A Reinforcement Learning Model of Temporal Difference Variations for Action-Selection and Action-Execution in the Human Brain

Ashar Y. Natsheh
Palestinian Neuroscience Initiative
Al-Quds University
Jerusalem, Palestine
asharnatsheh@gmail.com

Joman Y. Natsheh
Palestinian Neuroscience Initiative
Al-Quds University
Jerusalem, Palestine
jomannatsheh@gmail.com

Aya H. Mousa
Palestinian Neuroscience Initiative
Al-Quds University
Jerusalem, Palestine
mousa.aya1@gmail.com

Mahmoud H. Al-Saheb
Palestine Polytechnic University
Hebron, Palestine
alsaheb@ppu.edu

Ahmed A. Moustafa
School of Psychology
Faculty of Society and Design
Bond University
Gold Coast, Australia
amoustaf@bond.edu.au

Mohammad M. Herzallah
Palestinian Neuroscience Initiative
Al-Quds University
Jerusalem, Palestine
mohammad.m.herzallah@alquds.edu

***Abstract*— Temporal difference (TD) prediction error signal models are instrumental in simulating brain function during reinforcement learning (RL). Recent evidence suggests a significant role of TD prediction error signals in the action-selection and action-execution brain networks. We introduce a neurocomputational model that explores TD prediction error signal variations for action-selection and action-execution. The TD prediction error signal represents the dopamine neurotransmitter the basal ganglia and prefrontal cortex brain regions. The model incorporates dopamine genetic parameters in the two networks (*COMT* gene for action-selection; *DAT1* gene for action-execution) to generate four different parameter combinations. The model simulation showed that TD signaling in both networks plays a significant role in RL under optimal conditions of medium, not high, TD signals. Moreover, each parameter combination showed a unique pattern of RL, corresponding with experimental data obtained using a computer-based RL task.**

Keywords*— *reinforcement learning, computational modeling, dopamine, feedback-based learning

I. Introduction

A. The Neuroscience Background

The catecholamine dopamine is a neurotransmitter, which is simulated as a temporal difference (TD), or prediction error, training signal in model modules that correspond to specific brain areas. Dopamine is an integral neural substrate for mediating reinforcement learning (RL) [1], which refers to learning from positive, negative, and even salient stimuli in humans and animals [2].

RL models that are based on a reward TD signal, or prediction error, which encodes the difference in dopamine neuron firing rates in response to expected vs. received (actual) rewards [3]. Various direct and indirect approaches were used to measure dopamine prediction signal levels in animals and humans [4]. In human experiments, one of the main measurement techniques is the influence of naturally occurring genetic polymorphisms in genes that code for dopamine signal degradation and clearance proteins, such as the dopamine transporter gene (*DAT1*) and catechol-O-methyl transferase (*COMT*) gene [5]. *DAT1* plays an important role in clearing dopamine from synapses in the basal ganglia (BG). Studies have shown that carriers of the 9-tandem repeat *DAT1* parameter express lower levels of the dopamine transporter protein and thus exhibit higher levels of BG dopamine, while carriers of the 10-repeat parameter express higher levels of the dopamine transporter, and lower levels of BG dopamine.

On the other hand, the *COMT* gene has a Val158Met polymorphism that affects the clearance of dopamine from synapses in the prefrontal cortex (PFC). Val parameter carriers express a higher activity COMT and therefore have relatively lower concentrations of PFC dopamine. Met carriers; on the other hand, have lower activity COMT, and higher concentrations of PFC dopamine.

Several neuro-computational and RL models have examined BG- as well as PFC-dopamine and its contribution to different cognitive functions. Although many experimental studies have examined the role of dopamine in both the BG and PFC separately, most computational models simulate dopamine effects on the function of the BG, considering the BG and the PFC as one component. For example, Gurney et al. introduce a quantitative model for BG functional anatomy. The model addresses how the BG operates based on the computational hypothesis that the BG is the neurobiological substrate for action selection. This model is built by examining a comparison between action selection neural networks and the anatomy of the BG. It is based on the idea that the BG exerts, primarily, action selection [6]. Another BG model is the one introduced by Balaraman et al., The authors developed an RL model of the BG to understand impulse control disorder (ICD) in Parkinson’s disease (PD) patients. This model assumed that the dysfunction in both dopamine and serotonin systems accounts for ICD symptoms in PD patients, in which dopamine controls reward prediction and serotonin controls punishment prediction [7]. Another model that belongs to the actor-critic family is the one by Mandali, A. et al. The authors introduced a spiking network model of the BG to relate the subthalamic nucleus-globus pallidus externus synchrony levels to exploration in PD patients. This model provides a new understanding of the role of the BG in explorative behavior, as well as the occurrence of synchrony levels in PD conditions [8].

The BG consists of different anatomical modules; striatum, sub-thalamic nucleus (STN), and globus pallidum externa and interna (GPe and GPi) [9]. The striatum receives input from the PFC. It also receives dopamine projections that control values associated with sampled rewards [8]. Striatal projections project to GPe, GPi, and STN via direct and indirect pathways, which are important for action selection dynamics [10]. The striatum has two types of a receptor expressing neurons: D1 receptor (D1R) and D2 receptor (D2R). The neurons expressed by D1 and D2 receptors are medium-spiny neurons. D1R medium spiny neurons project through the direct pathway to GPi, while D2R medium spiny neurons project via the indirect pathway to GPe [11]. Two mechanisms regulate dopamine levels in the BG-PFC interactions: (1) phasic dopamine release in the BG and the PFC caused by dopamine neuron firing, and (2) tonic dopamine release, a background dopamine release regulated in the PFC. Tonic dopamine is thought to regulate phasic dopamine levels [12,13].

B. Paper Contribution

The BG is the brain area that represents the action-execution network, while the PFC is the brain area that represents the action-selection network.

In this paper, we introduce a new neurocomputational model of RL that addresses the effect of TD signal variations (dopamine variations). Our model simulates the interaction between the two networks of action selection and action execution signaling systems. We hypothesize that the TD signals in the action selection brain network are critical for the action selection process, while the temporal difference error signals in the action execution network are critical for executing actions in learning. We also hypothesize that the interacted TD signal in both the action-execution and the action-selection networks is the signal that mediates positive learning performance. We tested 99 healthy undergraduate subjects using a computer-based RL classification task to study feedback-based learning. We simulated TD signals using parameters that represent genetic polymorphisms of dopamine clearance systems in the action selection module (*COMT* gene) and the action execution module (dopamine transporter gene (*DAT1*)). The *COMT* gene parameters are (1) the Met parameter which is associated with a lower temporal difference signal and (2) the Val parameter which is associated with a higher temporal difference signal. The *DAT1* gene parameters are (1) the 9R parameter with the higher TD signal and (2) the 10R parameter with the lower TD signal. Each run of the model has one of the parameters in the action selection network and one of the parameters in the action execution network. Thus, the model simulates four different parameter combinations: (1) 9R-Met, (2) 10R-Met, (3) 9R-Val, and (4) 10R-Val.

To our knowledge, this is the first study to simulate learning from positive and negative feedback according to the function of action-execution and action-selection network parameters.

II. Model and Methods

A. Overview

We propose a new computational network model of action selection network and execution network to identify the effects of specific prediction error signal parameters on feedback learning. We merge various algorithms to form an innovative model architecture. The basic structure of the model uses the actor-critic architecture. The artificial intelligence algorithm that we use for training the model is TD, which is a prediction algorithm used to simulate various characteristics of dopamine firing [14] and to solve RL problems that are suitable for representing reward prediction errors.

In line with the TD algorithm for the execution network and action selection network; which simulates the dopamine signal as a reward prediction error, we borrow some model elements from a recent model that studies the relationship between two neurotransmitters in the execution network [15]. Testing the results of the model will be through (1) fitting them to experimental data results from human subjects and (2) applying the model to different learning tasks.

B. Participants

We recruited ninety-nine healthy undergraduates from Al-Quds University in the West Bank, Palestine. The age of participants ranged between 18 and 24 years. Subjects were excluded if they had psychotropic drug exposure, psychiatric disorders, current pregnancy, or breastfeeding. This study was carried out at the Palestinian Neuroscience Initiative following the approvals of the Al-Quds University Research Ethics Committee with written informed consent from all subjects.

C. The Computer-Based RL Task

We used a computer-based cognitive task that tests for category learning where subjects learn from either positive or negative feedback [16]. Category learning refers to learning which category a specific card (stimulus) belongs to sun or rain. Thus, subjects were asked whether a card predicts sun or rain (Fig.1). Subjects were asked to choose whether the stimulus predicts rainy weather (Rain) or sunny weather (Sun) (Fig.1-A) in which on each trial, the participant saw one of four stimuli and was asked whether this stimulus predicts rain or sun. The critical manipulation that differentiates this task from many previous studies of probabilistic category learning is that half the four presented stimuli (S1-S4) were trained using only positive feedback for correct answers (S1-S2, (Fig.1-C) and no feedback for incorrect answers (Fig.1-B) while the other half were trained using only negative feedback for incorrect answers (S3-S4, Fig.1-D) but no feedback for correct answers (Fig.1-B). Thus, across all stimuli, the no-feedback trials are ambiguous and can occur following correct responses for negative feedback

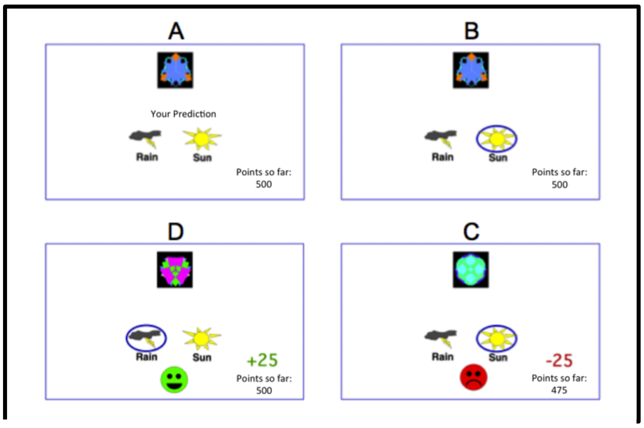

Figure 1. The probabilistic RL classification task.

stimuli or incorrect responses for positive feedback stimuli. This made it difficult for subjects to infer the implicit meaning of the no-feedback trials and encouraged them to focus, instead, on learning from the positive and negative feedback trials.

Across four blocks of 40 trials (160 trials total), subjects learned to categorize stimuli into the two outcome categories, Rain and Sun. This experimental design allows us to measure and compare individuals' sensitivity to learning from positive feedback versus negative feedback. Half of the four stimuli were trained using only positive feedback for correct answers (S1-S2) and no feedback for incorrect answers in 90% of the trials, while the other 10% received the opposite feedback (either positive feedback or no feedback). The same applies to stimuli that were trained using negative feedback for incorrect answers (S3-S4) and no feedback for correct answers. Table 1 summarizes the category and feedback structure of the probabilistic classification task according to 9:1 probability. A very similar table was reported in [3].
This probabilistic RL task has been validated and used extensively in the literature [17, 18, 19]. Imaging and animal studies have suggested that different brain structures, including both the BG and the medial temporal area, are involved in category learning; thus, this task can be used to examine the brain substrates of learning [17, 18, 19, 20].

Table 1 TASK CATEGORY AND FEEDBACK STRUCTURE

Stimulus	Probability Sun (%)	Probability Rain (%)	Feedback
S1	90	10	If correct: +25 If incorrect: null
S2	10	90	
S3	90	10	If correct: null If incorrect: -25
S4	10	90	

D. *Genotyping*

Subjects were asked to provide a 3-5 mL blood sample for genetic analysis. After working on these samples under suitable situations, genomic DNA was extracted using a specific Genomic DNA Purification Kit. Then after different operations, DNA polymerase was produced from the DNA. A DNA ladder was then used to identify the various parameters: 9-repeat and 10-repeat for the *DAT1* genotype, and Val, Met parameters for the *COMT* genotype. For each subject, four variants of interacted parameters were obtained: (1) 9R-Met, (2) 9R-Val, (3) 10R-Met, and (4) 10R-Val. There was a significant difference between the four groups in positive feedback learning performance while all groups showed very similar rates of negative feedback learning performance (Fig.2).

E. *The proposed Computational Network Model*

Model architecture follows the actor-critic architecture where the actor represents action selection and the critic represents feedback learning (Fig.3). The critic sends the signals to the actor and the actor strengthens or weakens action selection. The critic is informed whether the output of the action selected by the actor is rewarding, while it is not informed about the action itself. This model is trained using the temporal difference algorithm.

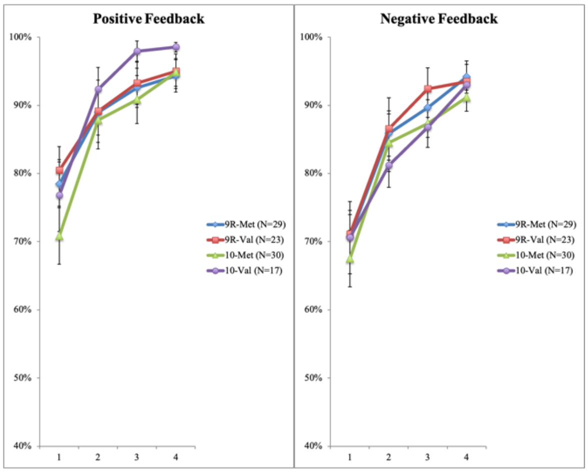

FIGURE 2. EXPERIMENTAL RESULTS FOR DAT1-COMT INTERACTION

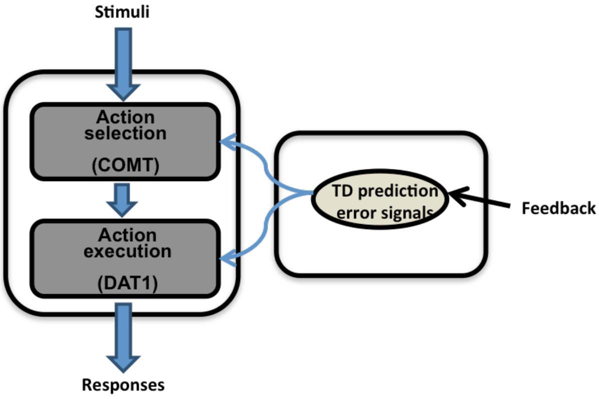

FIGURE 3. ACTOR-CRITIC ARCHITECTURE

The model has four modules: PFC/cognitive, BG/motor response, dopamine, and input. The PFC/cognitive layer is fully connected to the BG/motor response layer. The Input and PFC modules have the same number of nodes. Each unit in the Input module represents a cue/stimulus presented to the network. Input patterns presented to the network activate their corresponding units in the Input module. The Input module sends projections to the PFC layer. We use a winner-take-all network to simulate connectivity among PFC neurons. The BG module in the model learns to map input stimuli to responses [21]. Like the PFC module, we use a winner-take-all network to simulate connectivity among simulated BG neurons. At the cognitive level, the winning node represents the selected motor response. Unlike most existing models of the BG module, this module in our model learns to map representations of selected stimuli to motor responses.

In this model, the BG network is important for learning responses, whereas the PFC network is essential for action selection. This model assumes that D1 receptors in the BG network play a key role in positive feedback learning, whereas D2 receptors in the BG network have a critical role in negative feedback learning [20]. The model has several parameters that are manipulated depending on the simulated subject's dopaminergic system. We simulated the effects of phasic signals by manipulating the learning rate parameters in both the PFC and the BG modules. We also simulated the effects of tonic prediction error signals by manipulating the effects of gain parameters in sigmoidal activation in the simulated networks [22]. We simulated the implication of feedback-based learning via a cue category-learning task described in the previous section.

As mentioned above, the model has four modules interacting with each other under TD learning rules. TD algorithm implies having pairs of state-action (s, a) estimated by the value function Q(s, a), where ***s*** in our model represents the stimulus of the positive/negative feedback-based learning task which is the input to the system, ***a*** represents the selected category and Q(s, a) represents the expected reward. The expected reward is represented by the value function of the TD algorithm below [23]:

$$Q(s_t,a_t)=y_{D1}(s_t,a_t) \qquad (1)$$

The output of different types of medium spiny neurons (D1R and D2R) are represented by the variables y_{D1} and y_{D2} as follows:

$$y_{D1}(s_t,a_t)=w_{D1}(s_t,a_t)x(s_t) \qquad (2)$$

$$y_{D2}(s_t,a_t)=w_{D2}(s_t,a_t)x(s_t) \qquad (3)$$

The output of PFC nodes is represented in (4) below:

$$y_{PFC}(s_t,a_t)=w_{PFC}(s_t,a_t)x(s_t) \qquad (4)$$

Where x is modeled for the current state to be equal to 1 as in [21], and t denotes the trial.

Each stimulus has its weight for each category of the two categories A and B. This weight is updated once the system gets the feedback, where this feedback comes after the action is selected. The equation of weight update for a given pair (state, action) of different kinds of nodes (D1, D2, or PFC) can be computed as follows in (5) [14, 24]:

$$\Delta w_{node=}\ \eta_{node}\ \lambda_{node}(\delta\ (t))x(s_t) \qquad (5)$$

Where η is the learning rate for each neuron type and λ is the gain function (activation function) for different types of medium spiny neurons D1R, D2R, and PFC nodes. The values are calculated respectively as follows in (6):

$$\lambda_{node}(\delta)= (2c_1/(1+\exp(c_2(\delta+c_3))))-1 \qquad (6)$$

The δ's in weight update equations represent the classical TD error, which simulates the immediate reward for activity update. It is calculated according to (7) [14, 24].

$$\delta(t)= r-Q(s_t,a_t) \qquad (7)$$

Another form of TD used in the value function for action selection purpose is described in (8) [23].

$$\delta_Q(t)= Q_t(s_t,a_t)- Q_{t-1}(s_{t-1},a_{t-1}) \qquad (8)$$

Based on experimental findings, we simulated the 9R gene by increasing the weight of the D1 receptor for the positive feedback learning case and decreasing the D2 receptor weight for the negative feedback learning case. Accordingly, we simulated the 10R gene by decreasing the D1 weight for the positive feedback learning case, while increasing the D2 receptor weight for the negative feedback learning case.

1) *Dopamine Module*

Two types of dopamine signals are involved in the learning process; the immediate firing phasic signal and the running on the background tonic signal. We simulated the phasic signal using the TD algorithm prediction error equation (7) and (8). We simulated tonic dopamine using a gain function as proposed in the models of Moustafa et al. [23].

$$f(\delta)=1/(1+\exp(G_{tonic}(\delta)))-1 \qquad (9)$$

We simulated a rise in tonic TD prediction error signals by increasing the gain parameter Gtonic. We hypothesized that an increase in tonic TD prediction error signals firing decreases the magnitude of phasic TD prediction error signals firing as mentioned before in Taverna et al. [13]. Similarly, we simulated a reduction in tonic TD prediction error signal levels in a brain structure by decreasing the Gtonic parameter; thus, decreasing tonic signaling will increase phasic signaling magnitude [13, 22]. We simulated COMT parameters according to this gain function parameter. We simulated the Met parameter by increasing the value of the gain tonic parameter; more tonic and less phasic signals, and we simulated the Val parameter by decreasing the value of the gain tonic parameter which gives less tonic and more phasic TD prediction error signal.

2) *Basal Ganglia Module*

a) The direct pathway and indirect pathway projections to GPi.

The BG module is responsible for the motor response in the selection process; it includes different parts where the projections of the signals go through direct and indirect pathways. We borrowed the model of Balaraman et al. to manipulate our BG module and the calculations for the transition between STN and GPe, the transition of D1R neurons output via the direct pathway and the transition of D2R nodes output via the indirect pathway can be found in [10,11,15].

b) Response Selection at GPi

In GPi, the computations from direct and indirect pathways are combined in order to implement action selection and get the output as follows:

$$x^{GPi}=x^{DP}+w^{STN\text{-}GPi}y^{STN} \qquad (10)$$

$w^{STN\text{-}GPi}y^{STN}$ simulates the relative weightage of projections from STN to GPi. It is set to 1 for all nodes in the simulation.

3) *PFC Module*

We suggest a specific simulation for the PFC module layer. We suppose that the PFC module is similar to the BG module but with only one direct connection to the striatum. Further, it is mainly responsible for the action selection process. The dynamics of node projections from the PFC network to the BG network are calculated according to the equations below:

$$\tau_s\, dx_i^{PFC}/d_t = -xi^{PFC}+w^{PFC}yi^{PFC}-xi^{striatum} \qquad (11)$$

$$y^{PFC}=\tanh\ (\lambda_{PFC}\, xi^{PFC}\ ) \qquad (12)$$

We set the slope of $\lambda_{PFC} = 5$. The action selection process in the PFC that determines which action to follow is represented in (13):

$$xi^{PFC}= xi^{PFC}+wi^{PFC\text{-}striatum}yi^{PFC} \qquad (13)$$

Where wiPFC-striatum simulates the relative weightage of projections from the PFC module to the BG module. It is set to 1 for all nodes in the simulation. It is important to notice that λ's used in action selection purposes have different parameters from those used in λ's for weight update purposes. This model was conducted on a Macintosh MacBook Pro with OS X version 10.9.5, and processor 2.4 GHz Intel Core i5. Modeling was done using the MatlabR2013a environment.

III. Results

Simulation results for *DAT1* gene effects on RL gene show that the model learned from both reward and punishment trials with a percentage of 75%-90%, and gives good learning TD error. Moreover, runs with the 9R parameter learned better than those with the 10R parameter from reward trials. However, both 9R and 10R learned similarly from punishment trials.

For the simulation of *COMT* gene effects on RL, results show that the model learns from both reward and punishment trials with a percentage of 75%-90%, with a good TD error signal. In reward learning trials, subjects with the Val parameter showed enhanced learning compared to those with the Met parameter. Both groups show the same average of learning from punishment trials.

The effect of *DAT1-COMT* interaction on RL according to our model, subjects learned well from both reward and punishment trials. Fig. 4 illustrates the curves of reward and punishment learning for both. Results showed that the interaction between the two parameters of *DAT1* and *COMT* affects learning. In reward learning, the interaction between the 10R parameter and the Val parameter gives enhanced reward learning (about (80%)) as compared to the other interaction parameters, similar to behavioral results. The other three variants have almost the same performance rate (about (75%)). However, in a punishment-based learning, the model performed similarly in the four interactions.

In addition, the optimal reward responses were plotted against the dopamine signal on the second block (as the best measure of phasic dopamine); the four interacted parameters produced an inverted U-shaped function where reward accuracy and TD prediction error signal for each of the interactions was as follows: (1) 10R-Val showed the highest reward accuracy and an average value of TD prediction error signal, (2) 9R-Val showed an average reward accuracy and the highest TD prediction error signal, (3) 10R-Met showed an average reward accuracy and the lowest TD prediction error signal, and (4) 9R-Met showed an average reward accuracy and a low TD prediction error signal. These results are the same on the second, third, and fourth blocks of the model trials (Fig.5).

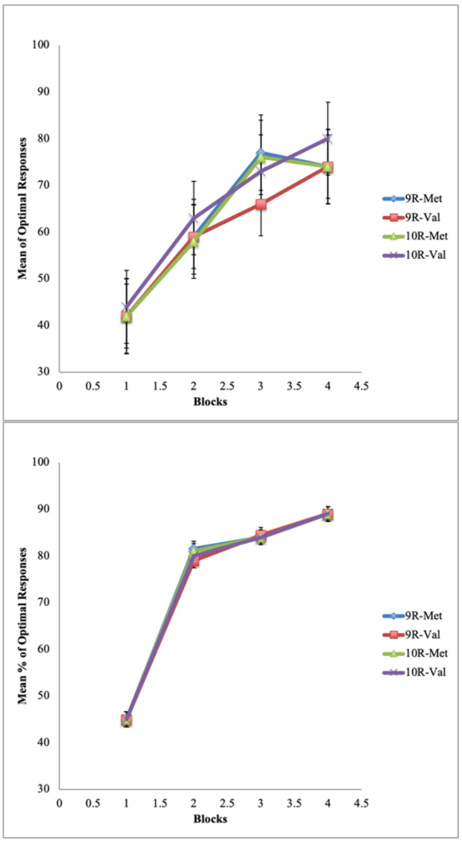

FIGURE 4. RESUTS FOR DAT1-COMT INTERACTION

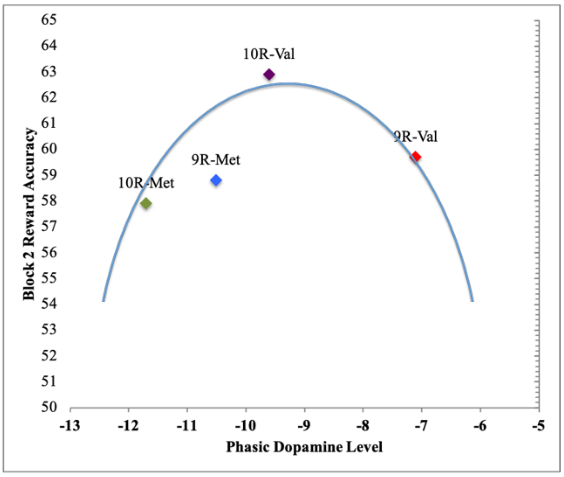

FIGURE 5. INVERTED U-SHAPED FUNCTION

IV. DISCUSSION

Our neuro-computational RL model simulates the differential effects of DAT1- COMT parameters interaction on cognition. We assume that DAT1 only exists in the execution module, while COMT only exists in the action selection module. Action execution and motor function are rooted in the basal ganglia, while higher cognitive functions are rooted in the PFC.

The proposed model provides an account of how *DAT1* and *COMT* produce an optimal learning performance according to their parameter interactions in the two modules.

In an experimental study of genetic imagining; Scott et al. explain that 9-repeat carriers have relatively higher activation of the prediction error signals and more brain activity compared to 10R subjects [27]. Further, Yacubian et al. found that 9R carriers have lower levels of DAT1 expression compared to 10R parameter carriers. Thus, 9R carriers have less TD prediction error signal clearance; thus, higher levels of phasic TD prediction error signal availability and activity [30]. The data from our model simulation for the effects of 9R-10R parameters on RL are consistent with these experiments.

In addition, our results of *COMT* simulation imply that Val parameter carriers have higher TD prediction error signal activity compared to Met parameter carriers. These findings are in line with previous findings which suggest that the *COMT* Met parameter is associated with low COMT activity; thus, an increased level of tonic TD prediction error signal and a decreased level of phasic TD prediction error signal, in the specified modules [25, 26, 27].

Further, the current proposed model of 9R-10R /Met-Val parameters interaction addresses crucial findings that are replicated theoretically and experimentally in different prior models. For example, Moustafa et al. proposed that action selection nodes underlie the action selection process, while execution nodes underlie the motor execution process. Also, it reports that the increased levels of TD prediction error signal in the simulated action selection module results in enhancing learning performance [23]. These assumptions are consistent with the results of our model.

Moreover, our model's outcomes of parameter interactions are following several experimental data. For example, both Yacubian et al. and Frank et al. suggested that reward can be modulated by the interaction between TD prediction error signal parameters [26]. Further, Dreher et al. demonstrated that the interaction between *DAT1* and *COMT* can control reward system activation, in which, the combination of *COMT*-Val and Met and *DAT1*-10R and 9R can reflect differences in signal levels [28]. The highest phasic TD prediction error signal obtained from our model is the one representing the 9R-Val variant. This result comes in agreement with strong evidence in the literature that suggests that the 9R parameter expresses a relatively higher activation of TD prediction error signal compared to the 10R parameter. On the other hand, individuals with the Val parameter have higher levels of TD prediction error signal activity compared to individuals with the Met parameter. Therefore, the interaction between 9R and Val parameters would result in the highest signal as compared to other interactions among *DAT1* and *COMT* genes [25, 26, 27, 30]. More and above, like our model, several studies have shown that the relationship between prediction error signal parameter variations and learning performance in RL is nonlinear, rather, it followed an inverted U-shaped function. Modulation of an inverted U-shaped theory has been reported previously in [29, 30, 31].

Our computational hypotheses about the different functions of *DAT1* and *COMT* variation interactions are based on findings of a previous experimental study conducted in the Palestinian Neuroscience Initiative at Al-Quds University on 131 healthy subjects. This experimental study aims to understand the underlying mechanism for RL by finding the effects of different genetic variants of naturally occurring polymorphisms.

Results showed specific learning patterns for each parameter interaction in reward and punishment feedback-based trials. On one hand, 10R-Val interaction has the best learning performance in reward, while other interacted parameters imply the same percentage of reward learning that is also less than that of the 10R-Val combination. On the other hand, the four interactions (10R-Val, 9R-Val, 10R-Met, and 9R-Met) attain the same percentage of punishment learning. Our model simulates these functional contributions of *DAT1* and *COMT* in both the execution network and action selection network and accounts for all the results obtained from the experimental study.

The proposed model has several limitations, although it can account for different RL problems tasks. For example, our model can only account for RL. It cannot model other kinds of learning. Further, this model was tested on a group of healthy subjects. Thus, it might not account for other subjects learning if they have cognitive deficits.

The model can account for TD prediction error signal function in reward and punishment trials. This might be an issue because several studies have shown that the signaling system is responsible for reward, but not punishment whereas other systems are involved in punishment learning [36]. Moreover, the model, in its current form, can only learn from one stimulus at a time. This is a very simple assumption as compared to animal and human learning which is much more complicated and might include multi-stimuli systems to learn from and adapt to environmental changes, for example, during the learning process. Some tasks in the literature included several stimuli, to pay attention to, during learning [33, 34]. However, our model cannot account for such tasks in its current form. Some studies have shown that other networks, such as the hippocampus, contribute to learning when having several configurations. Execution networks and action selection networks cannot do such configural learning without the contribution of other networks [35].

In addition, our model uses 160 trials for the tested task to learn properly. Although the learning becomes steady after 160 trials, it cannot learn when trials are less than 40. Furthermore, although the model gives a good fit for experimental data, the values assigned for model parameters such as learning rates and gain parameters are not the only parameters that can replicate experimental results, some other values of the parameters may give the same performance of the model. Moreover, they have not been used according to biological data and features. In addition to this, since the aim of the model is capturing experimental data in several variants, and more than one case; the results of the model and the performance of learning obtained from it do not fully resemble the results and performance of experimental data. However, the model was able to simulate them successfully. These limitations that we present here are reported too in similar models in the literature [23, 36, 37, 38,39].

V. Conclusion

Our model provides a RL framework to study the effect of the interaction of two TD prediction error signal parameters on learning from positive and negative feedback. Many previous studies have examined the effects of only one type of gene parameter on this type of learning. However, several experimental studies have shown that other parameters are involved in such a type of learning. Our model takes into consideration more than one parameter, including action selection *COMT* parameters in addition to *DAT1* execution selection parameters. We simulated the effect of the interaction of different parameters on RL. The two parameters of *DAT1* are (1) the 9R parameter and (2) the 10R parameter. These two variants have a key role in RL, specifically, in the action execution network and its BG module. On the other hand, the parameters of *COMT* are (1) Met and (2) Val. These parameters have a significant role in RL as well, but they are mostly implicated in action selection. The interaction between these four parameters gives the result for four variants among subjects: (1) 10R-Val, (2) 10R-Met, (3) 9R-Val, and (4) 9R-Met.

The results of running our model show that these variations can lead to differences in RL among subjects according to their parameter interaction, where the learning performance of the 10R-Val variant shows the best performance among other variants in reward learning, while the performance of punishment learning is the same for the four variants. The level of the TD signal as well as reward learning accuracy both depends on subjects' parameter variability, in which 9R-Val obtains the higher signal while 10R-Met obtains the lower signal, and the other two variants have a medium signal. These results promote experimental evidence that suggests that learning can be modulated by normal parameter variations in subjects.

References

[1] John Lenz. Reinforcement learning and the temporal difference algorithm, 2003.

[2] Umar Ali Syed. Reinforcement learning without rewards. Princeton University, 2010.

[3] Nikoletta B´odi, Szabolcs K´eri, Helga Nagy, Ahmed Moustafa, Catherine E Myers, Nathaniel Daw, Gyo¨rgy Dibo´, Annama´ria Tak´ats, Da´niel Bereczki, and Mark A Gluck. Reward-learning and the novelty-seeking personality: a between-and within-subjects study of the effects of dopamine agonists on young parkinson's patients. Brain, page awp094, 2009.

[4] Richard S Sutton and Andrew G Barto. Introduction to reinforcement learning, volume 135. MIT Press Cambridge, 1998.

[5] M Marinelli and James E McCutcheon. Heterogeneity of dopamine neuron activity across traits and states. Neuroscience, 282:176–197, 2014.

[6] Kevin Gurney, Tony J Prescott, and Peter Redgrave. A computational model of action selection in the basal ganglia. i. a new functional anatomy. Biological cybernetics, 84(6):401–410, 2001.

[7] Pragathi Priyadharsini Balasubramani, V Srinivasa Chakravarthy, Manal Ali, Balaraman Ravindran, and Ahmed A Moustafa. Identifying the basal ganglia network model markers for medication-induced impulsivity in parkinson's disease patients. PloS one, 10(6):e0127542, 2015.

[8] Alekhya Mandali, Maithreye Rengaswamy, V Srinivasa Chakravarthy, and Ahmed A Moustafa. A spiking basal ganglia model of synchrony, exploration and decision making. Frontiers in neuroscience, 9, 2015.

[9] Randall C O'Reilly, Y Munakata, MJ Frank, TE Hazy, et al. Computational cognitive neuroscience. PediaPress, 2012.

[10] V Srinivasa Chakravarthy and Pragathi Priyadharsini Balasubramani. Basal ganglia system as an engine for exploration. Encyclopedia of Computational Neuroscience, pages 315–327, 2015.

[11] Pragathi P Balasubramani, V Srinivasa Chakravarthy, Balaraman Ravindran, and Ahmed A Moustafa. A network model of basal ganglia for understanding the roles of dopamine and serotonin in rewardpunishment-risk based decision making. Frontiers in computational neuroscience, 9, 2015.

[12] Jakob K Dreyer, Kjartan F Herrik, Rune W Berg, and Jørn D Hounsgaard. Influence of phasic and tonic dopamine release on receptor activation. The Journal of Neuroscience, 30(42):14273–14283, 2010.

[13] Stefano Taverna, Ema Ilijic, and D James Surmeier. Recurrent collateral connections of striatal medium spiny neurons are disrupted in models of parkinson's disease. The Journal of neuroscience, 28(21):5504–5512, 2008.

[14] Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. Science, 275(5306):1593–1599, 1997.

[15] Pragathi Priyadharsini Balasubramani, V Srinivasa Chakravarthy, Manal Ali, Balaraman Ravindran, and Ahmed A Moustafa. Identifying the basal ganglia network model markers for medication-induced impulsivity in parkinson's disease patients. PloS one, 10(6):e0127542, 2015.

[16] Mohammad M Herzallah, Ahmed A Moustafa, Joman Y Natsheh, Salam M Abdellatif, Mohamad B Taha, Yasin I Tayem, Mahmud A Sehwail, Ivona Amleh, Georgios Petrides, Catherine E Myers, et al. Learning from negative feedback in patients with major depressive disorder is attenuated by ssri antidepressants. Front. Integr. Neurosci, 7:67, 2013.

[17] P Bolikal, CE Myers, R Patel, L Ropp, N Daw, and MA Gluck. Punishmentbased learning correlates with a putative index of serotonin in healthy young adults. In Cognitive Neuroscience Society Annual Meeting Program, page F148, 2007.

[18] Jeansok J Kim and Mark G Baxter. Multiple brain-memory systems: the whole does not equal the sum of its parts. Trends in neurosciences, 24(6):324–330, 2001.

[19] Mark G Packard, Richard Hirsh, and Norman M White. Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems. The Journal of neuroscience, 9(5):1465–1472, 1989.

[20] Russell A Poldrack and John DE Gabrieli. Characterizing the neural mechanisms of skill learning and repetition priming. Brain, 124(1):67–82, 2001.

[21] Ahmed A Moustafa, Izhar Bar-Gad, Alon Korngreen, and Hagai Bergman. Basal ganglia: physiological, behavioral, and computational studies. Frontiers in systems neuroscience, 8, 2014.

[22] Ivo Grondman, Lucian Bu¸soniu, Gabriel AD Lopes, and Robert Babuˇska. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 42(6):1291–1307, 2012.

[23] Ahmed A Moustafa and Mark A Gluck. A neurocomputational model of dopamine and prefrontal–striatal interactions during multicue category learning by parkinson patients. Journal of Cognitive Neuroscience, 23(1):151–167, 2011.

[24] JC Houk, C Bastianen, D Fansler, A Fishbach, D Fraser, PJ Reber, SA Roy, and LS Simo. Action selection and refinement in subcortical loops through basal ganglia and cerebellum. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362(1485):1573–1583, 2007.

[25] Bj¨orn H Schott, Constanze I Seidenbecher, Daniela B Fenker, Corinna J Lauer, Nico Bunzeck, Hans-Gert Bernstein, Wolfgang Tischmeyer, Eckart D Gundelfinger, Hans-Jochen Heinze, and Emrah Du¨zel. The dopaminergic midbrain participates in human episodic memory formation: evidence from genetic imaging. The Journal of Neuroscience, 26(5):1407–1417, 2006.

[26] Juliana Yacubian, Tobias Sommer, Katrin Schroeder, Jan Gl¨ascher, Raffael Kalisch, Boris Leuenberger, Dieter F Braus, and Christian Bu¨chel. Gene–gene interaction associated with neural reward sensitivity. Proceedings of the National Academy of Sciences, 104(19):8125–8130, 2007.

[27] Robert M Bilder, Jan Volavka, Herbert M Lachman, and Anthony A Grace. The catechol-o-methyltransferase polymorphism: relations to the tonic-phasic dopamine hypothesis and neuropsychiatric phenotypes. Neuropsychopharmacology, 29(11), 2004.

[28] Jean-Claude Dreher, Philip Kohn, Bhaskar Kolachana, Daniel R Weinberger, and Karen Faith Berman. Variation in dopamine genes influences responsivity of the human reward system. Proceedings of the National Academy of Sciences, 106(2):617–622, 2009.

[29] Roshan Cools, Roger A Barker, Barbara J Sahakian, and Trevor W Robbins. Enhanced or impaired cognitive function in parkinson's disease as a function of dopaminergic medication and task demands. Cerebral Cortex, 11(12):1136–1143, 2001.

[30] Graham V Williams and Patricia S Goldman-Rakic. Modulation of memory fields by dopamine d1 receptors in prefrontal cortex. Nature, 1995.

[31] Justin Zahrt, Jane R Taylor, Rex G Mathew, and Amy FT Arnsten. Supranormal stimulation of d1 dopamine receptors in the rodent prefrontal cortex impairs spatial working memory performance. The Journal of Neuroscience, 17(21):8528–8535, 1997.

[32] Pablo Henny, Matthew TC Brown, Augustus Northrop, Macarena Faunes, Mark A Ungless, Peter J Magill, and J Paul Bolam. Structural correlates of heterogeneous in vivo activity of midbrain dopaminergic neurons. Nature neuroscience, 15(4):613–619, 2012

[33] Timothy J Bussey, Rebecca Dias, Edward S Redhead, John M Pearce, Janice L Muir, and John P Aggleton. Intact negative patterning in rats with fornix or combined perirhinal and postrhinal cortex lesions. Experimental Brain Research, 134(4):506–519, 2000.

[34] Roger N Shepard, Carl I Hovland, and Herbert M Jenkins. Learning and memorization of classifications. Psychological Monographs: General and Applied, 75(13):1, 1961.

[35] Jerry W Rudy and Robert J Sutherland. The hippocampal formation is necessary for rats to learn and remember configural discriminations. Behavioural brain research, 34(1):97–109, 1989.

[36] Mark A Gluck and Catherine E Myers. Hippocampal mediation of stimulus representation: A computational theory. Hippocampus, 3(4):491–516, 1993.

[37] Andrew Amos. A computational model of information processing in the frontal cortex and basal ganglia. Journal of Cognitive Neuroscience, 12(3):505–519, 2000

[38] Roland E Suri and Wolfram Schultz. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience, 91(3):871–890, 1999.

[39] Michael J Frank. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism. Cognitive Neuroscience, Journal of, 17(1):51– 72, 2005.