A biologically inspired neuronal model of reward prediction error computation

Pramod Kaushik, Maxime Carrere, Frédéric Alexandre, Surampudi Raju

A biologically inspired neuronal model of reward prediction error computation

Pramod S. Kaushik[1,2], Maxime Carrere[3,4], Frédéric Alexandre[2,3,4], Surampudi Bapi Raju[1,5]
[1]International Institute of Information Technology, Hyderabad, India
[2]Inria Bordeaux Sud-Ouest, Talence, France
[3]LaBRI, Université de Bordeaux, Bordeaux INP, CNRS, UMR 5800, Talence, France
[4]Institut des Maladies Neurodégénératives, Université de Bordeaux,
CNRS, UMR 5293, Bordeaux, France
[5]School of Computer and Information Sciences, University of Hyderabad, Hyderabad, India.

***Abstract*—The neurocomputational model described here proposes that two dimensions involved in computation of reward prediction errors i.e magnitude and time could be computed separately and later combined unlike traditional reinforcement learning models. The model is built on biological evidences and is able to reproduce various aspects of classical conditioning, namely, the progressive cancellation of the predicted reward, the predictive firing from conditioned stimuli, and delineation of early rewards by showing firing for sooner early rewards and not for early rewards that occur with a longer latency in accordance with biological data.**

I. Introduction

One of the main motivations to design neural networks is to get adaptive functions mimicking natural learning. In addition, inspiration from biology can be deep and exploit the paradigm of neural computation to propose also a model of the underlying brain circuitry. One of the earliest attempts to understand how animals learn involved pairing an unconditioned stimulus (US) with a cue or conditioned stimulus (CS) and observing that animals start responding to the CS after some point in time [1]. This is the basis of pavlovian learning, a fundamental learning mechanism in animals, which has been addressed by several models of neural networks [2], [3]. The model described here focuses on the mechanism of reward prediction error within pavlovian learning and does not deal with other conditioning phenomena.

Concerning the link to the brain architecture, when neurons in a cerebral structure called VTA were recorded on a similar procedure in primates [4], dopaminergic neurons in VTA were found to shift their firing from the US to the CS allowing a link to the Temporal Difference (TD) learning algorithm put forward by Sutton and Barto as a possible mechanism that animals might use to learn [5]. In short, the algorithm says that the dopaminergic firing that the US causes is an error signal that the brain uses for learning. It is proposed to correspond to the Reward Prediction Error (RPE), comparing predicted and actual rewards. This error signal flows back in a recursive manner from the US to the CS, canceling the peak of dopamine at the time of the US and creating one at the time of the CS.

The TD framework has a high explanatory value while remaining simple. But, despite its usefulness, it remains a rather high level model that does not precisely account for the knowledge accumulated on the brain mechanisms associated to pavlovian learning.

The RPE computation involves a high number of neural structures which makes explaining it in a biologically plausible manner difficult. Still, considerable progress has been achieved [6], [7], [8], [9], [10], [11].

The available models possess certain mechanisms in common. One of them is the dual pathway mechanism proposing that distinct circuits cancel the peak of dopamine at the US and create another one at the CS. The models usually differ in the structures implicated and the origin of the timing signals. Some vary by the signal that cuts off the arrival of the reward: the O'Reilly PVLV model [8] uses a ramping expectation similar to the one in the TD algorithm while the Vitay model [11] proposes an oscillatory mechanism which peaks at the expected time through the ventral striatum (VS), inhibiting the reward signal. Analyzing the performances as well as the properties of these models can be interesting to decipher cerebral mechanisms but also to develop more powerful algorithms in Machine Learning.

To acquire a deeper understanding of the roles of the cerebral structures involved in the RPE computation, we need a neurocomputational model implicating them more faithfully in Pavlovian learning, The approach described here is a model that proposes a circuit involving a dissociation between magnitude expectation and timing expectation and explaining more precisely how the computation of the reward prediction error happens inside the VTA.

Our model dissociates the processing of reward magnitude and reward timing and delegates it to two structures of the Medial Temporal Lobe, the Basolateral Amygdala (BLA) and the Ventral Striatum (VS) respectively. Unlike other models, our expectation signal is based on VTA GABA neurons that ramps at the expected time to inhibit the US dopamine reward signal. These VTA GABA neurons receive their input from one of the sub-populations in the Peduncolopontine Nucleus (PPN) called PPN FT(Fixation Target), which exhibits persistent activity between the CS and US [12].

II. Model Overview

The model attempts to explain how the dopamine reward prediction error is computed in appetitive conditioning in the VTA. The model is shown in Figure 1. The functioning of the model can be explained in four phases of functioning, described in the following paragraphs, together with references to the main biological evidences supporting them:

A. US Firing and Learning

When reward is delivered, it is reported to fire the Lateral Hypothalamus (LH) and activate the LH $\rightarrow$ PPN RD (Reward delivery) $\rightarrow$ VTA Dopamine pathway resulting in US dopamine firing prior to any sort of learning [13] [14]. A pathway from LH to BLA ensures that BLA firing for CS has the exact amplitude as the US firing [15]. The VTA US dopamine firing alerts the BLA to recognize there is a reward and it progressively learns to associate the CS (reaching BLA through the inferotemporal cortex, IT) to the US gated by the US VTA dopamine [16]. There is concurrent learning in the VS for the time duration of the cue and reward delivery [17].

B. CS Firing

Another nucleus of the Amygdala, the Central Nucleus (CE) appears to get activated during this learning by the BLA [18], enabling the VTA dopamine to undergo phasic bursts of the same amplitude at the presentation of the CS through the IT $\rightarrow$ BLA $\rightarrow$ CE $\rightarrow$ PPN RD $\rightarrow$ VTA Dopamine pathway [19] at the end of conditioning. During conditioning, part of the reward magnitude is learnt resulting in partial conditioning causing partial cancellation of the US and also showing a partial firing at the arrival of the CS like other dual pathway models.

C. Expectation

The expectation signal is what ultimately cancels the predicted reward enabling the dopamine to fire its background rate at the time of reward delivery. The magnitude and timing of the reward are handled by BLA and VS respectively. The presentation of the CS, which reportedly fires the IT and thereby the Orbitofrontal Cortex (OFC) [20], activates the VS and its neurons learn the interval timing and it acts similar to a negative integrator and progressively lowers the inhibition that VS exerts on PPN FT, as reward delivery is approached. Thus, it conveys the precise time where the PPN FT can increase the inhibition through VTA GABA and cancel the dopamine.

The magnitude of expectation originates from the CS firing in the Central Amygdala (CE) and maintained in the PPN FT through a self sustaining mechanism [21] [22]. The GABA firing in the VTA is reflective of this [23] and the PPN FT integrates the magnitude from the Central Amygdala (CE) and timing information from the VS to achieve the ramping signal that encodes both time and magnitude of the reward delivery.

D. Early reward

It has been reported in the literature [4] as a hallmark of reinforcement learning in VTA that an omitted reward causes a dip of dopamine at the time of the expected reward, which can be easily explained by the dual pathway mechanism, dissociating the creation of a dopaminergic peak at the time of the CS and the cancellation of the peak at the time of the US. More difficult to explain is the fact that an early reward does not cause dips at the time of the expected US. Here it is explained using a different mechanism due to the sustained nature of the expectation signal. It posits that there is an inhibition from PPN RD to PPN FT and the early reward that flows through PPN RD resets the expectation of PPN FT.

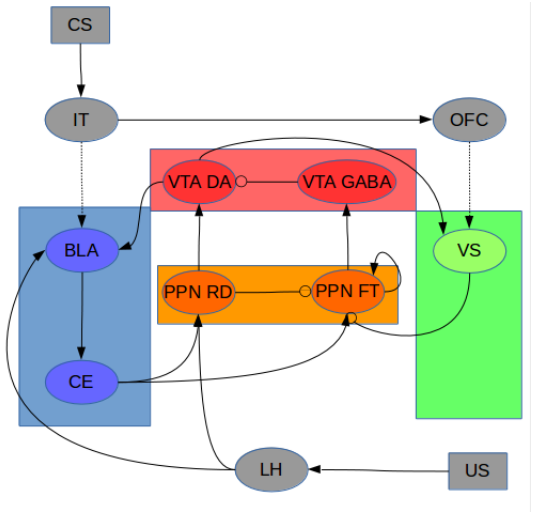

Fig. 1. Model diagram illustrating the neuronal structures and their connections involved in RPE computation. Pointed arrows represent excitatory connections, while rounded arrows represent inhibitory projections. Dashed lines represent learnable connections, while solid represent fixed connections in the model.

III. Model Description

A. Computational principles

The proposed model is composed of computational units where each unit represents a population and computes the mean activity of the population. A time-dependent firing rate describes the dynamics across time for each population. $V(t)$ represents the membrane potential of the unit and the firing rate is a positive scalar of $V(t)$ given by $U(t)$. Each unit is represented by the following equations:

$$\tau.\frac{dV(t)}{dt} = (-V(t) + g_{exc}(t) - g_{inh}(t) + B + \eta(t)) \quad (1)$$

$$U(t) = (V(t))^{+} \quad (2)$$

where τ is the time constant of the cell, B is the baseline firing rate and $\eta(t)$ is the additive noise term chosen randomly at each time step from an uniform distribution between -0.01 and 0.01. The incoming afferent currents g_{exc} and g_{inh} represent the weighted sum of excitatory and inhibitory firing

rates respectively, the weight representing the synaptic weights between the populations.

Some of the populations require an incoming tonic component converted to a short phasic transformation. This is done by the following equations

$$\tau.\frac{d\overline{x}(t)}{dt} = (-\overline{x}(t) + x(t)) \tag{3}$$

$$\phi_{\tau,k}(x(t)) = (x(t) - k.\overline{x}(t))^{+} \tag{4}$$

where $\overline{x}(t)$ integrates the incoming input $x(t)$ with a time constant τ, while $\phi_{\tau,k}(x(t))$ represents the positive part of the difference between $x(t)$ and $\overline{x}(t)$. The constant k is a constant that controls how much of the tonic component is kept, a k value of 0 indicates the entire tonic component to be preserved and a k value of 1 outputs the entire phasic component from the tonic input.

A Bound function is used when the firing of a population is described with an upper and a lower limit in certain populations

$$\psi(x) = \begin{cases} 0 & \text{if x} < 0 \\ \text{x} & \text{if } 0 < \text{x} < 1 \\ 1 & \text{if x} > 1 \end{cases} \tag{5}$$

There is also a threshold function used in some populations and it outputs 1 when the input exceeds a threshold Γ, 0 otherwise:

$$\Delta_{\Gamma}(x) = \begin{cases} 0 & \text{if x} < \Gamma \\ 1 & \text{otherwise} \end{cases} \tag{6}$$

The learning rule defined in the model is based on the Hebbian learning rule. The evolution over time of the weight $w(t)$ of a synapse between the neuronal population pre (presynaptic neuron) and the neuronal population $post$ (postsynaptic neuron) is governed by:

$$\frac{dw(t)}{dt} = (\alpha.U_{pre}(t).U_{post}(t)) \tag{7}$$

where w is the weight term, α the learning rate and $U(t)$ is indicating the firing rate of the presynaptic and postsynaptic neuronal populations.

B. Population definitions

1) Representations of inputs: IT and LH are the populations used for inputs for the CS and the US respectively and they are represented simplistically by a square wave signal as given below:

$$U(t) = I(t)^{+} \tag{8}$$

where $I(t)$ is an external input resulting either from a stimulus or from a reward.

2) Basolateral Amygdala: The BLA learns to associate the CS with the US and provides the magnitude expectation that eventually cancels the US dopamine. The BLA receives inputs from the IT.

$$\tau.\frac{dV(t)}{dt} = (-V(t) + \phi_{\tau exc,k}(g_{exc}(t)) + \eta(t)) \tag{9}$$

$$U(t) = (V(t))^{+} \tag{2}$$

with τ = 10ms, τexc = 10ms, k= 1.
The CS is learnt by updating the synaptic weights between IT and BLA and the learning rule is given by:

$$\frac{dw(t)}{dt} = D.\alpha.U_{pre}(t).(U_{mag} - U_{post}(t))^{+} \tag{10}$$

where D indicates the presence of the US corresponding to the dopaminergic neuronal modulation from the VTA, α is the learning rate equal to 0.003, U_{mag} is the magnitude of LH firing, U_{pre} and U_{post} are the firing rates of presynaptic and postsynpatic neurons respectively.

3) Central Amygdala: The CE is the output nuclei of the amygdala in this model and it projects to both the PPN nuclei, relaying information from the BLA. The CE projects to the PPN RD neurons that convey US and CS firing to the VTA dopamine neurons and PPN FT neurons that convey expectation.

The equations for the membrane potential and the firing rate are the same as Equation 9 and Equation 2 respectively, with τ = 20ms, τexc = 5ms, k = 1.

4) Peduncolopontine nucleus: The PPN has two distinct populations in this model for reward and expectation.

a) PPN RD: The PPN Reward Delivery neurons signal occurrence of the CS and the US from the CE and the LH respectively. It also has a sub-population of inhibitory neurons that inhibit the PPN FT expectation neurons.

The equations for the membrane potential and the firing rate are the same as Equation 9 and Equation 2 respectively, with τ = 5ms, τexc = 5ms, k = 1.

b) PPN FT: The PPN FT neurons encode the expectation and are subdivided into two populations, one holding the magnitude and the other delivering the expectation to the VTA GABA neurons.

PPN FT Magnitude: The PPN FT Magnitude neurons receive information from the CE and are inhibited by the PPN RD neurons. They serve to maintain a constant magnitude that is conveyed to the other population of PPN FT neurons (PPN FT Relay).

$$\tau.\frac{dV(t)}{dt} = (-V(t) + (g_{exc}(t)) - g_{inh}(t) + \eta(t)) \tag{11}$$

$$U(t) = (V(t))^{+} \tag{2}$$

with τ = 5ms.

PPN FT Relay: The PPN FT Relay population receives information from the PPN FT Magnitude population and is inhibited by the VS that conveys the timing signal and the output of these neurons is passed to the VTA GABA neurons enabling final cancellation.

The equations for the membrane potential and the firing rate are the same as Equation 11 and Equation 2 respectively,

with τ = 5ms.

5) Ventral Striatum and OFC: The Ventral Striatum handles the timing by reducing its inhibition at the required moment of reward delivery thereby conveying the precise moment to cancel the predicted reward. The OFC in this case, indicates the presence of the stimulus and has an excitatory effect on the Ventral Striatum. The timing model of VS described here is a simplified timing model comprising a negative integrator similar to the timing algorithm in [24]. The integrator here has an amplitude of 1 at the beginning of the trial and after weight updating, decreases its firing to 0 at the precise time of reward delivery adjusting its slope.

Mechanism of timing: The timing mechanism described here is an abstract method describing the time for a fixed interval with the weights encoding the duration of the interval.

$$\tau.\frac{dV(t)}{dt} = (g_{exc}(t)) - V.\Delta_\Gamma(\phi_{\tau mod,k}(g_{mod}(t)) - \overline{B}) + \eta(t)) \tag{12}$$

$$U(t) = (g_{exc}(t) - \Delta_\Gamma(\phi_{\tau mod,k}(g_{mod}(t)) - \overline{B}) - \psi(V(t))^+ \tag{13}$$

with τ = 1ms, τmod = 5ms, k = 1, Γ = 6 and $\overline{B}$ is the baseline firing rate from VTA dopamine to VS. Γ ensures a minimum threshold to be achieved for the VTA dopamine phasic firing to enable modulation. $\psi()$ is a bound function.

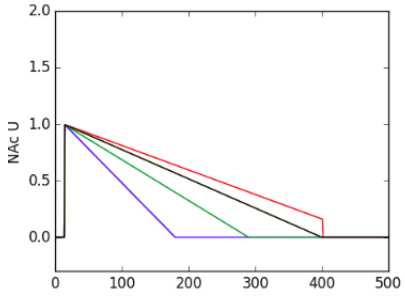

Fig. 2. The slope is decreased at every iteration until it exceeds the duration (the red line) enabling exact correction of the weight encoding the duration to be found (the black line). The colors indicate the progressive iterations

As described in figure 2, weight is updated after each iteration according to the following rule:

$$\frac{dw(t)}{dt} = (-\alpha.w + \Delta_\Gamma(U(t)).w.(U(t)/(1-U(t))) \tag{14}$$

where α is the learning rate equal to 0.4 The first term decreases the weights based on α and the weights keep decreasing until the bound is reached when $\Delta_\Gamma(U(t))$ becomes greater than 0 at the time of the reward. The correcting update is the second term of the weight updating and the slope is increased with a weight increase encoding the duration of the interval.

It should be noted that the model postulates the learning of time to be happening before the learning of value of the stimulus i.e. its magnitude.

6) VTA: The VTA in this model is divided into two populations based on the type of neurons as found in [25] and the neurobiological assumptions are derived from [26] where a ramping VTA GABA signal did not have a significant influence on tonic dopamine firing and only affected phasic dopamine. Refer [26] for details.

VTA Dopamine: The VTA dopaminergic neurons convey the final reward prediction error of the system. The VTA Dopamine neurons initially fire for the US reward which progressively gets canceled and at the same time predict the US through the phasic firing it undergoes upon arrival of a CS.

$$\tau.\frac{dV(t)}{dt} = (-V(t) + \phi_{\tau exc,k}(g_{exc}(t)) + \eta(t)) \tag{9}$$

$$U(t) = (V(t) + B)^+ \tag{15}$$

with τ = 5ms, τexc = 5ms, k = 1 and B is the baseline firing rate of the VTA Dopamine equal to 0.2

VTA GABA: The VTA GABA neurons encode expectation and receive their inputs from the PPN FT neurons.

$$\tau.\frac{dV(t)}{dt} = (-V(t) + (g_{exc}(t)) + \eta(t)) \tag{16}$$

$$U(t) = (V(t))^+ \tag{2}$$

with τ = 20ms

This model is implemented in Python, and is using the DANA library for neuronal computation [27].

IV. Evaluation of the Model

The paradigm used to evaluate the model is a simple CS-US associative learning task and considers also how the expectation cancels out the dopamine peak at the time of the reward. The trial duration is 500 time steps with each time step corresponding to 1ms. The stimulus is presented at the 10th time step and is kept switched on till the arrival of the reward at the 400th time step (400ms). The reward and the stimulus have by default a magnitude of 1. The number of trials for the entire conditioning to happen was 9 trials.

Architectural parameters		
Parameter	Meaning	Value
US input_size	size of input vectors from LH	1
CS input_size	size of input vectors from IT	4
VTA Dopamine_size	number of neurons in VTA Dopamine	10
VTA GABA_size	number of neurons in VTA GABA	5
BLA_size	number of neurons in BLA	1
CE_size	number of neurons in CE	1
OFC_size	number of neurons in OFC	1
PPN RD_size	number of neurons in PPN RD	4
PPN FT_size	number of neurons in PPN FT	4
PPN Magnitude_size	number of neurons in PPN Magnitude	4
Equation parameters		
BLA_CE	constant weights from BLA to CE	0.15
LH_PPN_RD	constant weights from LH to PPN_RD	1.2
LH_BLA	constant weights from LH to BLA	1
IT_OFC	constant weights from IT to OFC	.25
CE_PPN_RD	constant weights from CE to PPN_RD	2
CE_PPN_Mag	constant weights from CE to PPN_Mag	.3
PPN_RD_PPN_Mag	constant weights from PPN_RD to PPN_Mag	0.8
PPN_RD_VTA_Dop	constant weights from PPN_RD to VTA_Dop	1
PPN_Mag_PPN_Rel	constant weights from PPN_Mag to PPN_Rel	0.2
VS_PPN_Rel	constant weights from VS to PPN_Rel	1
PPN_Rel_VTA_GABA	constant weights from PPN_Rel to VTA_GABA	0.25
VTA_Dopamine_BLA	constant weights from VTA_Dopamine to BLA	1
VTA_Dopamine_BLA	constant weights from VTA_Dopamine to VS	1
VTA_GABA_VTA_Dopamine	constant weights between VTA_GABA and VTA_Dopamine	0.2
OFC_VS	initial weights between OFC and VS	0.006
IT_BLA	initial weights between OFC and VS	0.01

Fig. 3. Parameters describing network architecture and parameters used in activation and learning rules.

A. Initial Trial

During the first trial of the conditioning (Figure 4), the BLA hasn't yet learnt to associate the CS with the US. The BLA recognizes the reward signal through LH and the VTA. The CS firing gradually builds up and encodes the final magnitude of the US at the end of conditioning. The VTA fires on the delivery of the reward as shown in Figure 4. The Ventral Striatum is yet to learn the timing of the interval duration and the VTA dopamine US firing enables the Ventral Striatum to learn the duration of the interval in subsequent trials and this learning of the time happens before the learning of the magnitude. Since the CS has not been recognized as rewarding, there is no expectation at the arrival of the CS.

B. Partial Conditioning

After a few trials (four in our simulations), the magnitude of the reward is partially encoded in the BLA and the BLA fires upon the arrival of the CS. The synaptic weights between IT and BLA are updated after each rewarding trial. The learning of time of the US happens before the learning of magnitude and at this stage, the interval time has been completely learnt. The Ventral Striatum has no inhibition at the end of the interval time and allows the entire expectation to inhibit the VTA dopamine neurons. This results in partial CS firing and partial cancellation of dopamine. This partial cancellation is achieved through partial expectation developed as a result of CS firing through PPN FT neurons activating the VTA GABA. Partial conditioning results in both the CS and the US showing some firing (Figure 5).

C. Complete Reward cancellation

At the end of conditioning, the reward has been fully learnt and the VTA GABA neurons ramps to its maximum, canceling the entire US signal coming from the LH. The BLA neurons that drive firing for the CS has now encoded the magnitude of the US and drives the expectation that finally maintains the background firing of the dopamine at the point of reward delivery as shown in Figure 6.

D. Early Reward

The model is consistent with physiological data that does not treat all early rewards as the same. The expectation is substantial even at the half way point and dopamine firing is not observed for those "early" rewards that have a longer latency (that come after the half way mark). The earlier rewards fire more than the ones with longer latency and the earlier rewards that fire do not fire as much as the "unpredicted" reward. [28]

The Figure 7, shows a reward delivered before the half way point (at the 100th time step) invokes a dopamine firing but less than the firing showed for an "unpredicted" reward while Figure 8, shows an early reward delivered after the half way mark (at the 300th time step), does not invoke any firing.

V. Discussion

The model adds to the literature of computing the reward prediction error and is different from the other dual pathway models owing to its dissociation of magnitude and timing signals [6], [8]. This dissociation results in a distributed manner of processing and would enable the system to be more robust and maintain the information of one dimension even if the other is changed, say even if the previous time of the interval duration is changed for the same stimulus, the value of the stimulus need not be relearned again enabling faster transitions compared to the Temporal Difference (TD) algorithm. The simulated model has a continuous representation of time and evaluates how the inhibitory signal is modulated throughout the duration of CS and US differing it from other models such as PVLV [8] which has a single point of inhibition at the expected time of the US. The model also attempts to explain the early reward scenario in a biologically plausible manner by showing that not all early rewards are the same and the ones with longer latency (after the halfway mark of the interval) don't cause any firing at all.

Since the model is also an attempt to understand the circuits behind reward processing in the brain, it makes the following predictions:

1) CE and PPN FT encode magnitude of expectation
   This hypothesis does not have a separate CS-US processing and suggests the canceling of the US reward has its origins in the CS firing, one of the predictions is that inhibiting or partial lesions in the Central Amygdala (CE) in a conditioning task at the time of the CS would decrease the PPN FT expectation and possibly result in a positive prediction error in VTA dopamine instead of complete cancellation.
2) PPN through VTA GABA cancels dopamine
   The expectation that is encoded in PPN FT firing through VTA GABA provides the final cancellation signal required for maintaining the baseline firing of dopamine. There is a specific projection from PPN to VTA GABA that should implement this.

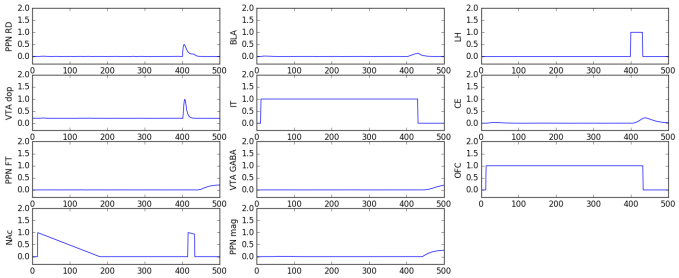

Fig. 4. Initial Trial shows arrival of the reward in LH (US) and the VTA dopamine neurons firing as a result. There isn't any firing in the PPN FT or VTA GABA since the CS is yet to be recognized as rewarding and no expectation has developed as a result. The VTA US firing also shows in the VS which enables the VS to finally learn the interval duration. All the firing rates of the populations are scaled down to fall between 0 and 1.

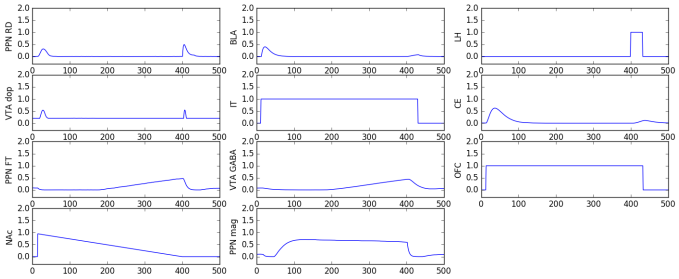

Fig. 5. Partial Conditioning shows the magnitude partially learnt but the timing fully learnt. The partial magnitude learning is reflected in an expectation that shows in the firing of the CS and in partial cancellation at the US resulting in the twin peaks as observed in physiological data

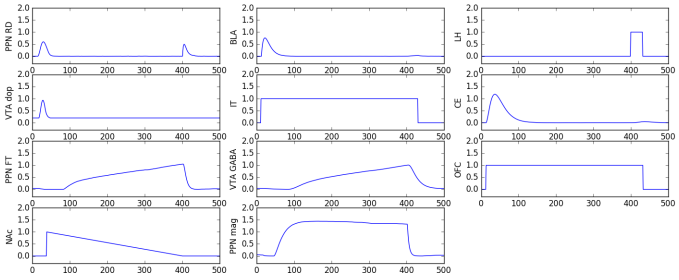

Fig. 6. Complete Reward Cancellation shows the end of conditioning when the magnitude of the stimulus has been fully learnt and the CS firing at the VTA has the same amplitude as the previous US firing, the firing for expectation from VTA GABA is maximal at this point and ramps to cancel the expected US at the point of reward delivery

3) VS encodes timing

The model hypothesizes VS to learn the timing and progressively lower the inhibition that it exerts on PPN FT causing the VTA GABA to ramp and cancel the final reward.

4) Early Reward

Upon receiving early reward through the LH $\rightarrow$ PPN RD pathway, these PPN RD neurons should inhibit and reset the PPN FT neurons canceling the expectation. PPN FT neurons wouldn't show any activity after the arrival

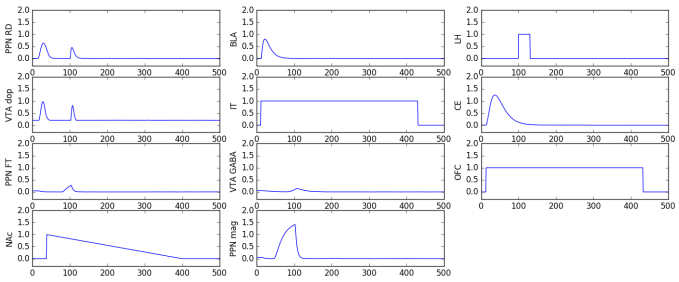

Fig. 7. Sooner Early Reward invokes a firing but less than the unpredicted reward in the VTA dopamine neurons

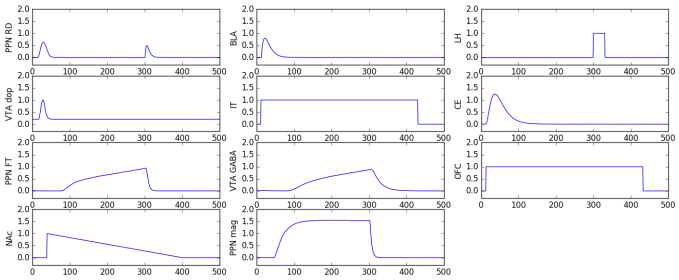

Fig. 8. The Early Reward with longer latencies does not invoke any firing in VTA dopamine neurons since the expectation from VTA GABA is already substantial at this point and does not allow the VTA dopamine to fire

of the reward due to this effect.

5) Learning of Time before Learning of Magnitude
   The model postulates that interval timing is learnt before magnitude of the stimulus.

The effects of the VS lesion experiments described in [17] could be reproduced by this model. A VS lesion in this model would make the system lose its ability to track time but magnitude expectation will be preserved, resulting in the VTA GABA neurons undergoing elevated firing throughout the duration of the CS and US. Thus, the VTA dopamine neurons would only fire for those rewards that have magnitude greater than the expected reward and not for early rewards as described in the study.

The model also contends that the computation of reward processing is highly complex in the brain and could involve synchronous circuits that excite VTA dopamine neurons and simultaneously inhibit the VTA GABA neurons to cause phasic firing in VTA dopamine neurons [29]. Such a synchrony is not in this model's current scope. Recent studies also show LH has direct afferents to VTA GABA and reward firing could be caused by the GABAergic neurons in the LH inhibiting the VTA GABA neurons [30] and thus disinhibiting VTA dopamine neurons. The model speculates the same mechanisms of this model could be used in such a case where the integration of time and magnitude could converge on VTA GABA instead of PPN FT which would then only have the magnitude component and the VS still providing the timing information through direct afferents to VTA GABA. Early rewards with longer latencies would still be unable to cause VTA dopamine firing since the LH GABA might not be able to fully suppress the VTA GABA to cause a disinhibition in VTA dopamine.

VI. Implications and Future Work

This model is a biologically inspired way of looking at reward prediction errors. It tries to approximate behavior in a computational manner but is also an attempt to explain the circuit and the computations involved in reward prediction error processing. It aims to describe the precise nature of how the circuits of the brain solve the phenomenon of classical conditioning. It could play the "critic" in the actor-critic reinforcement learning paradigm and could also be used to extend to those cases where the reward prediction error computation does not happen properly (for e.g., in addiction or anhedonia) [31]. The model currently does not account for aversive conditioning and hopes to include it in the future.

VII. Conclusion

The neurocomputational model described here represents a model-free reinforcement learning system and learns the CS-US association in classical conditioning. It achieves this by proposing a dissociation between magnitude and timing expectation separating it from traditional reinforcement learning approaches. The model posits the brain could be solving the dimensions involved in classical conditioning separately in such a distributed manner. Interestingly, in a modular view, such a system, having the components required to process a given natural phenomena broken down into its elemental dimensions, could enable the same dimensions to be combined with new elemental dimensions to process other natural phenomena.

Whereas the model remains simple in its nature and description and is based on homogenous units, it can reproduce a variety of aspects in classical conditioning in accordance with physiological data, namely, predicting the US by CS firing, canceling the expected US by a ramping expectation, the twin peaks of the CS firing and US firing during partial conditioning, sooner early rewards causing firing but not long latency early rewards and sooner early rewards not firing as much as "unpredicted" reward.

VIII. Acknowledgements

The authors would like to acknowledge the following grants which have been a major support to this research.

- Indo-French CEFIPRA Grant for the project Basal Ganglia at Large (No. DST-INRIA 2013-02/Basal Ganglia dated 13-09-2014)
- *Internships programme* at INRIA, 6 month Internship with Team Mnemosyne at INRIA Bordeaux - Sud-Ouest

References

[1] I. P. Pavlov, *Conditioned Reflexes (V.Anrep, trans.)*. London: Oxford University Press", 1927.

[2] N. Schmajuk and J. DiCarlo, "Stimulus configuration, classical conditioning and the hippocampus," *Psychological Review*, vol. 99, pp. 268–305, 1992.

[3] M. E. Le Pelley, "The role of associative history in models of associative learning: a selective review and a hybrid model." *The Quarterly Journal of Experimental Psychology*, vol. 57, no. 3, pp. 193–243, Jul. 2004. [Online]. Available: http://dx.doi.org/10.1080/02724990344000141

[4] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An introduction*. MIT Press, 1998.

[6] J. Brown, D. Bullock, and S. Grossberg, "How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues," *The journal of neuroscience*, vol. 19, no. 23, pp. 10 502–10 511, 1999.

[7] R. E. Suri and W. Schultz, "Temporal difference model reproduces anticipatory neural activity," *Neural computation*, vol. 13, no. 4, pp. 841–862, 2001.

[8] R. C. O'Reilly, M. J. Frank, T. E. Hazy, and B. Watz, "Pvlv: the primary value and learned value pavlovian learning algorithm." *Behavioral neuroscience*, vol. 121, no. 1, p. 31, 2007.

[9] C. O. Tan and D. Bullock, "A dopamine–acetylcholine cascade: simulating learned and lesion-induced behavior of striatal cholinergic interneurons," *Journal of neurophysiology*, vol. 100, no. 4, pp. 2409–2421, 2008.

[10] T. E. Hazy, M. J. Frank, and R. C. OReilly, "Neural mechanisms of acquired phasic dopamine responses in learning," *Neuroscience & Biobehavioral Reviews*, vol. 34, no. 5, pp. 701–720, 2010.

[11] J. Vitay and F. H. Hamker, "Timing and expectation of reward: a neurocomputational model of the afferents to the ventral tegmental area," *Frontiers in Neurorobotics*, vol. 8, no. 4, 2014.

[12] K.-i. Okada, K. Toyama, Y. Inoue, T. Isa, and Y. Kobayashi, "Different pedunculopontine tegmental neurons signal predicted and actual task rewards," *The Journal of Neuroscience*, vol. 29, no. 15, pp. 4858–4870, 2009.

[13] K. Semba and H. C. Fibiger, "Afferent connections of the laterodorsal and the pedunculopontine tegmental nuclei in the rat: A retro-and antero-grade transport and immunohistochemical study," *Journal of Comparative Neurology*, vol. 323, no. 3, pp. 387–410, 1992.

[14] S. Lokwan, P. Overton, M. Berry, and D. Clark, "Stimulation of the pedunculopontine tegmental nucleus in the rat produces burst firing in a9 dopaminergic neurons," *Neuroscience*, vol. 92, no. 1, pp. 245–254, 1999.

[15] P. Sah, E. L. Faber, M. L. De Armentia, and J. Power, "The amygdaloid complex: anatomy and physiology," *Physiological reviews*, vol. 83, no. 3, pp. 803–834, 2003.

[16] S. Bissière, Y. Humeau, and A. Lüthi, "Dopamine gates ltp induction in lateral amygdala by suppressing feedforward inhibition," *Nature neuroscience*, vol. 6, no. 6, pp. 587–592, 2003.

[17] Y. K. Takahashi, A. J. Langdon, Y. Niv, and G. Schoenbaum, "Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat vta depends on ventral striatum," *Neuron*, 2016.

[18] J. LeDoux, "Emotion circuits in the brain," *Annu. Rev. Neurosci.*, vol. 200, pp. 155–184, 2000.

[19] K. Cheng, K. Saleem, and K. Tanaka, "Organization of corticostriatal and corticoamygdalar projections arising from the anterior inferotemporal area te of the macaque monkey: a phaseolus vulgaris leucoagglutinin study," *The Journal of neuroscience*, vol. 17, no. 20, pp. 7902–7925, 1997.

[20] S. Carmichael and J. L. Price, "Sensory and premotor connections of the orbital and medial prefrontal cortex of macaque monkeys," *Journal of Comparative Neurology*, vol. 363, no. 4, pp. 642–664, 1995.

[21] K.-i. Okada and Y. Kobayashi, "Reward prediction-related increases and decreases in tonic neuronal activity of the pedunculopontine tegmental nucleus," *Frontiers in integrative neuroscience*, vol. 7, p. 36, 2013.

[22] Y. Kobayashi and K.-I. Okada, "Reward prediction error computation in the pedunculopontine tegmental nucleus neurons," *Annals of the New York Academy of Sciences*, vol. 1104, no. 1, pp. 310–323, 2007.

[23] H.-J. Yau, D. V. Wang, J.-H. Tsou, Y.-F. Chuang, B. T. Chen, K. Deisseroth, S. Ikemoto, and A. Bonci, "Pontomesencephalic tegmental afferents to vta non-dopamine neurons are necessary for appetitive pavlovian learning," *Cell Reports*, vol. 16, no. 10, pp. 2699–2710, 2016.

[24] F. Rivest and Y. Bengio, "Adaptive drift-diffusion process to learn time intervals," *arXiv preprint arXiv:1103.2382*, 2011.

[25] J. Y. Cohen, S. Haesler, L. Vong, B. B. Lowell, and N. Uchida, "Neuron-type-specific signals for reward and punishment in the ventral tegmental area," *Nature*, vol. 482, no. 7383, pp. 85–88, 2012.

[26] N. Eshel, M. Bukwich, V. Rao, V. Hemmelder, J. Tian, and N. Uchida, "Arithmetic and local circuitry underlying dopamine prediction errors," *Nature*, 2015.

[27] N. P. Rougier and J. Fix, "DANA: Distributed (asynchronous) Numerical and Adaptive modelling framework," *Network: Computation in Neural Systems*, vol. 23, no. 4, pp. 237–253, Dec. 2012.

[28] C. D. Fiorillo, W. T. Newsome, and W. Schultz, "The temporal precision of reward prediction in dopamine neurons," *Nature neuroscience*, vol. 11, no. 8, pp. 966–973, 2008.

[29] M. Aggarwal, B. I. Hyland, and J. R. Wickens, "Neural control of dopamine neurotransmission: implications for reinforcement learning," *European Journal of Neuroscience*, vol. 35, no. 7, pp. 1115–1123, 2012.

[30] E. H. Nieh, C. M. Vander Weele, G. A. Matthews, K. N. Presbrey, R. Wichmann, C. A. Leppla, E. M. Izadmehr, and K. M. Tye, "Inhibitory input from the lateral hypothalamus to the ventral tegmental area disinhibits dopamine neurons and promotes behavioral activation," *Neuron*, 2016.

[31] A. D. Redish, "Addiction as a computational process gone awry," *Science*, vol. 306, no. 5703, pp. 1944–1947, 2004.