Primate Motor Cortical Activity Displays Hallmarks of a Temporal Difference Reinforcement Learning Process

1st Venkata S Aditya Tarigoppula
dept. Biomedical Engineering
University of Houston
Houston, USA
aditya30887@gmail.com

2nd John S Choi
dept. Physiology
State University of New York
Downstate Medical school
NY, USA
jschoi831@gmail.com

3rd John P Hessburg
dept. Physiology
State University of New York
Downstate Medical school
NY, USA
jhess90@gmail.com

4th David B McNiel
dept. Physiology
State University of New York
Downstate Medical school
NY, USA
Dave.B.Mcniel@gmail.com

5th Brandi T Marsh
dept. Physiology
State University of New York
Downstate Medical school
NY, USA
btm2002@gmail.com

6th Joseph Thachil Francis
dept. Biomedical Engineering
University of Houston
Houston, USA
jtfrancis@uh.edu

Abstract* — Reinforcement learning (RL) models comprehensively describe neural dynamics of multiple brain regions at several spatiotemporal scales during reinforcement-based learning. One of the key components of RL models that capture the expected cumulative reward from a given state is the State-Value Function (SVF). We utilized a non-human primate (NHP) subject (Bonnet macaque) that was implanted with a 96-electrode array in the primary motor cortex. The NHP performed a reward level cued reaching task manually and passively observed such a task. Here we show that primary motor cortical (M1) activity in an NHP resembles an RL process, encoding a state value function. The motor cortex responds to reward delivery (US, unconditioned stimulus) and extends this state-value-related response earlier in a trial, becoming predictive of the expected reward when indicated by an explicit cue (CS, conditioned stimulus). This SVF is observed in tasks performed both manually and passively, that is, without agency. Here, we used the Microstimulus Temporal Difference RL (MSTD) model, reported to accurately capture RL-related dopaminergic activity, to parsimoniously account for both the phasic and tonic M1 reward-related neural activity. In the future, we will use this state value information towards autonomously updating brain-machine interfaces (BMIs) to maximize the total subjective reward expectation of the NHP user.

Keywords *—reinforcement learning, motor cortex, BMI, reward, mirror neuron*

I. INTRODUCTION

Primates can learn from observation and through direct experience. This learning leads to subjective environmental state-values and/or state action-values[1], where the individual's internal state, such as level of hunger, thirst, and preferences, can be included in the full state-space that informs the animal's (agent's) behavior (policy). Where state-space is the space of all states, and a state gives us all the information needed to best predict the subsequent state. In this work, the state space will include task-relevant variables (see methods). Reinforcement learning (RL) provides a parsimonious theory and set of models for learning both state values and state-action values, for either direct experience by the agent or via observation of other agents. We wished to determine if neural activity in the primary motor cortex (M1) encoded a State-Value-Function (SVF) as expected for an RL agent, as reward modulation of M1 has been demonstrated by several groups [2]–[6]. In our previous publications, we reported on neural correlates of reward expectation and reward delivery without regard to the temporal structure of the neural correlates, which we start to rectify here.

Dopaminergic neurons were one of the earliest reported brain structures to represent reward [7]. Dopaminergic neurons have been reported to respond initially to reward delivery (US, unconditioned stimulus) and, with experience in an operant conditioning task, shift their response earlier towards the reward-predicting stimuli (CS, conditioned stimulus)[1], [8]. Neural activity that predicts reward has been reported in the striatum[9]. As cortical structures communicate reciprocally with the basal ganglia directly (mesocortical pathway) or indirectly (nigrostriatal and striato-thalamo-cortical pathways), it is not surprising that reinforcement activity has been observed in cortical regions upstream of M1, such as orbitofrontal, prefrontal, frontal eye fields, supplementary eye fields, rostral supplementary motor areas and premotor cortices[10]. For example, the rat orbitofrontal cortex exhibits a shift in the initiation of chronic neural activation from the US to the CS with repeated performance of a discrimination-learning task with cued rewarding and aversive reinforcement[11]. These cortical regions continued to respond to the predicted reward delivery

post-conditioning, unlike dopaminergic neurons, which we see in the current work.

Dopamine receptors are found in M1[12] and are necessary for long-term potentiation[13], [14] in M1. Tonic dopaminergic activity has been shown to act as a value function[15]. In this regard, dopamine can "charge" the nervous system, acting as a motivational signal[16]. Thus, dopamine could have at least two influences on M1 - gating synaptic plasticity toward sensorimotor learning and "charging" neural activity, possibly priming the system for action. This latter activity could resemble or be a product of the SFV, which may, in part, drive the arousal/motivational state of the agent. Our lab, and others, have shown that reward modulates units and local field potentials in the primary- and pre-sensorimotor cortices (M1, S1, PMd, PMv) of NHPs[2]–[6], [17], [18]. None of the above work addressed the derivation and temporal evolution of the state-value predictive signal in M1 during CS-US conditioning. The work presented here attempts to fill this gap. It explores the application of a temporal difference reinforcement learning model (TDRL), proficient in capturing the dynamics of dopaminergic neurons, to capturing reward-related M1 neural dynamics. The results of this work are also available in a pre-print version [19].

II. Methods

A. Experimental design

The data utilized in this paper comes from 1 NHP's data (NHP A, male, *Bonnet macaque*) that has been reported previously, showing static differences between rewarding (R1) and non-rewarding (R0) trials[2]. Our work utilized a Reward Cued Reaching Task (ReCRT). NHP A sat in a primate chair with his right arm in an exoskeletal robotic manipulandum (KINARM BKIN). NHP A was proficient at performing an 8-target CRT task manually with his right arm before implantation in the left M1. NHP A was then introduced to the ReCRT (Fig.1) [2]. A hand-feedback cursor was displayed on a screen in the horizontal plane just above his right arm in alignment with his right hand during the manual task. In ReCRT, the reward level was cued first via the color of the center hold target, hold time 325ms, and subsequently by the peripheral reaching target, while the disappearance of the center hold target acted as the go cue.

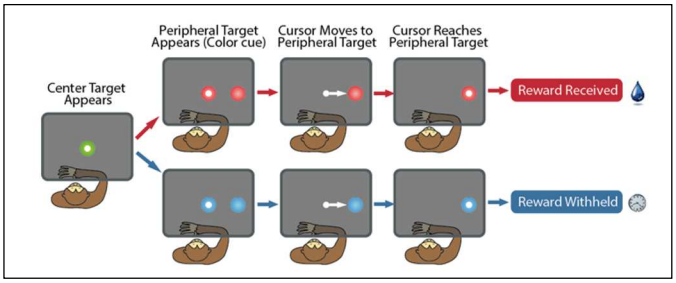

Figure.1. Observational Reward Cued Reaching Task (ReCRT). A color cue indicated the reward level for that trial, and the NHP simply had to watch the white cursor move to the right-hand target to receive a reward or complete a non-rewarding trial.

The NHP was required to move its hand to the peripheral target following the go cue and hold that position for 300ms in the manual ReCRT. In the observational ReCRT the NHP passively observed the cursor's constant speed movement to the peripheral target without any hold times. A successful reach to and hold on the peripheral target resulted in either a juice reward delivery (R1) or no reward (R0) depending on the color cue, the conditioned stimulus (CS), representing the current trial reward level (R0 vs. R1). Juice was delivered via a straw-like tube attached to a gravity feed dispenser with a control solenoid, Crist Instruments, at the end of rewarding trials. Progression from the current reward level trial to the next was only allowed following successful trials. We had to utilize this repetition scheme; otherwise, the NHP would skip R0 trials and wait for R1. NHP A performed manual and observational ReCRT with similar results for both. We only present the observational data here due to space limitations.

B. Implantation and neurophysiology

Micro-electrode arrays (4 MEAs 1.5mm length, 96 active channels, Blackrock microsystems) were implanted after NHP A could perform the manual version of the non-reward-level cued reaching task proficiently. Chhatbar et al.[20] describes the implantation procedure in detail. In summary, electrode arrays were implanted in rostral M1 corresponding to the contralateral arm/hand region. The array implantation procedure was as follows - 1) dissection of the skin above the skull, 2) craniotomy, 3) durotomy, 4) cortical probing in the primary sensory cortex (S1) to accurately locate the hand and the arm region 5) implantation of the electrode array in the forearm and hand region of the primary motor cortex and 6) closure. All surgical procedures were performed under the guidance of the State University of New York Downstate Medical Center Division of Comparative Medicine and were approved by the Institutional Animal Care and Use Committee in compliance with National Institutes of Health guidelines for the care and use of laboratory animals. Aseptic conditions were maintained throughout the surgery. Ketamine was used to induce anesthesia, and isofluorane and fentanyl were used to maintain the animal under anesthesia during the surgery. Possible cerebral swelling was controlled by using mannitol and furosemide, whereas dexamethasone was used to prevent inflammation during the surgery.

C. Data Analysis

The correlation of units with 'reward', a variable defined as +1 or -1, was assigned to each bin of R and NR trials, respectively. Data from cue presentation to the corresponding trial completion was considered for each trial. The correlation coefficient (*corrcoef*, MATLAB) was computed between each unit's firing rate and the reward variable in each block. The units were then sorted in a descending manner with respect to the correlation coefficient values. The unit with the highest positive correlation coefficient correlated most with rewarding trials. In contrast, the unit with the highest negative correlation coefficient correlated most with non-rewarding trials. The significance of the correlation coefficient was set at $p < 0.05$; see Fig.2.b.2-3.

Neural data were binned at 50ms. The average activity across R and NR trials for each unit was causally smoothed (500ms window) and concatenated to form a vector Y for a given unit. Subsequently, normalization was performed by using the below equation, resulting in a range from 0 to 1.

NIH 1R01NS092894-01, NSF IIS-1527558, DARPA N66001-10-C-2008

$Normalized\ Y = \frac{Y - \min(Y)}{\max(Y) - \min(Y)}$. ' $Normalized\ Y$ ' was decatenated to obtain the normalized average activity of the same unit across R and NR trials. The process was repeated for all units in the neural ensemble.

Microstimulus (MS) Temporal Difference (TD) Reinforcement Learning (RL) (MSTD-RL): We wished to determine if the neural activity changes we observed in M1 could be related to an RL process. We tested the MSTD-RL model against the real neural data, as seen in Figs.2-3. An RL agent's goal is to maximize its cumulative temporally discounted reward from the environment. The agent uses sensory information of states to complete this reward optimization task. During our observational task, the agent (NHP's brain) can use the state information, such as color cues, to build associations between states and value as the state value function, even when there is no agency, such as in the observational task. Dopaminergic centers of the brain have been shown to represent reward probabilities, the value of reward-predicting stimuli, and errors in reward expectation [1], [16], [21]. Recent work has reported a ramping up of dopaminergic activity as an animal approaches its goal [22]. All such modulations observed in the dopaminergic centers have been modeled and predicted using basic and modified TD-RL models [23]–[26]. In many RL learning scenarios, rewards are delayed with respect to the actions that caused them or the states that predict them. This leads to the credit assignment problem, which describes how an agent knows what actions and states to assign the credit for later rewards. Under the basic TD model, the stimulus, such as the conditioned stimulus, cue/CS, in our tasks, is represented as a complete serial compound [26], suggesting that the agent is aware of the exact amount of time that has elapsed since the CS. Such an assumption led to an incomplete encapsulation of dopaminergic neurodynamics [23]. The premise of a perfect clock was replaced with a coarse temporal stimulus representation using a temporal basis set representation in the MSTD-RL[23], [27], which we use here. We tested MSTD simulations, which mimicked the experiments of NHP A. For a comprehensive explanation of the MSTD model, please see [19], [23]

The temporal basis functions are defined as Gaussians, $f(y, \mu, \sigma) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$, where y is the trace height, μ is the center, and σ is the standard deviation of a particular basis function. Each stimulus (CS for R (CSR) and NR (CSNR), and US for R (USR) and NR (USNR) has its own memory trace and associated microstimuli. The trace height y was set to 1 at the onset of the corresponding stimulus and decayed at a rate of 0.985 on each time step. The level of the i^{th} microstimulus for a j^{th} stimulus at time t is given by $x_t^{S_j}(i) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-i/k)^2}{2\sigma^2}\right) y_t^{S_j}$, where $j = 1,2,\dots,m$, and $m \rightarrow$ number of stimuli $\in \{CSR, CSNR, USR, USNR\}$. $i = 1,\dots,k$, where; $i \rightarrow$ four microstimuli per stimulus, $k = 4$, and $S_j =$ Stimulus j. For more information on the MSTD model we used, please see [19].

MSTD simulations: Simulated ReCRT trial parameters were tailored to match the timelines of each trial experienced by NHP A while performing multiple blocks of the observational ReCRT. The first three observation ReCRT blocks performed by the NHP were considered for this analysis. Three simulated blocks were designed to match these three observations ReCRT blocks. The simulated and actual experimental blocks performed by the NHP had the same number of rewarding and non-rewarding trials. The length of each trial, the time of the cue presentation, and reward delivery in each trial were extracted for all observation ReCRT blocks. The trial length of the simulated trials was set as the trial length experienced by the NHP in each block. There were 4 stimuli and 4 microstimuli for each stimulus in the MSTD-RL simulations. Each microstimulus had a standard deviation (σ) that changed from one block to the next (block 1 = 0.1, block 2 = 0.2, block 3 = 0.3). The decay rate for the memory trace was maintained at 0.985, with the discount factor (γ) set at 0.98. The decay rate of the eligibility trace (λ) was set at 0.7 with a learning rate (α) of 0.7. The simulations mimicking each of the three experimental blocks were run separately. The immediate reward was 1 and -0.1 at the reward delivery time point in the R and NR trials, respectively. Every other time point was awarded an immediate reward of -0.001.

III. Results

We first show results from a single M1 unit's activity while the NHP performed the observational reward-cued reaching task (ReCRT) in Fig.2.a-b. These Peri-Event-Time-Histograms (PSTHs) are triggered by the cue presentation indicating rewarding (R) or non-rewarding (NR) trial types. The solid red line is for rewarding trials with the standard error of the mean shaded, while blue is for NR trials. The red dotted line is the time of reward delivery for R trials and the comparable time for NR trials. Notice as time within blocks and between blocks the region with significantly different (Wilcoxon rank sum test, $p < 0.05$) median firing rates between R and NR at a given time point (i.e. bin number, bin = 50ms), indicated by the black dots, moves further towards the cue presentation time from its initial post-reward time, Fig.2.a.1-b.6. This progression of reward level representation towards the CS is expected for an RL agent's representation of the state value function as we show in Fig.3. In Fig.2.b.1 we have plotted the percent of significantly different

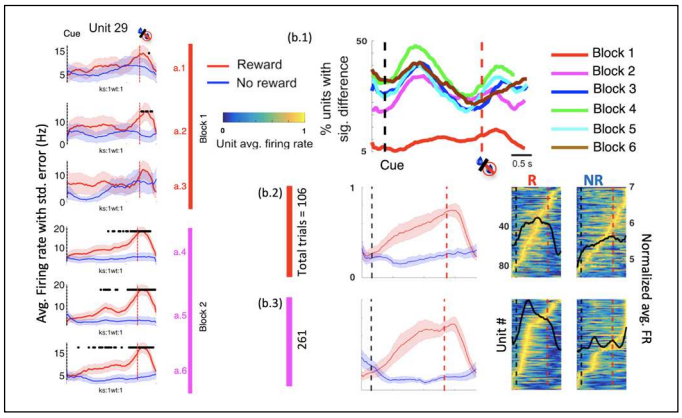

Figure.2. Single unit PSTH examples over the first (a.1 – a.3) and the second block of trials (a.4-a.6). Black dots indicate significant differences between rewarding (red) and non-rewarding (blue) trials. b.1 shows the number of units with significant differences between R and NR trials over the given blocks 1-6.

The average population PSTHs and the population of PSTHs in pseudocolor are shown in b.2 and b.3 for blocks one and two.

units between R and NR trials. Note that in block one, only a small percentage of units show such significance (Wilcoxon rank sum test, $p < 0.05$), but in subsequent blocks, the percentage is raised with phasic increases in the immediate post-cue and post-reward periods. Fig.2.b.2-3 show results from the population average PSTH between R and NR trials as line plots for a subpopulation that was from the top 10% of units with respect to their correlation with the reward sequence (see Methods). The population of PSTHs for all individual units is shown in pseudocolor separately for R and NR trials. The black lines are the average of all units' activity. Units were sorted for each block in decreasing order with respect to the time required for a given unit to reach the maximum average firing rate across R or NR trials (Fig.2.b.2-3). Therefore, units at the top of the PSTHs 'neurograms' reach the maximum average firing rate later in the trial, whereas the units at the bottom of the neurogram reach the maximum average firing rate earlier in the trial. There was no requirement for maintaining the sorting order across blocks.

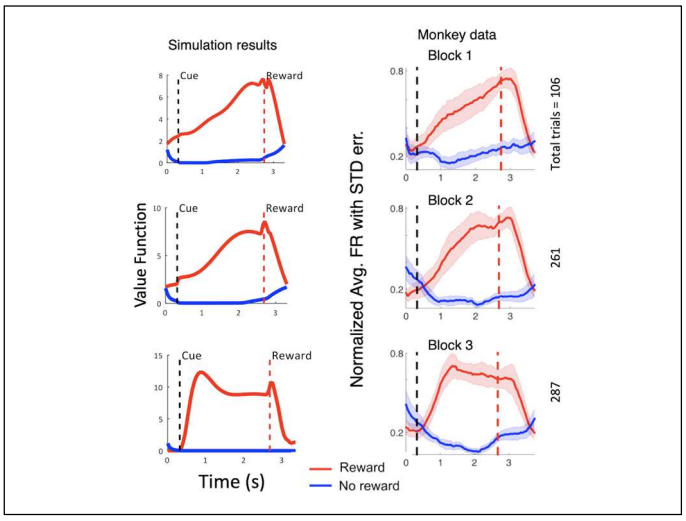

Figure. 3. Results from MSTD-RL simulations' value functions are in the left column for blocks 1-3, while the actual neural activity from the top 10% of reward correlated units are shown in the right column. Note the similarities between the simulated value functions and the neural data.

IV. Discussion

In work presented here we have shown what appears to be a state-value function in neural activity from the primary motor cortex of an NHP, observing relationships between environmental states and actions from an external agent that predict differing values to the subject. We have shown that this state-value activity changes with experience over blocks of trials, such that the reward-modulated neural activity first shows significant separability between the R and NR trials in the post-reward period Fig.2.a, and with experience, this significant separability extends into the post-cue period. Looking at the percentage of units with such significant separability at a bin-by-bin level, we showed that with learning, the peak in separability moves from the post-reward period to the post-cue period Fig.2.b.1 block one as compared to the remaining blocks.

In addition to showing significant changes in the reward-modulated activity with experience, that is, learning of a putative state-value function, we have modeled this neural behavior with the MSTD-RL model. The model shows that such a state-value function is consistent with TD-RL. There are clear limitations to work presented here. One is the lack of changes in behavioral variables, such as shorter reaction times or times in trials between the R and NR trials with state-value learning. We have seen such associated changes in trial time for the manual version of the ReCRT (data not shown); however, such data is unavailable for the observational task shown here. The data used in this report was taken for brain-machine-interface work and was not optimized to answer questions on RL. We are currently analyzing data from choice-based experiments that will more clearly address the behavioral aspects of RL in the primary motor cortex.

Our future work will utilize this M1 state-value activity to autonomously update a brain-machine interface (BMI), such that the BMI works towards maximizing the user's cumulative subjective reward.

References

[1] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997, doi: 10.1126/science.275.5306.1593.

[2] B. T. Marsh, V. S. A. Tarigoppula, C. Chen, and J. T. Francis, "Toward an autonomous brain machine interface: Integrating sensorimotor reward modulation and reinforcement learning," *Journal of Neuroscience*, vol. 35, no. 19, pp. 7374–7387, 2015, doi: 10.1523/JNEUROSCI.1802-14.2015.

[3] A. Ramakrishnan, Y. W. Byun, K. Rand, C. E. Pedersen, M. A. Lebedev, and M. A. L. Nicolelis, "Cortical neurons multiplex reward-related signals along with sensory and motor information," *Proceedings of the National Academy of Sciences*, vol. 114, no. 24, pp. E4841–E4850, 2017, doi: 10.1073/pnas.1703668114.

[4] P. Ramkumar, B. Dekleva, S. Cooler, L. Miller, and K. Kording, "Premotor and motor cortices encode reward," *PLoS ONE*, vol. 11, no. 8, p. e0160851, 2016, doi: 10.1371/journal.pone.0160851.

[5] D. B. McNiel, J. S. Choi, J. P. Hessburg, and J. T. Francis, "Reward value is encoded in primary somatosensory cortex and can be decoded from neural activity during performance of a psychophysical task," in *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 3064–3067. doi: 10.1109/EMBC.2016.7591376.

[6] J. An, T. Yadav, J. P. Hessburg, and J. T. Francis, "Reward expectation modulates local field potentials, spiking activity and spike-field coherence in the primary motor cortex," *eNeuro*, vol. 6, no. 3, p. ENEURO.0178-19.2019, 2019, doi: 10.1523/ENEURO.0178-19.2019.

[7] J. Olds and P. Milner, "Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain.," *Journal of Comparative and Physiological Psychology*, vol. 47, no. 6, pp. 419–427, 1954, doi: 10.1037/h0058775.

[8] T. Ljungberg, P. Apicella, and W. Schultz, "Responses of monkey dopamine neurons during learning of behavioral reactions," *Journal of Neurophysiology*, vol. 67, no. 1, pp. 145–163, 1992, doi: 10.1152/jn.1992.67.1.145.

[9] W. Schultz, "Predictive reward signal of dopamine neurons," *Journal of Neurophysiology*, vol. 80, no. 1, pp. 1–27, 1998, doi: 10.1152/jn.1998.80.1.1.

[10] M. R. ROESCH and C. R. OLSON, "Neuronal Activity Related to Anticipated Reward in Frontal Cortex: Does It Represent Value or Reflect Motivation?," *Annals of the New York Academy of Sciences*,

vol. 1121, no. 1, pp. 431–446, Sep. 2007, doi: 10.1196/annals.1401.004.

[11] G. Schoenbaum, M. R. Roesch, T. A. Stalnaker, and Y. K. Takahashi, "A new perspective on the role of the orbitofrontal cortex in adaptive behaviour," *Nature Reviews Neuroscience*, vol. 10, no. 12, pp. 885–892, 2009, doi: 10.1038/nrn2753.

[12] G. W. Huntley, J. H. Morrison, A. Prikhozhan, and S. C. Sealfon, "Localization of multiple dopamine receptor subtype mRNAs in human and monkey motor cortex and striatum," *Molecular Brain Research*, vol. 15, no. 3–4, pp. 181–188, 1992, doi: 10.1016/0169-328X(92)90107-M.

[13] K. Molina-Luna *et al.*, "Dopamine in motor cortex is necessary for skill learning and synaptic plasticity," *PLoS ONE*, vol. 4, no. 9, p. e7082, 2009, doi: 10.1371/journal.pone.0007082.

[14] J. A. Hosp, A. Pekanovic, M. S. Rioult-Pedotti, and A. R. Luft, "Dopaminergic projections from midbrain to primary motor cortex mediate motor skill learning," *Journal of Neuroscience*, vol. 31, no. 7, pp. 2481–2487, 2011, doi: 10.1523/JNEUROSCI.5411-10.2011.

[15] A. A. Hamid *et al.*, "Mesolimbic dopamine signals the value of work," *Nature Neuroscience*, vol. 19, no. 1, pp. 117–126, 2015, doi: 10.1038/nn.4173.

[16] Y. Niv, N. D. Daw, D. Joel, and P. Dayan, "Tonic dopamine: Opportunity costs and the control of response vigor," *Psychopharmacology*, vol. 191, pp. 507–520, 2007, doi: 10.1007/s00213-006-0502-4.

[17] M. M. U. Atique and J. T. Francis, "Mirror Neurons are Modulated by Grip Force and Reward Expectation in the Sensorimotor Cortices (S1, M1, PMd, PMv)," *Scientific Reports*, vol. August 2021, Aug. 2021, doi: 10.1038/s41598-021-95536-z.

[18] Z. Yao, J. P. Hessburg, and J. T. Francis, "Normalization by Valence and Motivational Intensity in the Sensorimotor Cortices (PMd, rM1, and cS1)," Sep. 2021. doi: 10.1101/702050.

[19] V. S. A. Tarigoppula, J. S. Choi, J. H. Hessburg, D. B. McNiel, B. T. Marsh, and J. T. Francis, "Motor Cortex Encodes A Value Function Consistent With Reinforcement Learning," *bioRxiv*, p. 257337, Jan. 2018, doi: 10.1101/257337.

[20] P. Y. Chhatbar, L. M. von Kraus, M. Semework, and J. T. Francis, "A bio-friendly and economical technique for chronic implantation of multiple microelectrode arrays," *Journal of Neuroscience Methods*, vol. 188, no. 2, pp. 187–194, 2010, doi: 10.1016/j.jneumeth.2010.02.006.

[21] Y. Niv, N. D. Daw, and P. Dayan, "How fast to work: Response vigor, motivation and tonic dopamine," in *Advances in Neural Information Processing Systems 18*, 2005.

[22] M. W. Howe, P. L. Tierney, S. G. Sandberg, P. E. M. Phillips, and A. M. Graybiel, "Prolonged dopamine signalling in striatum signals proximity and value of distant rewards," *Nature*, vol. 500, pp. 575–579, 2013, doi: 10.1038/nature12475.

[23] E. A. Ludvig, R. S. Sutton, and E. J. Kehoe, "Stimulus representation and the timing of reward-prediction errors in models of the dopamine system," *Neural Computation*, vol. 20, no. 12, pp. 3034–3054, 2008, doi: 10.1162/neco.2008.11-07-654.

[24] H. W. Chase, P. Kumar, S. B. Eickhoff, and A. Y. Dombrovski, "Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis," *Cognitive, Affective and Behavioral Neuroscience*, vol. 15, no. 2, pp. 435–459, 2015, doi: 10.3758/s13415-015-0338-7.

[25] S. J. Gershman, "Dopamine ramps are a consequence of reward prediction errors," *Neural Computation*, vol. 26, no. 3, pp. 467–471, 2014, doi: 10.1162/NECO_a_00559.

[26] R. E. Suri, "TD models of reward predictive responses in dopamine neurons," *Neural Networks*, vol. 15, no. 4–6, pp. 523–533, 2002, doi: 10.1016/S0893-6080(02)00046-1.

[27] E. A. Ludvig, R. S. Sutton, and E. J. Kehoe, "Evaluating the TD model of classical conditioning," *Learning and Behavior*, vol. 40, pp. 305–319, 2012, doi: 10.3758/s13420-012-0082-6.