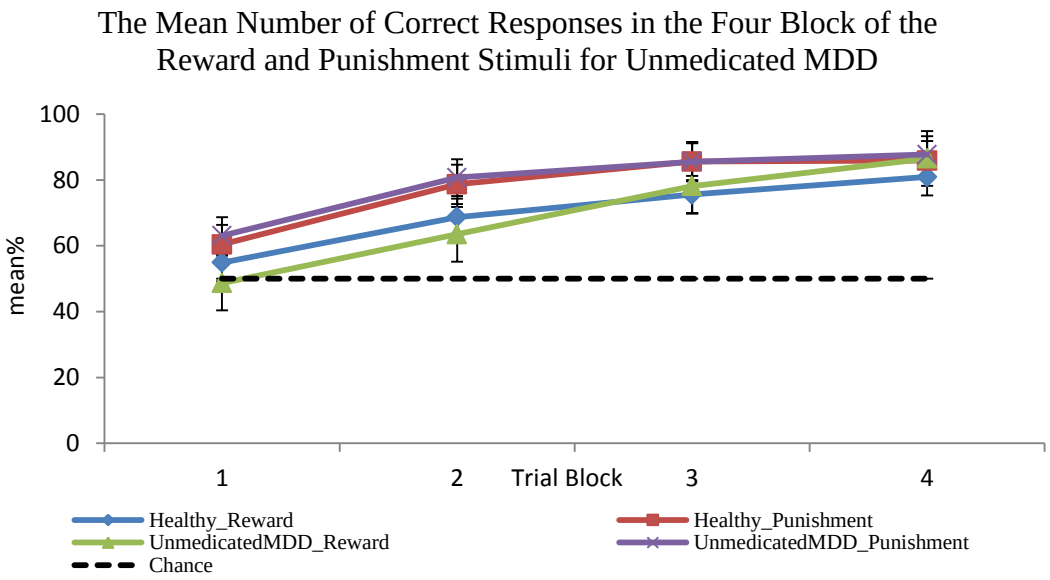

Figure 6.8 Simulation results for modeling MDD using Approach#2. We modeled MDD by decreasing the risk prediction value (5HT). Results show that decreasing the risk prediction value (5HT) didn't affect learning from punishment or reward feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

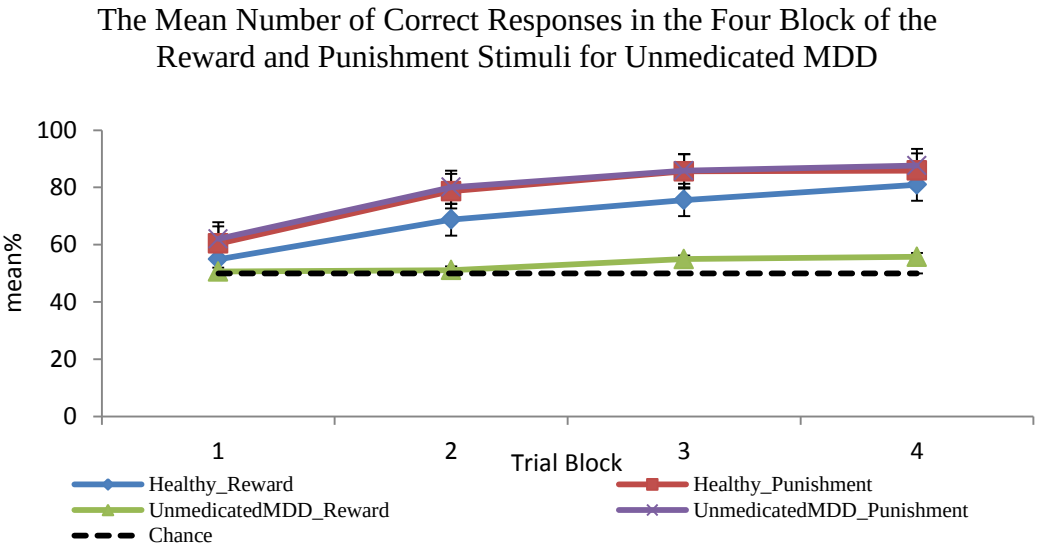

Figure 6.9 Simulation results for modeling MDD using Approach#2. We modeled MDD by limiting the TD signal only (DA) without altering the risk prediction value (5HT). Results show an impairment in learning from reward feedback similar to the experimental results. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

Also, we used multiple simulation approaches to study the effects of SSRI antidepressants in medicated patients with MDD:

1. Increasing the risk prediction value (5HT concentration) alone. (Figure 6.10).
2. Increasing the TD signal (DA concentration) alone. (Figure 6.11)
3. Decreasing the D2 learning rate alone. (Figure 6.12)
4. Increasing both the TD signal and the risk prediction value. (Figure 6.13)
5. Increasing the risk prediction value and decreasing the D2 learning rate. (Figure 6.14).
6. Increasing the TD signal and decreasing the D2 learning rate. (Figure 6.15).

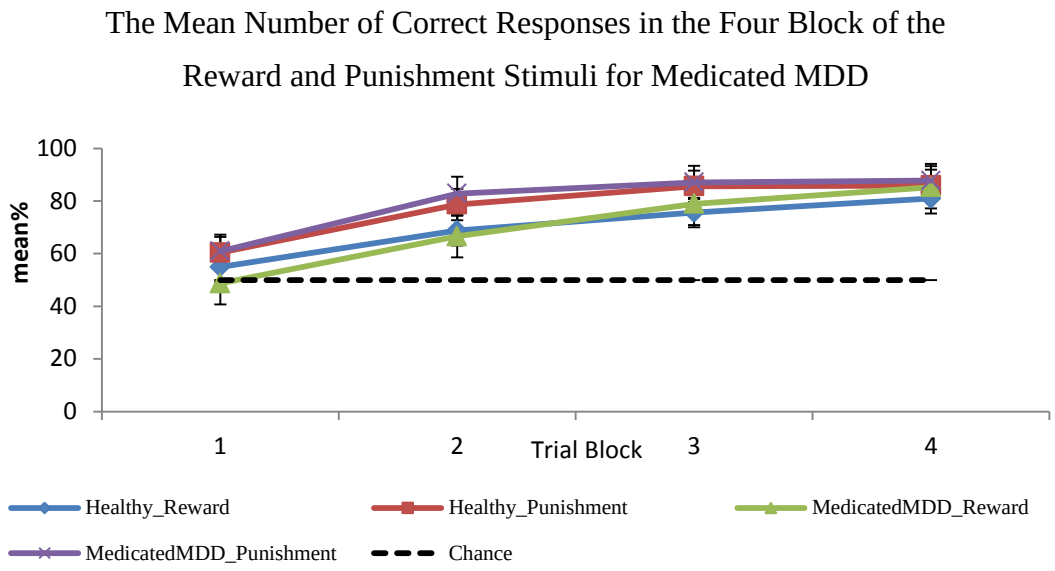

Figure 6.10 Simulation results for modeling medicated MDD using Approach#2. We hypothesized that MDD could be represented by a low value for risk prediction and modeled the effects of SSRIs by increasing the risk prediction value (5HT). Increasing the value of risk prediction (5HT) doesn't significantly affect learning from reward and punishment feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

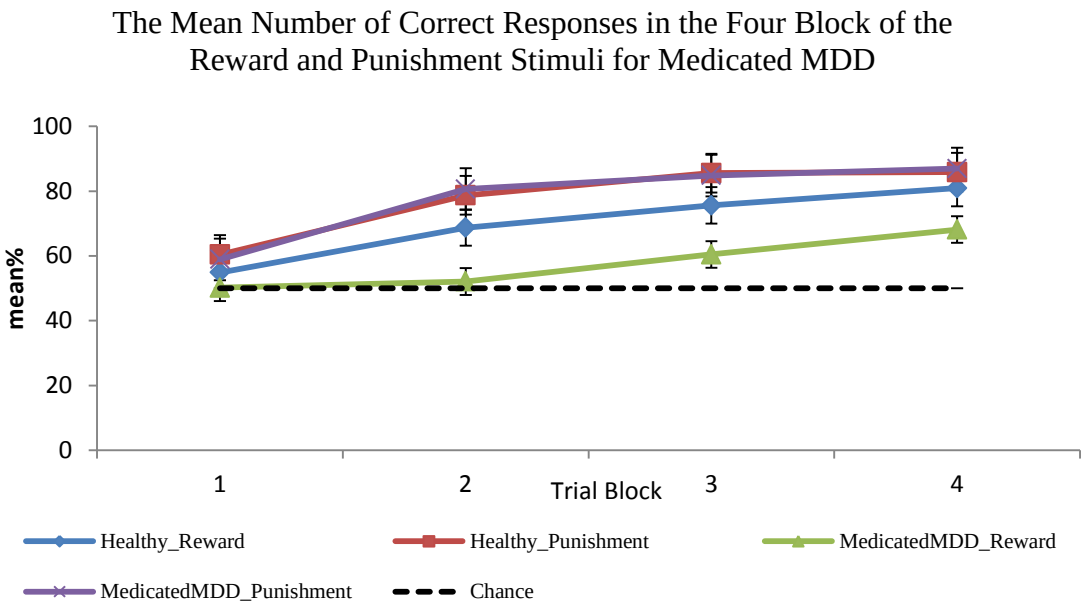

Figure 6.11 Simulation results for modeling medicated MDD using Approach#2. We modeled MDD by limiting the value of TD signal (DA). We modeled the effects of SSRIs by slightly increasing the TD signal (DA). The figure shows a slightly enhanced learning from reward feedback and does not affect learning from punishment feedback. This is different from the experimental findings in Herzallah et al. 2013. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

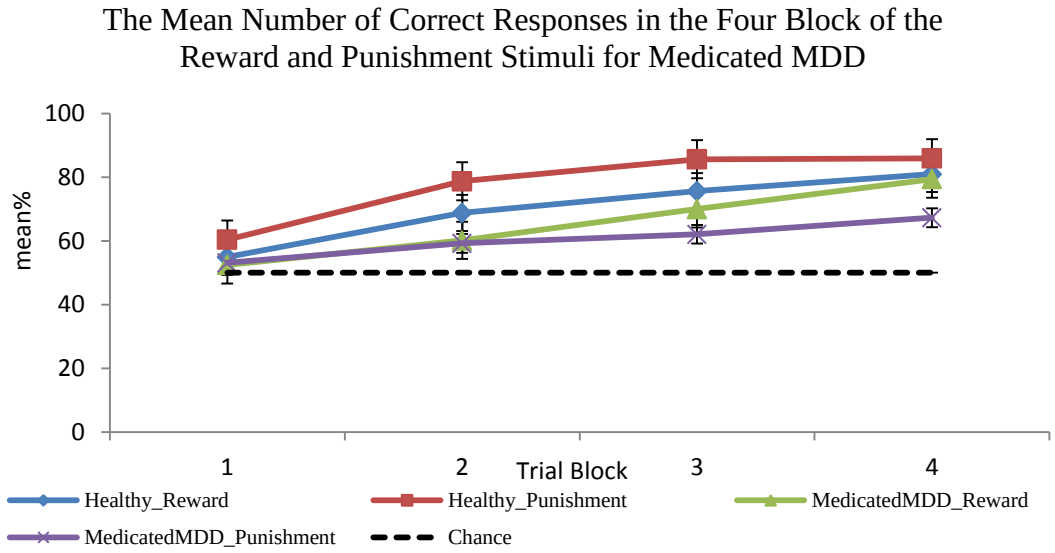

Figure 6.12 Simulation results for modeling medicated MDD using Approach#2. We hypothesized that MDD could be represented by a low value for risk prediction and modeled the effects of SSRIs by decreasing the D2R learning rate. This causes impairment in learning from the punishment feedback

and it doesn't significantly affect learning from reward feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

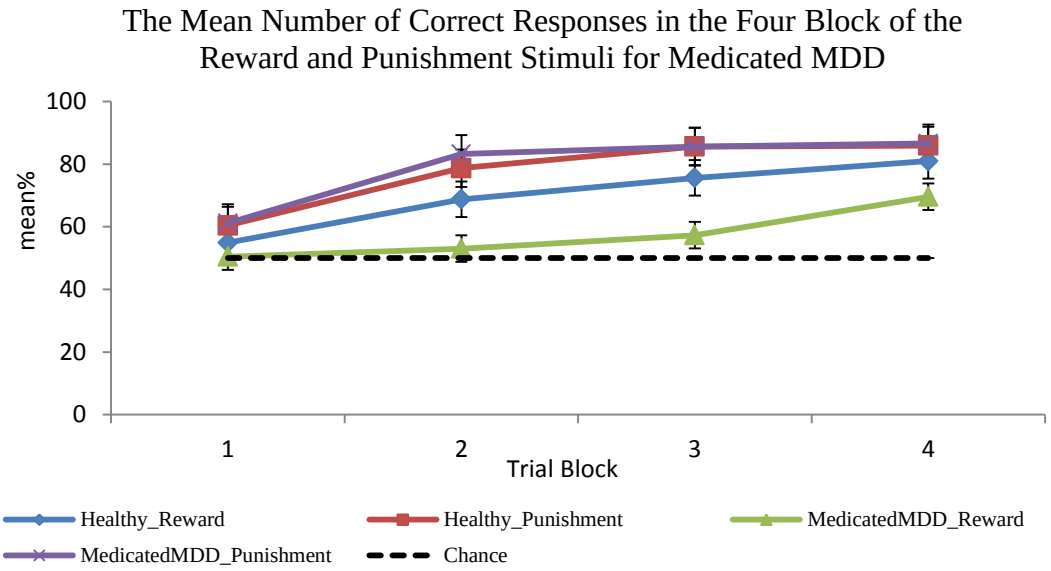

Figure 6.13 Simulation results for modeling medicated MDD using Approach#2. We hypothesized that MDD could be represented by a low value for risk prediction and modeled the effects of SSRIs by increasing both the TD signal (DA) and risk prediction (5HT). This slightly enhanced learning from reward feedback but had no effect on learning from punishment feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

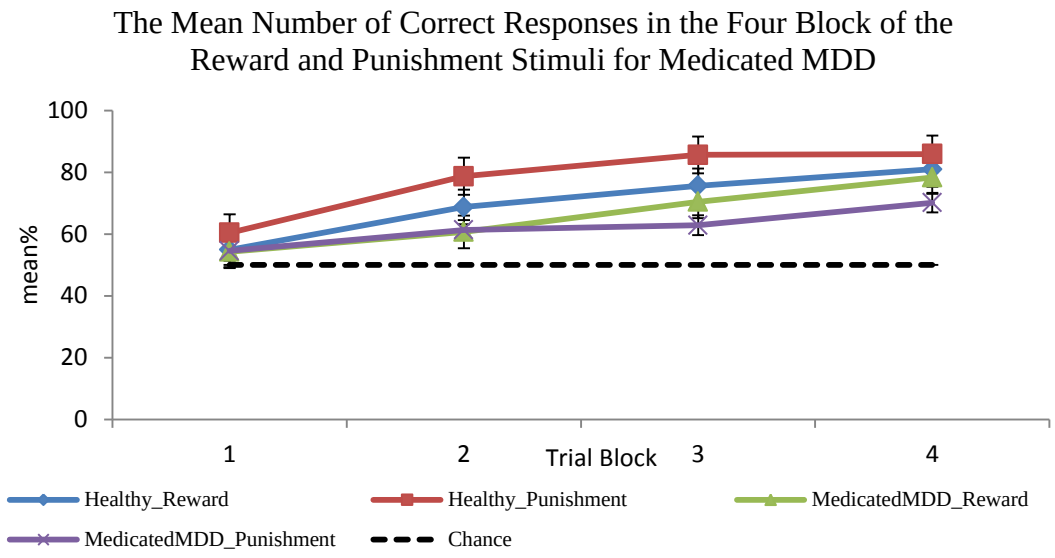

Figure 6.14 Simulation results for modeling medicated MDD using Approach#2. We hypothesized that MDD could be represented by a low value for risk prediction and modeled the effects of SSRIs by increasing risk factor (5HT) and decreasing learning rate of D2. It shows impaired learning from punishment feedback and a slight impairment in learning from reward feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

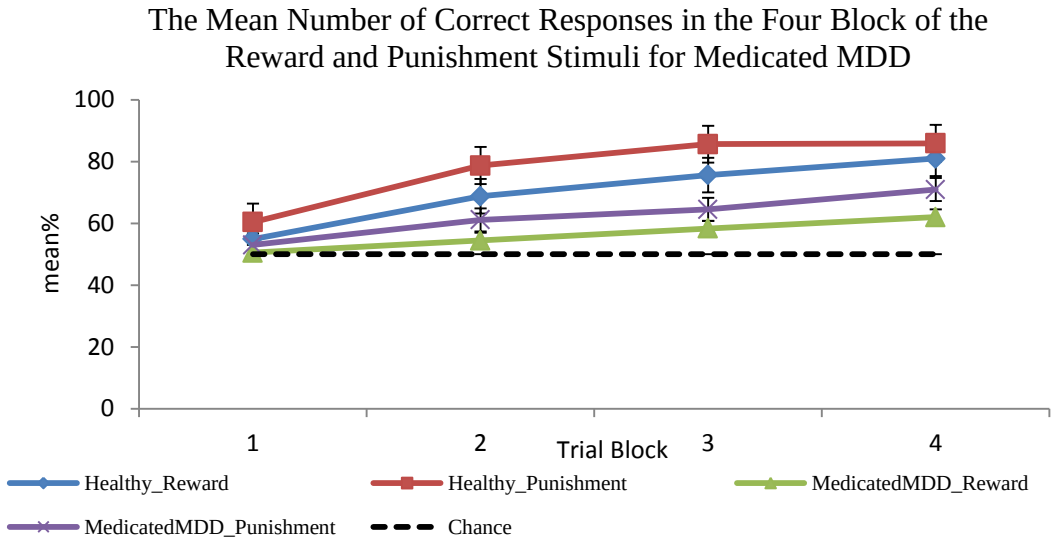

Figure 6.15 Simulation results for modeling medicated MDD using Approach#2. We modeled MDD by limiting the TD signal (DA) and we modeled the effects of SSRIs by decreasing learning rate of D2. It shows impairment in learning from both reward and punishment feedback in line with the experimental results. The lines represent the average learning accuracy (+SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

6.3 Simulation of DA/5HT TD and Behavioral Inhibition in RL in MDD (Approach#3)

In our proposed model, healthy people are learned from both reward and punishment trials feedback. Unmedicated patients with MDD learned from punishment feedback but not from reward feedback. Medicated MDD patients did not learn from weather reward and punishment feedback.

To simulate the effects of DA and 5HT on patients with MDD, the concentration of DA was limited by limiting the TD signal which represents DA signal only, thus, leading to a decrease in the behavioral inhibition signal (5HT) given its TD-dependent activity. To simulate the effects of SSRI antidepressants, we increased the activation function of 5HT receptors so that it mimics an increase in the 5HT concentration.

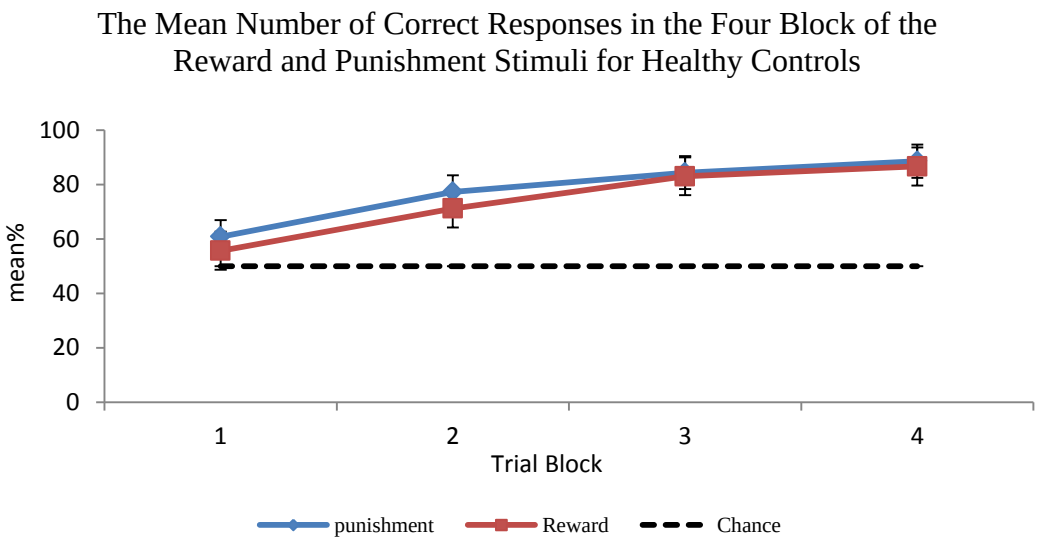

Figure 6.16 Simulation results for healthy controls using Approach#3; learned from reward and punishment feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%)

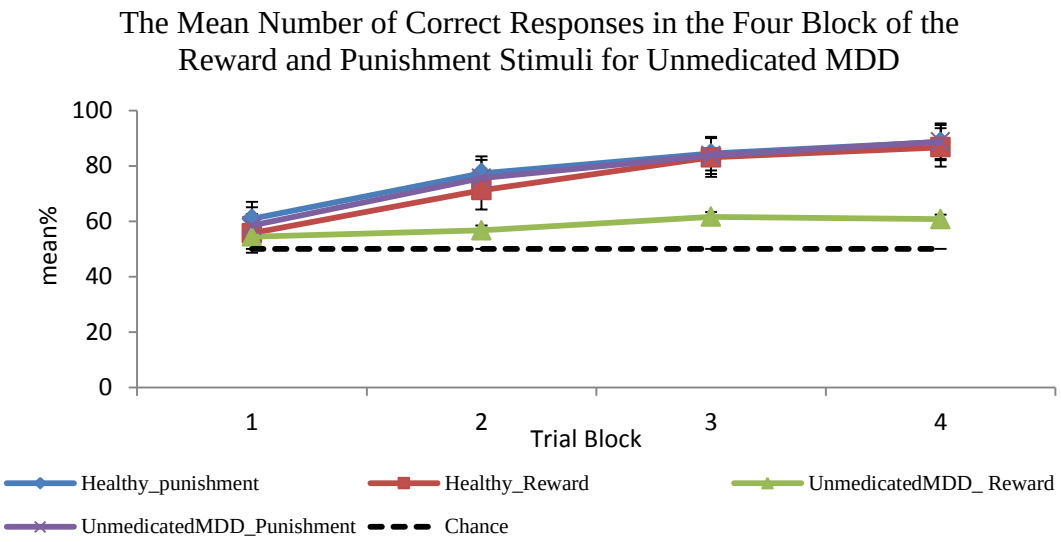

Figure 6.17 Simulation result for modeling MDD using Approach#3. We modeled MDD by limiting the DA-TD signal, thus decreasing both DA-TD and 5HT behavioral inhibition. It shows impaired learning from reward feedback but not from punishment feedback, in line with experimental results. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

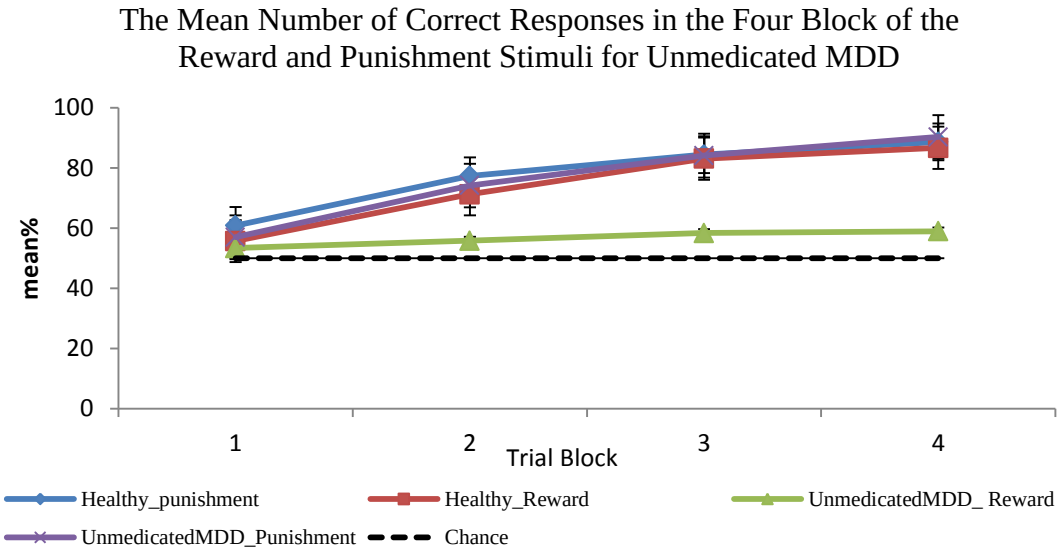

Figure 6.18 Simulation result for modeling MDD using Approach#3. We modeled MDD by limiting the DA-TD signal, and decreasing the activation function of 5HT receptors thus decreasing both DA-TD and 5HT behavioral inhibition. It shows impaired learning from reward feedback but not from punishment feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%).

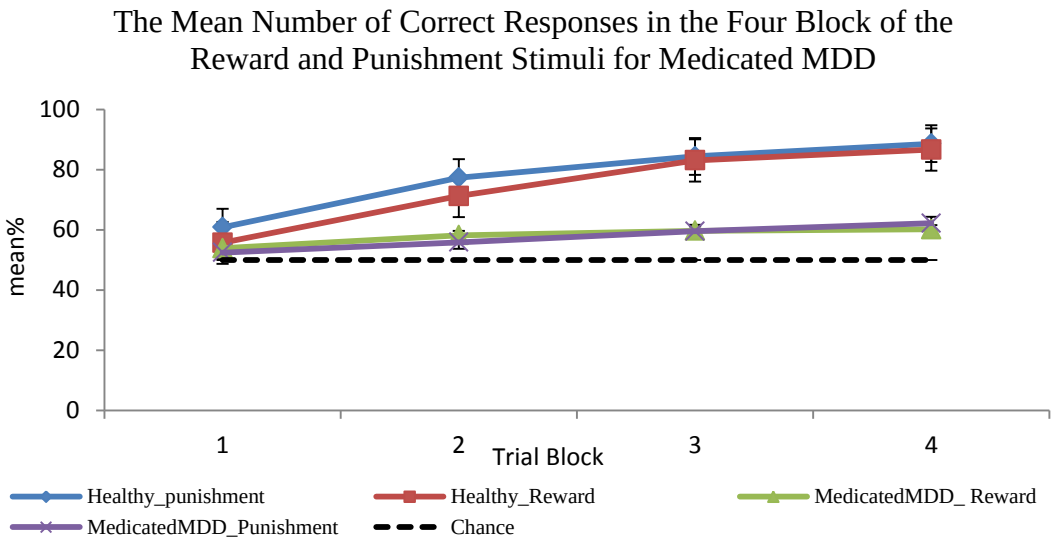

Figure 6.19 Simulation results for medicated MDD using Approach#3. We modeled the effects of SSRIs by increasing the behavioral inhibition signal (5HT) which impaired learning from both reward and punishment feedback. The lines represent the average learning accuracy (±SEM) from reward feedback and punishment feedback per group. The black dashed line represents chance level (50%)

6.4 Model Accuracy

For testing the accuracy of different models approaches, we used the normalized error algorithm to find the difference between our simulation results and the experimental results.

$$Normalized\ Error = ((expt_{rew}\text{-}sims_{rew})/expt_{rew})^2 + ((expt_{pun}\text{-}sims_{pun})/expt_{pun})^2 \quad (5.22)$$

The following figures (6.18, 6.19, 6.20) for healthy controls, unmediated patients with MDD, medicated patients with MDD represent the normalized error per modeling approach.

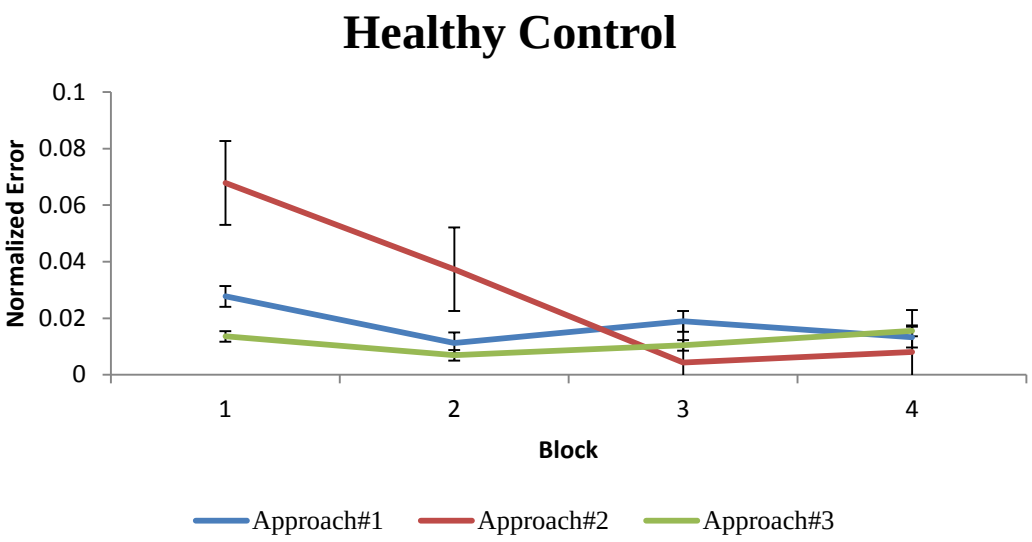

Figure 6.20 Normalized modeling error for healthy control. Approach#1 and Approach#2 have higher normalized error in first and second blocks compared with Approch#3 but there is no significant difference between approaches in the third and fourth blocks. Overall, Approach#3 has the best fitting curve to the experimental data.

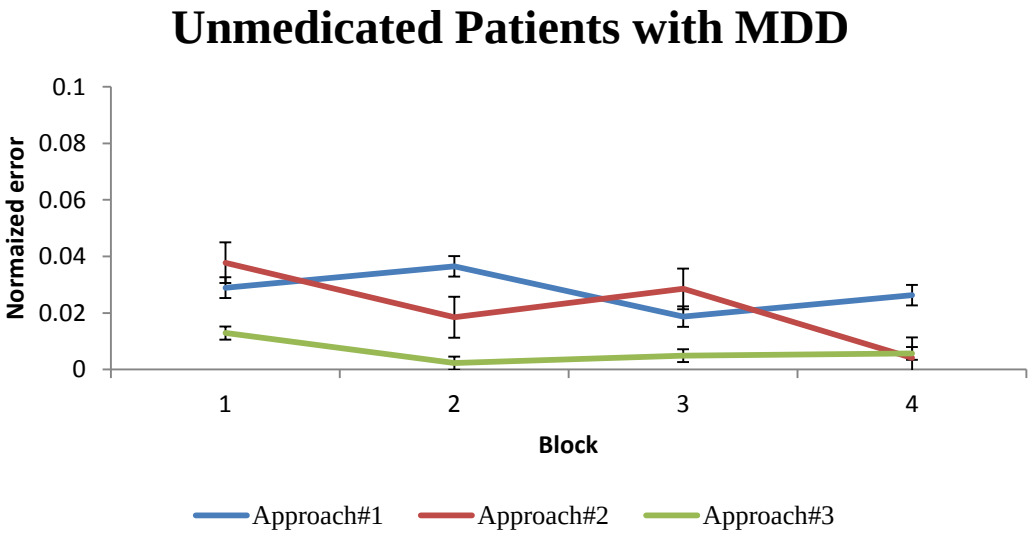

Figure 6.21 Normalized modeling error for unmedicated patients with MDD. It shows that our Approach#3 had the best fitting results compared to Approach#1 and Approach#2, Regarding Approach#1, simulation results that were used to estimate the normalized prediction error are shown in figures (6.1 & 6.2) while the simulation results that were used for Apprach#2 were shown in figures (6.7, 6.9, 6.14).

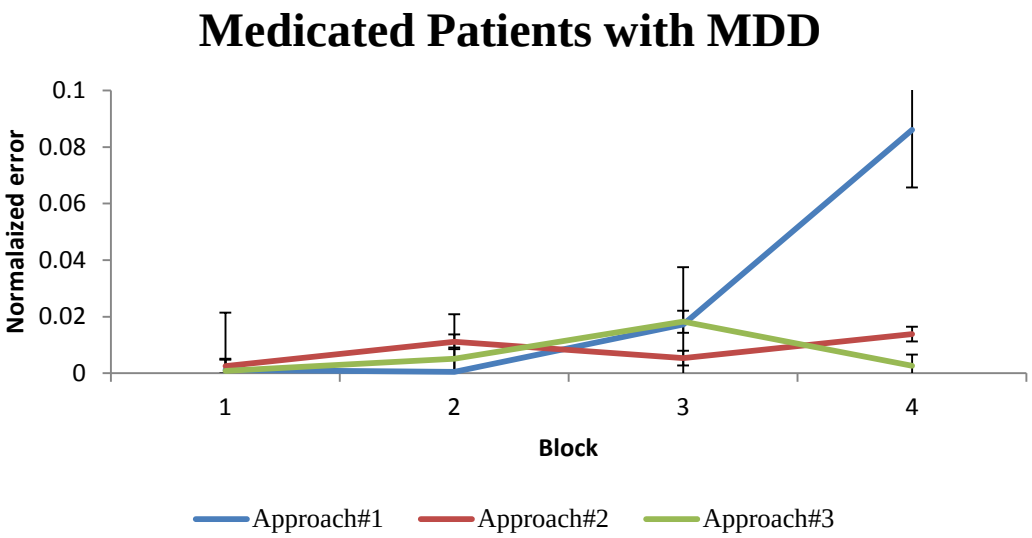

Figure 6.22 Normalized modeling error for medicated patients with SSRI. It shows at first and second block there is no significant difference between the three modeling approaches. In the third and fourth blocks, Approach#3 has the least normalized error, closely followed by Approach#2. Overall, Approch#3 provided the best fit for experimental data. Regarding Approach#1, simulation results that were used to estimate the normalized prediction error are shown in figures (6.1 & 6.2) while the simulation results are used for Apprach#2 were shown in figures (6.7, 6.9, 6.14)

Chapter Seven

Discussion and Limitations

Section 1: Discussion

Section 2: Limitations

Chapter Seven:

Discussion and Limitations:

In this chapter, we discuss the results of the different models and compare it with literature. Then, the limitations and constraints of the models were illustrated.

7.1 Discussion

We can argue that our extended actor-critic architecture successfully implemented a new control approach on reinforcement learning signals, i.e., behavioral inhibition. By generating the behavioral inhibition signal that phasically opposes the TD prediction error signal, our model, unlike the classical actor-critic architecture, addresses the role of modulatory control systems in the reinforcement learning process. From our simulation results, we found that increasing the behavioral inhibition signal affects both learning from reward and punishment feedback. On the other hand, decreasing the TD signal only affected learning from reward feedback. Given the reliance of the behavioral inhibition signal on the TD signal, the initial value of TD signal controls the intensity of behavioral inhibition signal.

Here, we discuss the simulation results of modeling MDD using the three approaches that were summarized in the methods and results chapters.

7.1.1 The Effects of DA-Only TD Signal in RL in MDD

In Approach#1 model, we simulated the effects of DA which was represented as the TD signal in learning from reward and punishment in patients with MDD. RL studies explain that there is an abnormality in prediction error signal in patients with MDD. Also, Herzallah et al., 2013 found that unmediated patients with MDD learned from punishments feedback but not from reward feedback. Furthermore, it is known that patients with MDD exhibit DA dysfunction (TD signal) [40]. Our simulation results are in line with these findings as our model learned from punishment feedback but not from reward feedback. We simulated this by limiting the value of TD signal.

Imaging studies show that SSRI antidepressants decrease the availability of D2 receptor[29]. We used this finding to simulate the effects of SSRI by decreasing the weight of activation function of D2R. Also, in our simulations, SSRI didn’t significantly improve

learning from reward feedback. Our simulation results reproduced the experimental results reported in Herzallah et al., 2013. We showed that medicated patients with MDD failed to learn from both reward and punishment feedback.

7.1.2 The effects of DA/5HT TD and Risk Prediction in RL in MDD

In the Approach#2 model, the 5HT concentration was represented as the risk prediction value (α_{D1D2}) and the DA signal was represented as the TD signal. In this model, we used different simulations to study the effects of the TD signal and the risk prediction in RL in patients with MDD. Different studies showed that patients with MDD have low 5HT concentration. Hence, we decreased the value of the risk prediction value α_{D1D2} to simulate this deficiency. However, our simulation results showed that decreasing risk prediction value dose not effect learning from reward or punishment feedback. On the other hand, to simulate the effects of SSRI antidepressants, we increased the value of the risk prediction value α_{D1D2}. Similarly, this did not affect learning from reward and punishment feedback. Only the TD signal and value function of D1R and D2R (DA parameters) controlled learning from reward and punishment feedback regardless the value of the risk factor (5HT) in this modeling approach. In the architecture of this model, the TD signal controls the value of the risk factor whereas the risk factor does not affect the TD signal. This contradicts with a wealth of studies that support the critical role of 5HT in modulating the DA signal.[37] In this approach, Balasubramani et al. considered that risk prediction (5HT) plays a critical role in the action selection network. Based on our different simulations, the risk prediction (5HT) signal in this model may affect learning from punishment but cannot affect learning from reward feedback. In conclusion, we can argue that this approach failed to simulate the interaction between TD signal and the risk prediction error in patients with MDD.

7.1.3 The Interaction of DA/5HT TD and Behavioral Inhibition in RL in MDD

As we discussed in Chapter 4, the role of 5HT is to oppose the DA signal. In this model, DA is represented as TD signal and 5HT is simulated as a behavioral inhibition signal. Experimental studies show that patients with MDD have low DA and 5HT concentrations. To simulate this interaction, we limited the value of the DA TD signal. As a result, the behavioral inhibition signal was also reduced given that the value of behavioral inhibition signal depends on the value of DA TD signal. Our simulation results were in line with

experimental results; unmedicated patients with MDD did not learn from reward feedback and but learned from punishment feedback.

To simulate the effects of SSRI antidepressants, we increased the activation function of the 5HT receptors to increase the behavioral inhibition signal. This led to inhibiting learning from both reward and punishment feedback. These results are in line with experimental results where medicated MDD patients did not learn from reward or punishment feedback. We can conclude that patients with MDD exhibit dysfunction in the DA system. This deficit can cause dysregulation in the 5HT system. Our model proposed that the behavioral inhibition signal (5HT) functions in synchrony with the TD signal (DA). Therefore, any deficit in the TD signal will affect the behavioral inhibition signal. We also found that the core module in the MDD system is the TD signal module, i.e., the DA signal, while the behavioral inhibition signal, i.e., the 5HT signal represented a reaction rather than a standalone response. Based on these findings, we can argue that SSRI antidepressants might not be the most effective medication for patients with MDD. Our model argues that that TD module plays the most critical role in controlling cognition in patients with MDD. Thus, it is very likely that DA boosting medications can have a more favorable effects on cognition in MDD.

7.2 Limitations

Our models in this form have several limitations. Similar models can simulate only RL but they cannot simulate other kinds of learning such as supervised and unsupervised learning. Moreover, our models can learn from one stimulus at a time and this assumption is oversimplified given that the learning process in humans and animals is much more complicated as they can evaluate multiple stimuli almost simultaneously.[42]

Regarding the DA only model (Approach#1), it is oversimplified as it uses one signal to simulate a cognitive task in the BG model. However, there are different signals that could contribute to simulating cognition. Although this model provides a good fit to experimental data in modeling MDD, it neglects the effects of SSRI in increasing the 5HT signal, and therefore and we assumed that the pooled TD signal can simulate the effects of SSRI.

Although second model (Approach#2) simulates the effects of 5HT as a risk factor but this assumption failed in modeling MDD as we explained earlier. Any change in the risk prediction α_{D1D2} did not improve or impair learning from reward and punishment

feedback. Only the reinforcement parameters of the TD signal could control the learning from reward and punishment feedback.

Finally, our proposed model of the interaction suggests that the role of 5HT is in expressing a behavioral inhibition signal to oppose the DA TD signal. However, different studies show that 5HT has different roles in regulating cognitive function, such as risk avoidance. Moreover, 5HT neurons have more than 17 subtypes of receptors. In this context, our model is oversimplified as we categorized these receptors into excitatory and inhibitory subtypes. Despite all limitations, our model provided the best simulation results that matches those from experimental data.

Chapter Eight

Conclusions and Future Directions

Section 1: Conclusions

Section 2: Future Directions

Chapter Eight:

Conclusions and Future Directions:

Here, we summarized the conclusion of this thesis and we propose different ways that may improve our research in future.

8.1 Conclusions

In this thesis, we proposed an extended version of actor-critic architecture. Our model simulates the effects of the interaction between two kinds of reinforcement learning signals; the TD signal and our proposed behavioral inhibition signal. The behavioral inhibition signal is synchronized to the TD signal and generated a phasic signal that matches and opposes the TD signal.

We used this extended version of actor-critic to construct a neurocomputational model of the BG. Earlier neurocomputational model of the BG simulate the DA signal as TD signal but ignore the effects of other neurotransmitters in regulating the reinforcement learning process. Moreover, previous models ignore the interaction between DA and 5HT in the BG. We constructed our model based on the anatomical pathways that govern the interaction between DA and 5HT. The TD signal represented DA while 5HT represented a reactive behavioral inhibitions signal that relies on the initiation of the DA TD signal.

We used our neurocomputational model of the BG to study the effects of DA and 5HT in learning from reward and punishment feedback in patients with MDD. We found that the DA TD played a critical role in learning from reward feedback whereas an increase in 5HT (behavioral inhibition) similar to what happens with SSRI administration inhibited learning from both reward and punishment feedback. These simulation findings are consistent with cognitive results reported by Herzallah et al. in 2013. [28]

We conclude that while the TD signal plays critical role in leaning from positive feedback, an active behavioral inhibition signal is essential in impairing learning from positive and negative feedback.

8.2 Future Directions

Future work plan will entail improving this extended version of actor-critic architecture to develop a critic-critic-actor architecture. In this proposed model, there is a bidirectional relationship between the two critics where each critic can control the value function of other critic and both critics update the actor policy. Given the perplexing complexity of our brains and cognition, one or two signals are not sufficient to simulate RL in human or animals. Future modeling efforts ought to focus on searching for and implementing other reinforcement signals that are essential to the learning process.

References

[1] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*, vol. 135. MIT press Cambridge, 1998.

[2] T. J. Sejnowski, C. Koch, and P. S. Churchland, "Computational neuroscience," *Science (80-.).*, vol. 241, no. 4871, pp. 1299–1306, 1988.

[3] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, 1988.

[4] P. Faulkner and J. F. W. Deakin, "The role of serotonin in reward, punishment and behavioural inhibition in humans: Insights from studies with acute tryptophan depletion," *Neurosci. Biobehav. Rev.*, vol. 46, no. P3, pp. 365–378, 2014.

[5] Di Giovanni, G., Di Matteo, V., & Esposito, E. (Eds.). (2008). Serotonin-dopamine interaction: experimental evidence and therapeutic relevance (Vol. 172). Elsevier.

[6] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, pp. 237–285, 1996.

[7] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, pp. 79–86.

[9] A. State and E. Action, "Reinforcement Learning: Predicting Rewards 1."

[10] J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, and R. J. Dolan, "Temporal difference models and reward-related learning in the human brain," *Neuron*, vol. 38, no. 2, pp. 329–337, 2003.

[11] S. Bhatnagar, M. Ghavamzadeh, M. Lee, and R. S. Sutton, "Incremental natural actor-critic algorithms," in *Advances in neural information processing systems*, 2008, pp. 105–112.

[12] D. Bahdanau *et al.*, "An actor-critic algorithm for sequence prediction," *arXiv Prepr. arXiv1607.07086*, 2016.

[13] Y. Chen, "Mechanisms of winner-take-all and group selection in neuronal spiking networks," *Front. Comput. Neurosci.*, vol. 11, p. 20, 2017.

[14] B. Yegnanarayana, *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[15] W. Maass, "Neural Computation with Winner-Take-All as the Only Nonlinear

Operation," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 293–299, 2000.

[16] R. Rescorla, "Rescorla-Wagner model," *Scholarpedia*, vol. 3, no. 3, p. 2237, 2008.

[17] P. Dayan, L. F. Abbott, and L. Abbott, "Theoretical neuroscience: computational and mathematical modeling of neural systems," 2001.

[18] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, "Neurotransmitters, synapses, and impulse transmission," 2000.

[19] A. A. Grace, "Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia," *Neuroscience*, vol. 41, no. 1, pp. 1–24, 1991.

[20] J.-C. Dreher, P. Kohn, B. Kolachana, D. R. Weinberger, and K. F. Berman, "Variation in dopamine genes influences responsivity of the human reward system," *Proc. Natl. Acad. Sci.*, vol. 106, no. 2, pp. 617–622, 2009.

[21] P. De Deurwaerdère and G. Di Giovanni, "Serotonergic modulation of the activity of mesencephalic dopaminergic systems: therapeutic implications," *Prog. Neurobiol.*, vol. 151, pp. 175–236, 2017.

[22] G. Di Giovanni, V. Di Matteo, M. Pierucci, and E. Esposito, "Serotonin--dopamine interaction: electrophysiological evidence," *Prog. Brain Res.*, vol. 172, pp. 45–71, 2008.

[23] L. W. Swanson, "The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat," *Brain Res. Bull.*, vol. 9, no. 1–6, pp. 321–353, 1982.

[24] G. Di Giovanni, E. Esposito, and V. Di Matteo, "Role of serotonin in central dopamine dysfunction," *CNS Neurosci. Ther.*, vol. 16, no. 3, pp. 179–194, 2010.

[25] S. Matias, E. Lottem, G. P. Dugue, and Z. F. Mainen, "Activity patterns of serotonin neurons underlying cognitive flexibility," *Elife*, vol. 6, p. e20552, 2017.

[26] P. Redgrave, "Basal ganglia," *Scholarpedia*, vol. 2, no. 6, p. 1825, 2007.

[27] N. G. Stevens, "Guidelines for the diagnosis and treatment of major depression.," *J. Am. Board Fam. Pract.*, vol. 7, no. 1, pp. 49–59, 1994.

[28] M. M. Herzallah *et al.*, "Learning from negative feedback in patients with major depressive disorder is attenuated by SSRI antidepressants," *Front. Integr. Neurosci.*, vol. 7, p. 67, 2013.

[29] E. Therapeutics, "D 2 -Like Dopamine Receptors Depolarize Dorsal Raphe Serotonin Neurons through the Activation of Nonselective Cationic Conductance," *Pharmacology*, vol. 320, no. 1, pp. 376–385, 2007.

[30] K. Gurney, T. J. Prescott, and P. Redgrave, “A computational model of action selection in the basal ganglia. I. A new functional anatomy,” *Biol. Cybern.*, vol. 84, no. 6, pp. 401–410, 2001.

[31] A. A. Moustafa and M. A. Gluck, “A neurocomputational model of dopamine and prefrontal-striatal interactions during multicue category learning by Parkinson patients,” *J. Cogn. Neurosci.*, vol. 23, no. 1, pp. 151–167, 2011.

[32] A. A. Moustafa, M. M. Herzallah, and M. A. Gluck, “Dissociating the cognitive effects of levodopa versus dopamine agonists in a neurocomputational model of learning in Parkinson’s disease,” *Neurodegener. Dis.*, vol. 11, no. 2, pp. 102–111, 2013.

[33] Ashar-Al-Natshah, “A Neuro-Computational Model of DAT1 and COMT Contributions to Learning from Positive and Negative Feedback Ashar Yousef Al-Natsheh,” 2016.

[34] N. D. Daw, S. Kakade, and P. Dayan, “Opponent interactions between serotonin and dopamine,” *Neural Networks*, vol. 15, no. 4, pp. 603–616, 2002.

[35] K. Doya, “Metalearning and neuromodulation,” *Neural Networks*, vol. 15, no. 4–6, pp. 495–506, 2002.

[36] J. Best, M. C. Reed, and H. F. Nijhout, “Computational studies of the role of serotonin in the basal ganglia,” *Front. Integr. Neurosci.*, vol. 7, p. 41, 2013.

[37] P. P. Balasubramani, V. S. Chakravarthy, B. Ravindran, and A. A. Moustafa, “A network model of basal ganglia for understanding the roles of dopamine and serotonin in reward-punishment-risk based decision making,” *Front. Comput. Neurosci.*, vol. 9, p. 76, 2015.

[38] P. P. Balasubramani, V. S. Chakravarthy, M. Ali, B. Ravindran, and A. A. Moustafa, “Identifying the basal ganglia network model markers for medication-induced impulsivity in Parkinson’s disease patients,” *PLoS One*, vol. 10, no. 6, pp. 1–23, 2015.

[39] C. Chen, T. Takahashi, S. Nakagawa, T. Inoue, and I. Kusumi, “Reinforcement learning in depression: a review of computational research,” *Neurosci. Biobehav. Rev.*, vol. 55, pp. 247–267, 2015.

[40] P. Kumar, G. Waiter, T. Ahearn, M. Milders, I. Reid, and J. D. Steele, “Abnormal temporal difference reward-learning signals in major depression,” *Brain*, vol. 131, no. 8, pp. 2084–2093, 2008.

[41] M. M. Herzallah *et al.*, “Depression impairs learning, whereas the selective

serotonin reuptake inhibitor, paroxetine, impairs generalization in patients with major depressive disorder," *J. Affect. Disord.*, vol. 151, no. 2, pp. 484–492, 2013.

[42] T. J. Bussey, R. Dias, E. S. Redhead, J. M. Pearce, J. L. Muir, and J. P. Aggleton, "Intact negative patterning in rats with fornix or combined perirhinal and postrhinal cortex lesions," *Exp. Brain Res.*, vol. 134, no. 4, pp. 506–519, 2000.