Motor Cortex Encodes A Value Function Consistent With Reinforcement Learning

Venkata S Aditya Tarigoppula[1,2], John S Choi[1], Jack H Hessburg[1], David B McNiel[1], Brandi T Marsh[1] and Joseph T Francis[1-3]

Correspondence

joey199us@gmail.com

Summary

Temporal difference reinforcement learning (TDRL) accurately models associative learning observed in animals where they learn to predict the reward value of an unconditioned stimulus (US) based on a conditioned stimulus (CS), such as in classical conditioning. A key component of TDRL is the value function, which captures the expected temporally discounted reward from a given state. The value function can also be modified by the animal's knowledge and certainty of its environment. Here we show that not only do primary motor cortex (M1) neurodynamics reflect a TD learning process, but M1 also encodes a value function in line with TDRL. M1 responds to the delivery of an unpredictable reward, and shifts its value related response earlier in a trial, becoming predictive of an expected reward, when reward is predictable, such as when a CS acts as a cue predicting the upcoming reward. This is observed in tasks performed manually or observed passively, as well as in tasks with explicit CS predicting reward, or simply with a predictable temporal task structure, that is a predictable environment. M1 also encodes the expected reward value associated with a CS in a multiple reward level CS-US task. The Microstimulus TD model, reported to accurately capture RL related dopaminergic activity, extends to account for M1 reward related neural activity in a multitude of tasks.

[1] Department of Physiology and Pharmacology, The Robert F Furchgott Center for Neural and Behavioral Science, State University of New York Downstate Medical Center, Brooklyn, NY11203, United States;

[2] Department of Biomedical Engineering, University of Houston, Houston, TX77204, United States;

[3] Lead Contact - joey199us@gmail.com

When learning to ride a bike, we learn through trial-and-error, with delays between actions and consequences, such as rewards. Trial-and-error learning, as well as forming passive stimulus-reward associations, are well modeled by reinforcement learning (RL) [1,2,3,4 5]. RL relies on reward as feedback to the learning agent, where the goal of the agent is to accumulate the maximum amount of temporally discounted reward [6]. In most scenarios, reward outcome may be subject to delays. These delays make it difficult to determine how to assign credit to actions, and or states, that have come before a reward. This is called the credit assignment problem. Temporal Difference (TD) learning methods can address this problem. Under TDRL, the agent learns the expected discounted reward (value function) for each of the states visited leading to reward. The agent can then utilize these learned estimates of the state value function at a given state to select an appropriate action to maximize reward [6,7].

Phasic neural activity in dopaminergic brain centers is similar to the TD error signal ($\delta_t = r_t + \hat{V}(S_t) - \hat{V}(S_{t-1})$), which is the difference between the current state's value estimate $\hat{V}(S_t)$, the previous state's value estimate $\hat{V}(S_{t-1})$, and the immediate reward r_t [8,9 10]. Dopamine has been shown necessary for long-term potentiation in the motor cortex associated with sensorimotor learning [11,12], possibly bridging TDRL theory with sensorimotor learning. Tonic dopaminergic activity has been shown to act like a value function, in this regard, dopamine can "charge" the nervous system, acting as a motivational signal [13,14]. Thus, dopamine could have two influences on the motor cortex, one gaiting synaptic plasticity toward sensorimotor learning, and the other "charging" neural activity, possibly priming the system for action.

Previously we, and others, have shown that reward modulates the primary sensorimotor cortices (M1,S1) [2,15–17] and frontal regions influencing M1 [18]. We have proposed that this activity could be used toward an autonomously updating brain machine interface [2,17]. Here we expand on this previous work, showing that M1 activity displays all the hallmarks of a value function as expected under TDRL. We show that M1 activity first responds to an unpredictable reward (sup fig.2), but when a conditioned stimulus (CS) is used to cue reward, M1 responds to the CS following conditioning (Fig.2). Reward prediction error is also encoded in M1[15], and seen following cue-reversal, or after an unexpected reward omission (sup fig.8). Relearning of the value function in M1 is visible following cue-reversal (sup fig.7). If reward is temporally predictable, then M1 tracks the underlying temporal structure as well. These activity patterns are seen during both manual and observational tasks bilaterally in M1 (Fig.2-3, sup fig.2-4). In addition, if multiple levels of reward are used, M1 activity encodes the expected value following CS presentation in a linear fashion, similar to that reported in the striatum [19] (Fig.3).

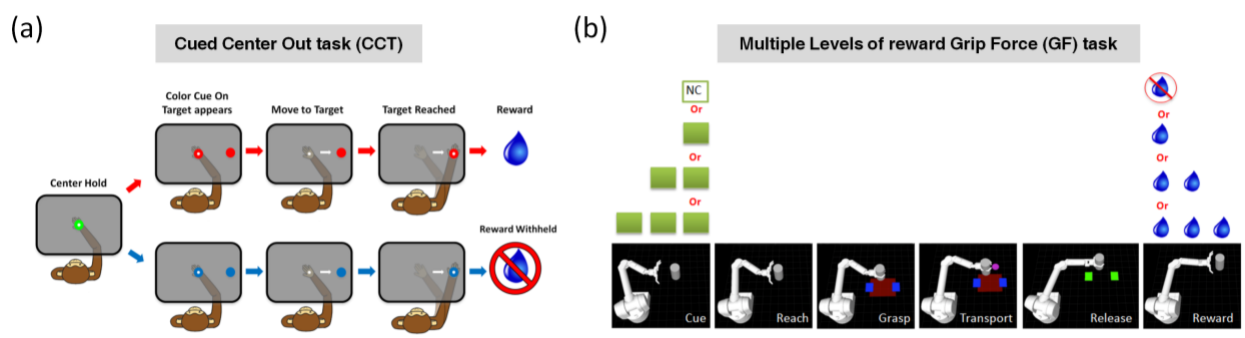

Figure 1: Behavioral tasks.

(a) Manual and observational (OT) versions of a Cued Center Out Reaching task (CCT) [2]. Manual task required a reaching movement from center to peripheral target. A cue (CS),

target color, informed the monkey of the trials value if successful. Observational CCT required the monkey to simply watch as the task played out.

(b) Cued manual and observational Grip Force (GF) task (see methods). The conditioned stimulus (CS) cued the NHP of the value. Monkeys controlled the amount of grip force necessary to pick up and move the target item.

In Fig.2, we present results from an operant conditioning task (Fig.2.c), or a classical conditioning variant of it (Fig.2.a-b). The task involved either making a cued reaching movement to a single visual target, or observing such movements (Fig.1.a, Cued Center Out task (CCT)). Trials were color cued (CS) via the reaching target as to the trials value. The sequence of trial values, rewarding (R, red) or non-rewarding (NR, blue), could be random or predictable. We found several consistencies in M1 neural responses that are expected based on TDRL. In Figure 2.a we have plotted results from the observational (OT) CCT task (Fig.1.a), which is a classical conditioning paradigm, for an example single unit from M1 that displays an activity pattern consistent with the evolution of a value function while learning a cue-reward association. We have plotted the unit's peri-cue-time-histogram (PCTH) for R and NR trials, for three sessions, broken up into three equal parts (Fig.2.a.1 - a.9). As this data comes from the OT version of the task, confounds of movement related activity are reduced. In session 1 the only time bins with significant differences between R and NR trials are post reward (Fig.2.a1 - a.3), and as experience is gained, from sessions 2-3, and putative learning of the association builds, there is a movement of significant differences propagating forward in time toward the presentation of the CS. At the end of these three sessions one can see a peak of activity post CS and before/at reward delivery (Fig.2.a.7-9), as expected for a value function from TDRL (see Fig.3.a). We have plotted the average neural activity for a subpopulation of single/multi units that correlated with reward value, separately for R and NR trials, in red and blue respectively (Fig.2.b.2-6). Specifically, this subpopulation is the average of the top 10% of the population after rank ordering with respect to individual unit's correlation with reward. In both the OT Fig.2.b, and manual, Fig.2.c, versions of the CCT task, we see that the difference between R and NR trials, grows earlier in the trials as the subject continues the task over time, as in Fig.2.a for the single unit.

We have plotted the PCTH for R and NR trials for all units as false color plots in Fig.2.b,c.2-6. In each false color subplot, the black line is the mean of the population. Learning also took place for the population mean in the OT-CCT where the trial type sequence was completely predictable with R-trials always followed by NR-trials, and repeating. Thus, this is a very stable and predictable environment. Notice that in the OT-CCT predictable sequence task (Fig.2.b) that the NR population mean activity, black lines (Fig.2.b.2-6), become more and more linearly increasing with time to the next trial, which is a rewarding trial, thus the full motor cortical population average activity is tracking the time to the next R associated CS starting from the previous trial. This population activity peaks post R-CS, as seen in the OT-CCT R trials (Fig.2.b.2-6). This population tracking of the trial sequence is also clear in Fig.2.b.1 where we have plotted the % units in the full population that show significant differences between R and NR trials. Note, that many units show separability pre-CS in Fig.2.b.1 as compared to the manual task (Fig.2.c.1), which had a random trial value sequence. For both the manual and OT-CCT task there is an increase in the number of units that show significant differences between R and NR trials from session 1 to sessions 2, and onward, indicating the animal is learning the association between the CS and the reward (US, unconditioned stimulus) regardless of the environmental stability, that is fully predictable vs. random trial value sequences (Fig.2.b.1-c.1, sup fig.4)). Both the manual and OT-CCT task data show two peaks in these % unit plots (Fig.2.b,c,.1, sup fig.4)), one post cue (CS) and one post reward delivery. Differences between

the manual and OT plots are likely due to the uncertainty that is present in the manual task vs. the OT task.

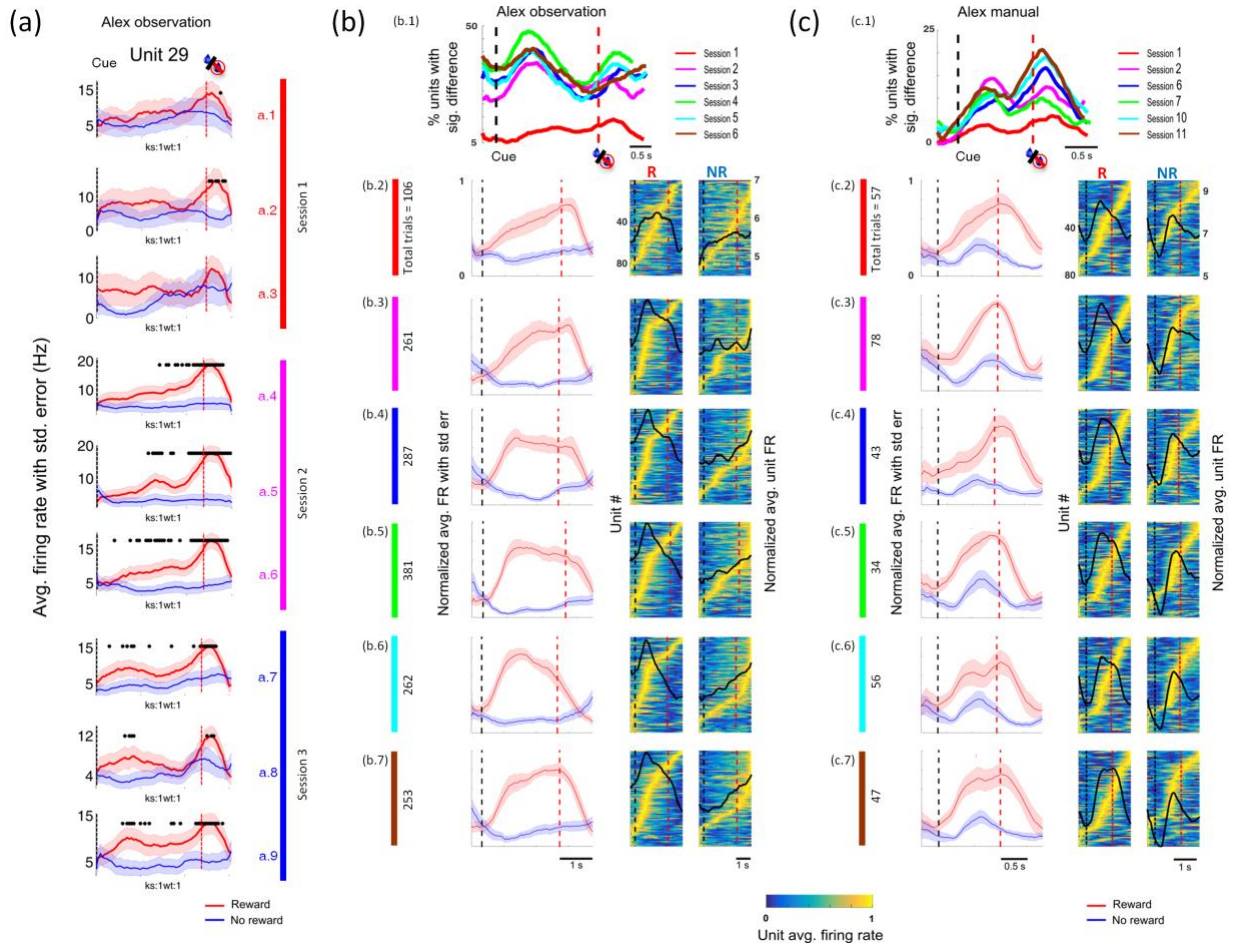

Figure 2 Single unit and population activity showing value function like evolution with learning. (a) Single unit's example peri-cue-time-histograms (PCTH) for rewarding and non-rewarding trials of the OT-CCT task. (b.1,c.1) % units with significant differences between R and NR trials for OT and Manual CCT. (b.2-7, c.2-7) sub-population PCTH, red and blue line plots, and full population PCTH in false color with black line showing the mean of the population.

The results in Fig.2 were recorded from M1 contralateral to the arm used by the NHP in the manual CCT. We hypothesized that the reward modulation signal would be broadcast to M1 in both hemispheres[2] and show support of this in sup fig.3-4. The results shown in Fig.2.b.1 indicate that the M1 population is clearly separable between R and NR trials even before the CS is shown due to the predictable sequence of the trial's value, that is R trials followed by NR and repeating. We show further support of this in sup fig.2, where we tested this hypothesis in the grip force (GF) task seen in Fig.1.b. The task had no explicit CS (un-cued), indicating the trial value, and the trial value (R, NR) sequences could either be fully predictable, or random. In sup fig.2 we present results from these GF tasks (sup fig.1) consistent with the CCT results, that again M1 showed clear separability between R and NR trials when the trial value was predictable, even in the absence of a CS that indicated the trial value.

In order to further test our hypothesis that M1 holds a value function we ran simulations of a Microstimulus TDRL model[20] using the same experimental structure as the real experiments (see Methods for model description). In Fig.3.a we have plotted the value functions from this TDRL model (Fig.3.a left column) as well as data from Fig.2.b.2-4 for comparison. There is clearly a strong resemblance between the predicted value function by the model and the neural data (see sup fig.5-7 and the corresponding sections for further supporting results). The cross-correlation between the model value function and the neural data in Fig.3.a is on average r = 0.91. In Fig.3.b we have plotted the PCTH for example single units from M1 with an average cross-correlation between these units and the value function for the Microstimulus TDRL model of r = 0.92.

Furthermore, if M1 activity is holding a value function then we should be able to test this using multiple levels of reward. Thus, we used the cued GF task with different levels of reward as seen in Fig.3.c. The monkeys experienced approximately an equal number of trials for each reward level presented randomly. Here, we have plotted the average firing rate and standard error of the mean (window size ~ 1s post cue) for M1 example units with respect to the cued reward level. We see that there are units that either represent increasing reward value levels in a linearly increasing, or decreasing fashion. Also shown are the total number of units that had significant positive and negative linear correlation with reward value. Note that the number of units that increase firing rate with value is much higher than for units that decrease Fig.3.c. These relationships are seen during both manual and fully observational versions of the GF task.

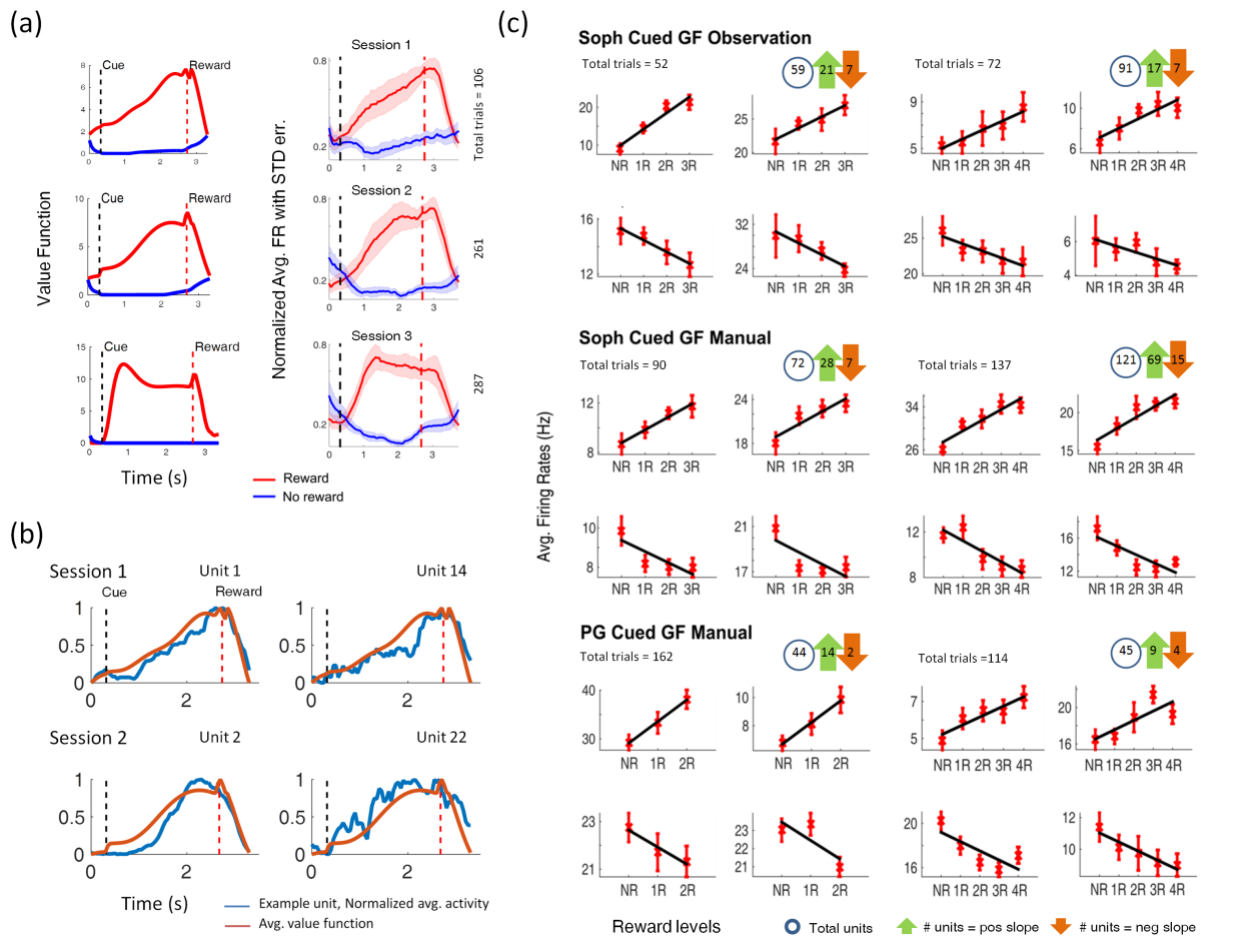

Figure 3 Value representation in M1 in time and reward level. (a) TDRL simulation value function and M1 output (see Fig.2.b.2-4). (b) Single unit examples from M1 with the TDRL value function overlaid. (c) M1 mean firing rates, with SEM, for the cued reward value on the x-axis from observational and manual tasks for monkey Soph and monkey PG.

In short M1 activity holds all the hallmarks of temporal difference reinforcement learning as seen during the learning process of a CS-US relationship during action and observation bilaterally. M1 holds information on the value of the trial when multiple reward levels are used and shows reversal learning when the CS-US association changes. M1 also tracks the stability and predictability of the reward environment without a CS in order to build a state value function. We have compared the M1 data with the MSTD model with great agreement between the value function of the model and the neural data. In addition, we have shown that M1 represents reward prediction error.

Methods:

Cued-Center Out-Reaching task (CCT)

Monkeys Alex and Zee were proficient in performing an 8-target center-out reaching (COR) task before implantation. Alex and Zee were implanted in the contralateral and ipsilateral M1 (with

respect to the right arm) respectively. These monkeys were then introduced to the cued center out reaching task (CCT) (Fig.1.a) [2] . A hand-feedback cursor was displayed on a screen in the horizontal plane just above their right arm in alignment with their right hand during the manual task. The monkeys were asked to perform cued reaching tasks, where the reward level was cued via the color of the reaching target. In an observational version of the task the monkey passively observed constant speed cursor trajectories to the cued target. Progression from the current trial to the next was only allowed following successful completion of the current trial. The data considered here from the monkey Alex corresponds to the days when it was relatively new to the CCT. Alex performed manual and observational CCT with chance or complete trial value predictability (TVP) respectively. The chance TVP session had a random sequence of R and NR trials in a given session whereas, an R-NR sequence was repeated in a completely predictable manner in the complete TVP session. Monkey Zee had some experience with the CCT post implantation. The data considered in this paper from Zee are from days with an inherent bias in the number of R to NR trials presented randomly (66% R trials). These sessions were from days after Zee had a break of 20 days from performing the CCT, during which it performed BMI tasks. In addition, Zee did multiple types of tasks on these days - CCT, manual COR and BMI experiments. Monkey Bond, our pseudo naive monkey, was under an unsuccessful training to perform the COR task manually. It never performed a single successful manual reach to the target in 18 days of training spread across 2 months. We stopped the training as considered it as our naive animal given the near zero R values of the predicted kinematics from the neural data. The data considered in this work from Bond corresponds to the second day ever of it experiencing the observational CCT. It was required to maintain its gaze on the task plane throughout the trial period. The virtual cursor moved only when the gaze was maintained on the task plane.

Grip Force (GF) task

Monkeys Soph and PG were required to apply and maintain an instructed (with tolerance) amount of grip force (applied force shown as red bars, instructed force shown as blue bars Fig.1.b) following initialization of the grasping phase until the robot had automatically moved the cylinder to a target position. Monkeys were required to release the grip to successfully complete the task and receive / not receive juice based on the trial type, that is rewarding or non-rewarding. Progression from the current trial to the next required successful completion of the current trial. The difference between cued and uncued GF tasks was the presence of a conditioned stimulus (CS) indicating the trial's value. The presence or absence of a cue at the beginning of a cued GF trial informed the monkey on whether it would be rewarded or not at the end of a successful trial. There was no CS during uncued GF tasks. Both monkeys were required to perform the grip manually in the manual version of the task whereas they passively observed while the grasping was performed automatically in the observation version of the GF task. All monkeys performed the grip force task with their right hand. The GF task had the following 6 stages. 1.) Cue - This stage is observed only during the cued GF task. A cue (CS) explicitly informed the monkey of the reward value it would receive at the end of a successfully completed trial. Cues fly into the task plane during this stage. The cue was maintained in the task plane throughout the trial period. 2.) Reaching - The cue presentation is complete before the task enters into this stage. The virtual robot moved automatically at a constant speed from the rest position to the cylinder location. 3.) Grasping - Monkey is allowed to apply grip force, virtually represented in real time as a force bar in red (Fig.1.b & sup fig. 3.a). Blue bars in Fig.1.b represent the instructed force required to be applied and maintained during the task by the monkey. The width of the blue bars instructed the tolerance allowed in matching the applied grip force to the instructed force. The grip force is considered valid as long as the applied force (red bar) is within the lower and upper boundary of the blue bars. Over or under application of the force resulted in a failure and a repeat of the same trial type. The grip force is applied

automatically and the monkey is required to watch the task being performed passively during the observation GF trials. 4.) Transport - Virtual robot automatically moved the cylinder at a constant speed from the start to the target position given that the monkey maintains the instructed force i.e. the monkey is required to maintain the instructed amount of force throughout this period. 5.) Release - Monkey is required to release its grip to successfully complete the trial and either receive or not receive juice reward. The release scene is automatically executed during the observational GF trials. 6.) Reward- Juice was delivered to the monkey following a successfully completed rewarding trial whereas no juice was awarded following a non-rewarding trial.

Surgery

Electrode array implantation was performed after the monkeys manually performed the GF tasks (monkeys PG and Soph) or the COR tasks (monkeys Alex and Zee) proficiently as described in the earlier sections. The monkey Bond was implanted having never sucessfully performed a COR task. Chhatbar et. al. [21] describe the implantation procedure in further details. Electrode arrays were implanted in M1 either contralateral (monkeys Soph, PG, Alex and Bond), or ipsilateral (monkey Zee), to the right arm used by the monkeys to perform the CCT or GF tasks. In short, 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 arm/hand region of the primary motor cortex and 6) closure. Following craniotomy and durotomoy, markers such as the Central Sulcus, Arcuate Sulcus, Arcuate Spur and the Intraparietal Sulcus were used to confirm our location on the cortex. Cortical probing was performed using Neuronexus silicon neual probes. They were inserted into S1 and real time neural activity was heard on speakers while a surgery assistant touched the right/left arm and hand of the monkey. Once the hand and the arm regions were recognized in S1, chronic 96 channel platinum microelectrode array (Utah array with ICS-96 connectors, 1.5mm electrode length, Blackrock Microsystems) were implanted in the primary motor cortex mirroring the S1 hand/arm region across the Central Sulcus. All surgical procedures were performed under the guidance of the State University of New York Downstate Medical center Division of Comparative Medicine (DCM) and were approved by the Institutional Animal Care and Use Committee in compliance with NIH 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 by and under the guidance of DCM. Possible cerebral swelling was controlled by the use of mannitol and furosemide whereas dexamethasone was used to prevent inflammation during the surgery. A titanium post (Crist Instruments) was implanted on the skull with an attached platform built in house in order to house the 4 ICS-96 connectors form the Utah arrays [21]. Antibiotics (Baytril and Bicilin) and analgesics (Buprenephrine and Rimadyl) were administered in line with the DCM veterinarian staff recommendations.

Electrophysiology

Monkeys were allowed to recover from surgery for 2-4 weeks. Following the recovery period, single- and multi-unit activity and LFPs were recorded from M1 while the monkeys performed the behavioral tasks. Multichannel Acquisition Processor systems (Plexon Inc.) were used to record neural data. The unit activity was captured at a sampling frequency of 40KHz. Offline sorting of the neural data was performed using template sorting in Offline Sorter (Plexon Inc.) to identify individual single- and multi-units. Parameters of the templates used for sorting in the first file of a day were saved and used for sorting remaining data from that day. This provided us with units across multiple files on the same day that had similar sorting templates and parameters.

Microstimulus Temporal Difference model (MSTD)

The reinforcement learning problem for an agent is to maximize its cumulative temporally discounted reward from the environment. The agent uses information, such as from sensory systems, which we call states, in order to complete this reward optimization task, that is if the agent is actively choosing actions, such as in our operant conditioning tasks (manual tasks). During our observational tasks (classical conditioning) the agent can still use the state information to build state/value associations that is the state value function. Dopaminergic centers of the brain have been shown to represent reward probabilities, value of reward predicting stimuli and error in reward expectation [1,14,22] . Recent work has reported a ramping up of dopaminergic activity as an animal approached its goal i.e. reward [23] . All such modulations observed in the dopaminergic centers have been modeled and predicted well using basic and modified Temporal Difference (TD) reinforcement learning models [20,24,25] [26] .

In most trial and error learning scenarios rewards are delayed with respect to the actions that caused them, or the states that predict them. This leads to what is known as the credit assignment problem, which is how does the agent know what actions and states to assign the credit for later rewards. Under the basic TD model, the stimulus, such as the CS in our tasks, is represented as a complete serial compound [26] , suggesting that the agent is aware exactly of the amount of time that has elapsed since the CS. Such an assumption led to an incomplete encapsulation of dopaminergic neurodynamics, especially when the reward timing was varied [20] . Therefore, the assumption of a perfect clock in the basic TD model was replaced with a coarse temporal stimulus representation captured using a temporal basis set representation in the microstimulus TD model (MSTD) [20,27] . The MSTD model utilizes a coarse temporal stimulus representation to overcome the shortcomings of the basic TD model. The MSTD model suggested by Ludvig et. al. [20] was tested in our current work using simulations which mimicked various experiments performed by monkeys Alex, Zee, Bond, Soph and PG.

Under MSTD, stimulus onset results in a decaying memory trace. Such a trace when encoded with equally spaced temporal basis functions results in what were termed microstimuli. Each microstimulus's value at any given time point keeps track of how close the memory trace height of a given stimulus is to the center of the corresponding basis function. Therefore, it acts as a temporal proximity measure and as a confidence measure that the memory trace has reached a certain height, or in other words that the stimulus happened a given period of time ago. For further details please refer to [20] .

The basis functions are defined as simple Gaussian functions

$$f(y,\mu,\sigma) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) \qquad (1)$$

Where, y is the trace height, μ is the mean and σ is the standard deviation of the Gaussian. Each stimulus (CS for R and NR in cued tasks simulations, US for R and NR in cued and un-cued tasks simulations) has its own memory trace. The trace height y of the memory trace was set to 1 at the onset of a stimulus and decayed at a rate of 0.985 on each time step following the stimulus onset.

The level of the ith microstimulus at time t is given by:

$$x_t(i) = f(y_t, i/m, \sigma)\, y_t \qquad (2)$$

Where, m is the total number of microstimuli per stimulus. Therefore, the mean of each consecutive microstimulus is defined by i/m.

The coarse representation of the trace height, indirectly capturing the time since the stimulus, was then calculated using equation 2. The model attempts to learn the optimal estimate of the current state value. The state value can be thought of as a summation of discounted future

reward given the current time step in the trial. Therefore, the model is attempting to maximize its cumulative reward over 'some' time by better estimating the state values. Given the microstimuli levels, state values were estimated as an absolute value of the linear combination of weighted microstimuli levels. The vector of adaptable weights mapping the microstimuli levels to the state value estimates are represented as W and its elements are represented as $w_t(i)$. The state value at a given time step is:

$$V_t = W_t^T x_t = \left(\sum_{i=1}^{n} w_t(i)\, x_t(i)\right) \qquad (3)$$

Where, n is the total number of microstimuli across all stimuli. The lower boundary for the state value is maintained at 0 resulting in no negative state value.

The difference between the variables m and n can be better understood as follows. Let's assume an example trial where we have a cue and a reward. The cue and the reward both are considered as individual stimulus. The memory trace of each stimulus is encoded by m number of microstimuli each but it is not required for each stimulus's memory trace to be encoded by the same number of microstimuli. Assuming the simple scenario where traces corresponding to each stimuli are represented by the same number of microstimuli - m, the total number of microstimuli levels being mapped to estimate the current state value is $n = m + m$.

The TD error, which encapsulates the error in the estimated value, at every time step is calculated as follows -

$$\delta_t = r_t + \gamma V_t - V_{t-1} \qquad (4)$$

Where, r_t is the actual reward awarded during the reward outcome period of the trial, γV_t is the current discounted value and V_{t-1} is the previous value. TD error is then utilized to update the weights, which map the microstimulus values to the estimated state value V as shown below.

$$W_{t+1} = W_t + \alpha \delta_t e_t \qquad (5)$$

Where, α is the learning rate and e_t is the eligibility trace [6] updated as shown below.

$$e_{t+1} = \gamma \lambda e_t + x_t \qquad (6)$$

Eligibility traces are necessary for faster learning in a temporal credit assignment problem. It allows the propagation of the sparse rewards in the environment to the rest of the experienced state space. The decay of the credit such that the recently visited states are assigned more credit is represented by a decay factor λ. γ is the discount factor. γ encodes how fast the future rewarding events lose their value with time. The MSTD model was validated as detailed in the supplementary MSTD model validation section.

Correlation of units with 'reward'

A variable ‘reward’ was defined such that +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. Correlation coefficient (*corrcoef*, MATLAB) was computed between each unit’s firing rate and the variable ‘reward’ in a given session. The unit with the highest positive correlation coefficient was most correlated with rewarding trials whereas the unit with the highest negative correlation coefficient was most correlated with non-rewarding trials.

Normalization

Neural data acquired while monkeys performed various experiments were binned at 50ms. The average activity across R and NR trials for each unit was casually smoothed (window of 500ms) and concatenated to form a vector Y for a given unit. Subsequently, normalization was performed using equation (7), resulting in a range from 0 to 1. '$Normalized\ Y$' 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.

$$Normalized\ Y = \frac{Y - \min(Y)}{\max(Y) - \min(Y)} \qquad (7)$$

Peri-cue-time-histogram (PCTH) of all units with false colors

Binning of the neural data was performed. The casually smoothed average activity across R and NR trials for each unit were normalized individually. Units were sorted for each session in a decreasing order with respect to the time required for a given unit to reach the maximum average firing rate across rewarding and non-rewarding trials respectively. Therefore, units at the top of PCTH reach the maximum average firing rate across trials later in the trial whereas the units at the bottom of the 'neurogram' reach the maximum average firing rate earlier in the trial. Units in PCTH for R trials were sorted such that units at the top reached the maximum average firing rate across R trials later in the trial and similar logic was applied to the PCTH for NR trials. There was no requirement for maintaining the sorting order across sessions.

1. Schultz, W., Dayan, P. & Montague, R. A Neural Substrate of Prediction and Reward. *Science* **275,** 1593–1599 (1997).
2. Marsh, Tarigoppula, V., Chen & Francis. Toward an Autonomous Brain Machine Interface: Integrating Sensorimotor Reward Modulation and Reinforcement Learning. *Journal of Neuroscience* **35,** 7374–7387 (2015).
3. Sutton, RS & the ninth annual conference of the … B.-A. A temporal-difference model of classical conditioning. *… of the ninth annual conference of the …* (1987). at <http://www.incompleteideas.net/papers/sutton-barto-TD-87.pdf>
4. Ludvig, E. Reinforcement learning in animals. *Springer* 2799–2802 (2012).
5. O'Doherty, J., Dayan, P., Friston, K., Critchley, H. & Dolan, R. Temporal difference models and reward-related learning in the human brain. *Neuron* **38,** 329–37 (2003).
6. Sutton & Barto. Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks* **9,** (1998).
7. Walsh & Anderson. Learning from delayed feedback: neural responses in temporal credit assignment. (2011).
8. Hollerman, J. & Schultz, W. Dopamine neurons report an error in the temporal prediction of rewardduring learning. *Nature Neuroscience* **1,** 304–309 (1998).
9. Bayer, H. & Glimcher, P. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron* **47,** 129–141 (2005).
10. of the of Glimcher - PW. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of …* (2011). doi:10.1073/pnas.1014269108

11. Molina-Luna, Pekanovic, Röhrich & Hertler. Dopamine in motor cortex is necessary for skill learning and synaptic plasticity. (2009).
12. Hosp, Pekanovic & Rioult-Pedotti. Dopaminergic projections from midbrain to primary motor cortex mediate motor skill learning. (2011). doi:10.1523/JNEUROSCI.5411-10.2011
13. Hamid, A. *et al.* Mesolimbic dopamine signals the value of work. *Nat Neurosci* **19,** 117–126 (2015).
14. Niv, Y., Daw, N., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* **191,** (2007).
15. Ramakrishnan, A. *et al.* Cortical neurons multiplex reward-related signals along with sensory and motor information. *Proc National Acad Sci* **114,** E4841–E4850 (2017).
16. Ramkumar, Dekleva, Cooler, Miller & Kording. Premotor and motor cortices encode reward. (2016). doi:10.1371/journal.pone.0160851
17. McNiel, DB, Choi, JS & and … H.-J. Reward value is encoded in primary somatosensory cortex and can be decoded from neural activity during performance of a psychophysical task. *… in Medicine and …* (2016). at <http://ieeexplore.ieee.org/abstract/document/7591376/>
18. ROESCH, M. & OLSON, C. Neuronal Activity Related to Anticipated Reward in Frontal Cortex. *Annals of the New York Academy of Sciences* **1121,** 431–446 (2007).
19. Cromwell, H. & Schultz, W. Effects of Expectations for Different Reward Magnitudes on Neuronal Activity in Primate Striatum. *J Neurophysiol* **89,** 2823–2838 (2003).
20. Ludvig, E., Sutton, R. & Kehoe, J. Stimulus Representation and the Timing of Reward-Prediction Errors in Models of the Dopamine System. *Neural Computation* **20,** 3034–3054 (2008).
21. Chhatbar, von Kraus & Semework. A bio-friendly and economical technique for chronic implantation of multiple microelectrode arrays. (2010).
22. Niv, Daw & Dayan. How fast to work: Response vigor, motivation and tonic dopamine. (2005).
23. Howe, M., Tierney, P., Sandberg, S., Phillips, P. & Graybiel, A. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* **500,** 575–579 (2013).
24. Chase, H., Kumar, P., Eickhoff, S. & Dombrovski, A. Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. *Cognitive, Affective, & Behavioral Neuroscience* **15,** 435–459 (2015).
25. Gershman, S. Dopamine Ramps Are a Consequence of Reward Prediction Errors. *Neural Computation* **26,** 467–471 (2014).
26. Suri. TD models of reward predictive responses in dopamine neurons. (2002).
27. Ludvig, E., Sutton, R. & Kehoe, J. Evaluating the TD model of classical conditioning. *Learn Behav* **40,** 305–319 (2012).

Supplementary Materials

Here in this supplementary material we present further support that M1 activity represents, in part, a value function in line with TDRL. We present further results from the NHPs and from an RL model that line up with the data from the NHPs, and does so for many different tasks.

We first present results from un-cued experiments where the monkeys were not shown any conditioned stimuli that could be used to determine a given trial's value. However, in some cases the trial value sequence was fully predictable, and in other cases it was fully random. We show that the monkeys can track the underlying trial value sequence, and utilize such information to develop an appropriate value function. In other words, they are tracking the trial value sequence, if it can be tracked, and using that information to develop their neural value function. If there is no predictable trial value sequence in an un-cued task then we would expect the neural correlate of value to appear post reward, but, if they can predict the value of a trial then the neural correlate of this value should appear before the reward is received. This is precisely what we see and present supporting results below.

Un-cued Grip Force (GF) task with two reward levels

Monkeys performed a two-reward level (R/NR) un-cued GF task manually or passively (observation). Monkeys Soph and PG were proficient in performing the GF task before they were implanted in the contralateral M1 (contralateral to the right arm). The ability of our subjects to predict the value of a given trial was dependent on two main factors. First, whether or not there was a conditioned stimulus (CS) that predicted a trial's value, and secondly, whether or not the sequence of trial values had a predictable structure or not. The latter can be thought of as either a stable/predictable environment, or an unstable/random environment from a foraging point of view. We ran experiments that used un-cued trials, where no CS was presented to the monkey at the beginning of the trial. In addition, we utilized differing degrees of trial value structure from completely predictable sequences of R-NR (complete TVP, trial value predictability), to completely random sequences (Chance TVP). Monkeys PG and Soph performed the manual and observational GF task with chance and complete TVP (supp.fig.2). The un-cued GF tasks were designed to investigate whether the NHPs were able to track the trial reward values even without a CS based solely on the underlying predictability of the environment.

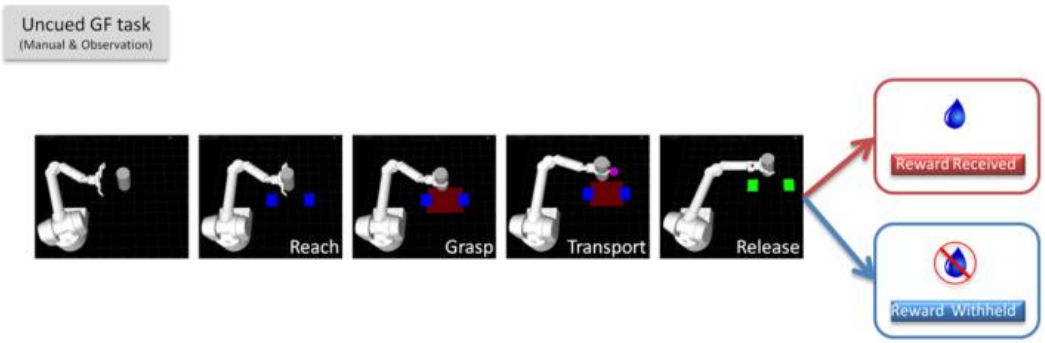

Figure 1: Behavioral task. Grip Force (GF) task. Monkeys PG and Soph performed the un-cued GF task with chance and complete TVP. No explicit cue (CS) was presented in the Un-cued GF task. Manual GF task required the monkeys to apply and maintain an instructed (with tolerance) amount of grip force (applied force shown as red bars, instructed force shown as blue bars) following the initialization of the grasping phase until the robot had automatically moved the cylinder to that target location from the start position. Monkeys were required to release the grip to successfully complete the task and receive/not receive juice based on the task paradigm. Observation GF task required the monkey to passively observe while the GF task was performed automatically for the monkey. Progression from the current trial to the next trial type required successful completion of the current trial type.

M1 can use a predictable reward environment to predict state values in the absence of an explicit conditioned stimuli

Here we give further support (see main Supp.fig.2.b.1 pre-reach separability) that the NHPs were able to track the trial values even without a CS based on the predictability of the environment. Trial value predictability of a given session was not explicitly presented to the monkeys in any task (GF or CCT). Therefore, TVP had to be inferred and tracked internally.

In Supp.fig.2, we present results from both the manual and observational versions of the un-cued GF task. Few M1 units showed a significant difference in their activity across various time windows of the trial until the post reward window in sessions with chance TVP. In contrast, a large percentage of units had significant differences in their R vs. NR activity at the beginning of the un-cued trials in with complete TVP sessions. This suggests that M1 inferred and tracked the underlying TVP in a given session and encoded the expected state value at the beginning of a trial based on the inferred trial value.

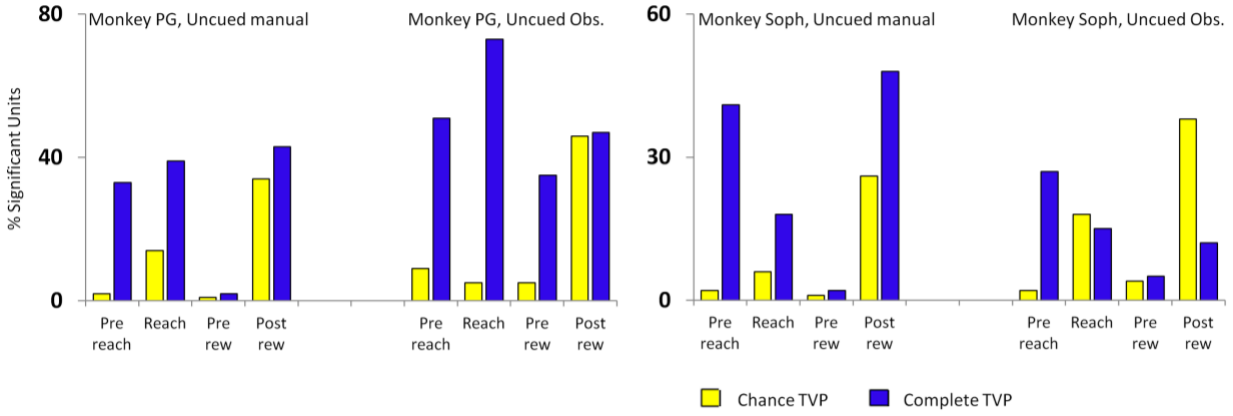

Figure 2: M1 units encode the expected reward value in un-cued GF manual and observational tasks when the reward environment is predictable. Percentage units with significantly different mean activity (ttest2, $p<0.05$) between R and NR trials for four task windows, Pre-reach (500ms before the reaching phase to the beginning of the reaching phase), Reach (500 ms from the start of the reaching phase), Pre-reward (500ms before reward to reward delivery) and Post-reward (500ms window following the completion of reward delivery/withholding). Note the large number of units with significantly different activity between R and NR at the beginning of the un-cued trials even in the absence of a reward value predicting CS in tasks with complete TVP (Blue bars).

The value function in M1 is represented bilaterally and in a naive monkey.

Monkey Zee was proficient in performing an 8-target center-out reaching (COR) task before implantation. Zee was implanted in the ipsilateral M1 (ipsilateral to the right arm). Zee also had some experience with the cued center out task (CCT) post implantation. The results shown here for Zee are from sessions with an inherent bias in the number of R trials to NR trials in a session (66% R trials). Moreover, these sessions are from days right after Zee had a break for few days. In addition, the monkey wasn't solely performing the CCT. The monkey was performing brain machine interface (BMI) tasks, manual tasks, and the CCT all on the same day. We show here that the reward related dynamics in Zee's ipsilateral M1, performing the CCT, followed a similar trend as observed in Alex (main paper, Fig. 2). For this NHP the data sessions come from 3 separate days due to the fact that this NHP was also part of BMI experiments on the same days as these CCT tasks were run. (see Supp.fig.3 below)

As a reminder, monkey Bond underwent training to perform a center out reaching (COR) task manually for 18 days spread across two months. The monkey never performed any successful manual COR task by itself during the training period. Bond was then implanted in the contralateral M1, approximately two months post the cessation of training. Bond then performed the observational CCT tasks following 4 weeks post-surgery. Therefore, we consider Bond to be a pseudo-naïve animal, and in support of this, the R value for predicted kinematics of the feedback cursor from the neural data using a linear regression model were close to 0. We investigated whether a reward signal would be observed in Bond and if Bond's M1 also represented reward related dynamics similar to that observed in monkeys Alex and Zee trained to manually perform the COR task. We show results from the second day of Bond performing the observational CCT task, which was the first day where Bond was required to maintain its gaze on the task throughout the trial, from initiation of the cursor movement post color cue till the time the virtual cursor reached the target. The cursor was allowed to move only when the monkey maintained its gaze on the task plane. This was done to make sure the visual cues and the corresponding cursor movements were in the visual field of the monkey and by extension in its attention.

In Supp.fig.3 we have plotted the average neural activity for a subpopulation of single/multi units that correlated with reward value, separately for R and NR trials, in red and blue respectively (Supp.fig.3.a-b.1-3). Specifically, this subpopulation is the average of the top 10% of the population after rank ordering with respect to individual unit's correlation with reward. In both the OT Supp.fig.3.b, and manual, Supp.fig.3.a, versions of the CCT task, we see that the difference between R and NR trials, grows earlier in the trials as the subject continues the task over time. We have plotted the PCTH for R and NR trials for all units as false color plots in Supp.fig.3.a-b,c.1-3. In each false color subplot, the black line is the mean of the population. Learning also took place for the population mean with increased trial experience. An increase in the percentage of units with significantly different activity at each bin between R and NR trials was observed earlier in the trial on day 2 and 3 compared to day 1 (Supp.fig.4). Therefore, the value function in M1 is represented bilaterally and in an untrained naive monkey.

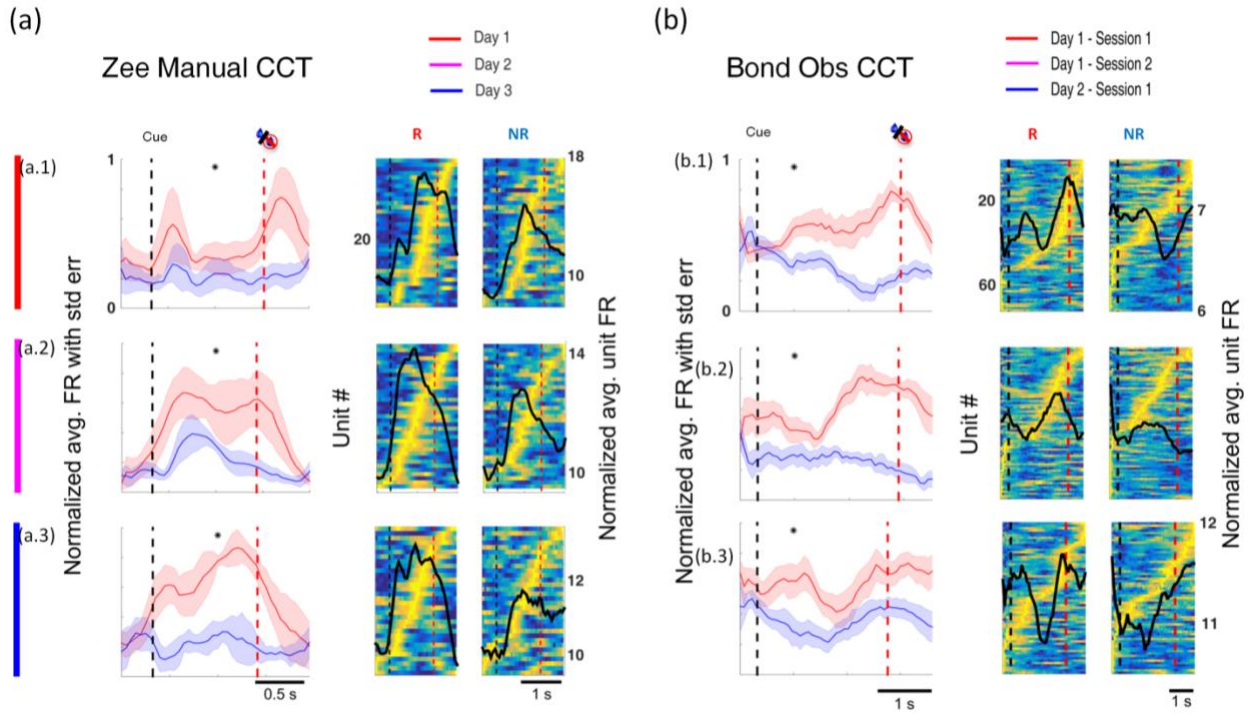

Figure 3: Population activity showing value function like evolution with learning in monkey Zee and Bond. Monkey Zee and Bond were implanted in the primary motor cortex ipsilateral and contralateral to the arm (right) respectively used by the monkey to manually perform the cued center out reaching task. (a.2-4, b.2-4) top 10 % sub-population PCTH, red and blue line plots, and full population PCTH in false color with black line showing the mean of the population.

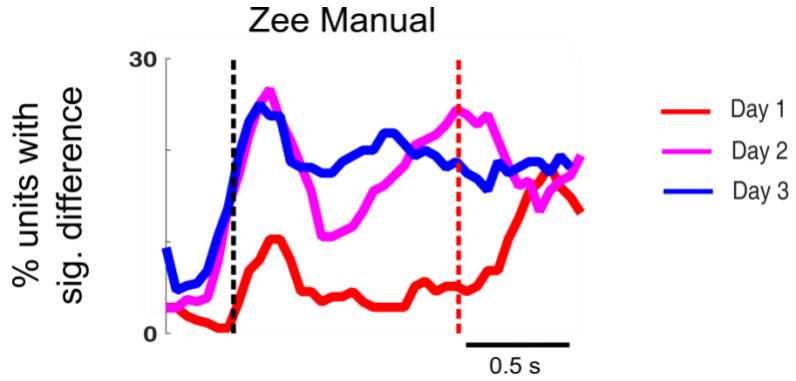

Figure 4: Percentage units that have significantly different (Wilcoxon rank sum test, $p < 0.05$) median firing rates between R and NR at a given time point. The figure shows an increase in the percentage of units whose median activity across R and NR trials is significantly different (Wilcoxon rank sum test, $p < 0.05$) post cue on day 2 and later compared to day 1. Therefore, the M1 neural ensemble represents a reward signal which is increasingly predictive of the yet to be delivered juice reward as monkey Zee progressed to later days.

Microstimulus Temporal Difference (MSTD) Value function predicts the expected reward value in an un-cued (no CS) and cued (CS) task as seen above for the NHPs.

We performed simulations of the above un-cued reward tasks with chance or complete trial type predictability (see Supp.fig.2). Please see Methods in the main paper for MSTD model details. The simulated task was arbitrarily 70 time steps long with the reward delivery/withholding at time step 55. The delivery of reward was simulated with a +1 feedback for 5 time steps to the Reinforcement Learning (RL) agent, whereas the reward withholding was simulated with a feedback of -0.1 for 5 time steps starting at time step 55. We performed simulations where the times of the reward delivery and the total trial time were varied in each trial to bridge the gap between the simulations and the actual experiments performed by the monkeys. Noise was added such that total trial time and reward time were centered at 70 and 55 respectively with a spread ranging from time - 2 to + 2 time steps. The feedback (as mentioned in the methods) is the immediate reward used to calculate the Temporal Difference (TD) error which is then used to adapt the weights mapping the current state of the RL agent to an estimated value function given its current state. The goal of the RL agent is to learn and predict the optimal value function given any time point in the task. The reward delivery during rewarding trials and the reward-withholding period during the non-rewarding trials were each considered to be an individual stimulus. The number of microstimuli per stimulus was 4 with a standard deviation (sigma) of 0.1. The decay rate for the memory trace was maintained at 0.985 with the discount factor (gamma) set at 0.95. The decay rate of the eligibility trace (lambda) was set at 0.7 with a learning rate (alpha) of 0.7. There were a total of 360 trials in the simulations.

Two simulations, one with a chance trial type predictability and another with complete trial type predictability were executed.

- Chance trial type predictability – The number of rewarding and non-rewarding trials were equal in the session. The trials were presented randomly. Supp.fig.5 (a) shows that the value function learned under the Microstimulus TD (MSTD) model only differentiates post reward delivery/withholding period. Such a prediction matches the observed neural response in monkeys Soph and PG to a similar task. The value function is incapable of predicting the expected reward given an unpredictable reward landscape in the task until the reward delivery period. The value function is less phasic in its response as compared to the TD error that has been shown to predict/model the neural dynamics of the deep brain dopaminergic neurons.

- Complete trial type predictability – The number of rewarding and non-rewarding trials were equal in a session. The trials were presented in a completely predictable sequence such that a rewarding trial always followed a non-rewarding trial and vice versa. The simulation was performed to address the question on how the RL agent would respond to a completely predictable reward landscape in a trial without a CS. Supp.fig.5 (b) shows that the value function encoded the expected reward in an un-cued task even before the delivery of reward. Such a prediction of the value function learned under the MSTD model was similar to the M1 activity dynamics reported in the un-cued GF task with complete trial predictability experiments.

The occurrence of a stimulus resulted in a trace value of 1. The time since the stimulus was coarsely coded as a combination of the decaying trace and the equally spaced bases functions covering the length of the trial resulting in micro stimuli (MS). These MS when multiplied with a weight matrix provided an estimated value of the current time step (state) of the trial. The trace decay rate of 0.985 allowed the memory of the trace to last longer since the presentation of the

corresponding stimulus i.e. a decay rate of 0.3 resulted in a faster decaying of the trace than 0.9. The weights associated with each state of the environment (in our environment, every state is visible and visited exhaustively by the model) are updated using the eligibility trace at a learning rate of alpha.

Therefore, when a non-rewarding trial stimulus gets activated at the trial's outcome time, it results in a decaying trace, and by extension the corresponding micro stimuli values. The trace with a decay rate of 0.985 has above zero values going into the next trial. Therefore, the model continued to remember the occurrence of the NR stimuli and coarsely encodes the time since the NR stimulus in the form of the MS amplitudes. The immediate reward of +1 awarded at the reward time in an R trial is propagated back in time and the weights associated with the "active" MS values are credited. Therefore, the MS values corresponding to the R stimuli are set at 1 whereas the decaying MS values corresponding to the NR stimuli still have above zero values leading to the assignment of partial credit of the obtained reward to them. Such updates over multiple trials result in an RL agent that can now start to predict the expected value before the delivery of the reward itself in un-cued tasks with complete trial type predictability. This can be thought of as the model being able to remember the time since the NR stimuli and then associate some value awarded in the R trial back in time allowing it to learn, in a predictable situation like this, that the NR stimuli is more often than not followed by a R stimuli.

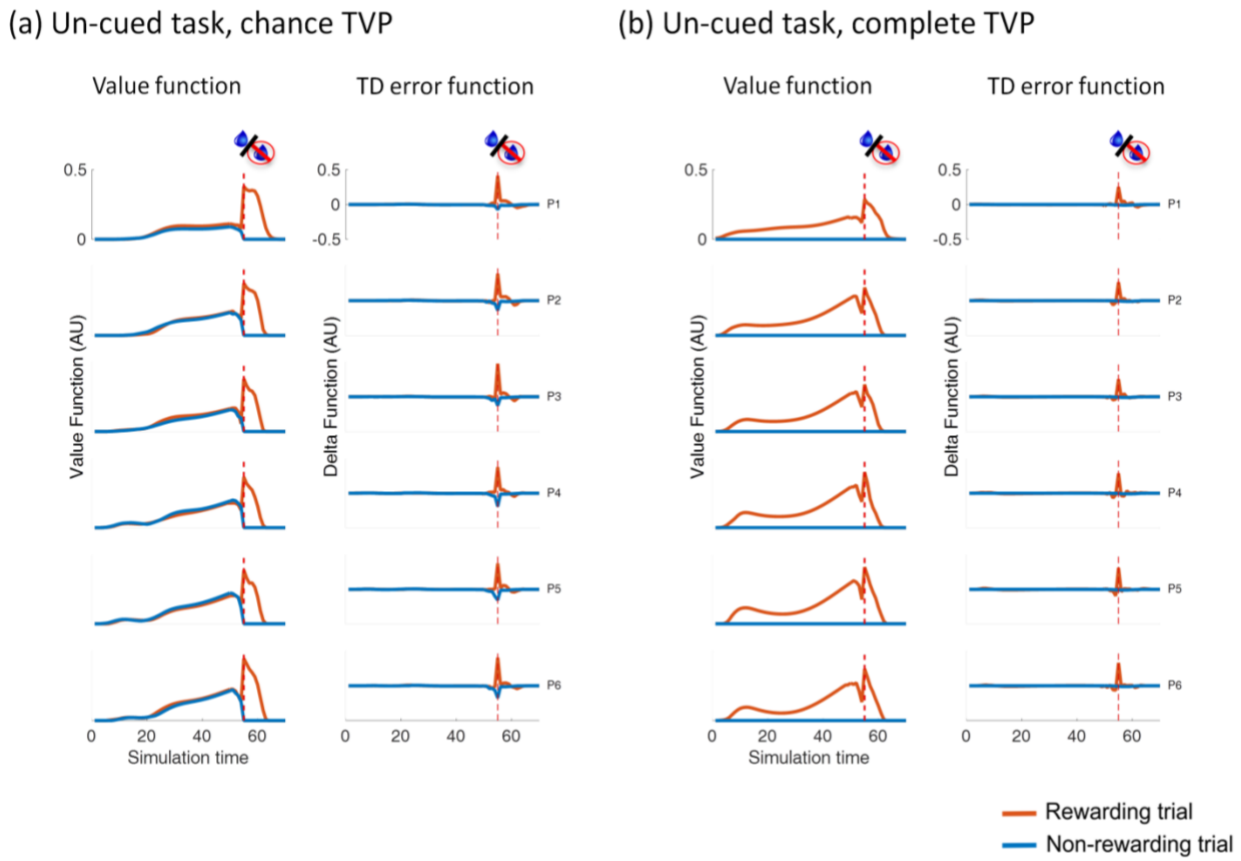

Figure 5: MSTD Value function predicts the expected reward in an un-cued (no CS) task. (a) MSTD simulation of un-cued tasks with chance trial value predictability. The value function and the TD error function are presented here for six successive sections of the total number of trials (360 trials). The red doted vertical line represents the reward delivery/withholding period. (b) MSTD simulation of un-cued tasks with complete trial type predictability. The value function and the TD error function are presented here for three successive sections of the total number of trials (360 trials). Red dotted vertical line represents the reward delivery/withholding period.

Simulations were also performed containing cued rewarding and non-rewarding trials. Rewarding, non-rewarding cues and the reward delivery/withholding periods each were considered as an individual stimulus. Therefore, there were 4 stimuli in total. Time since a stimulus was coarsely coded with four equally spaced micro stimuli. We performed these simulations with chance and complete trial value predictability. The value functions for R and NR trials were different primarily following the reward delivery in the first few trials of the session. As the session progressed, the time steps leading to the reward delivery in R trials were partially credited for the reward, thus leading to an increased expected reward (value) earlier in the trial itself irrespective of the TVP (Supp.fig.6). Supp.fig.6(b) shows that the value function in sessions with complete trial value predictability captures the expected reward value post conditioning even before the presentation of the CS compared to Supp.fig.6(a). The value function continued to respond to the actual delivery of reward along with predicting the expected reward. These predictions by the MSTD model are comparable to what were observed empirically in our experiments (Supp.fig. 3,4, main paper Fig. 2).

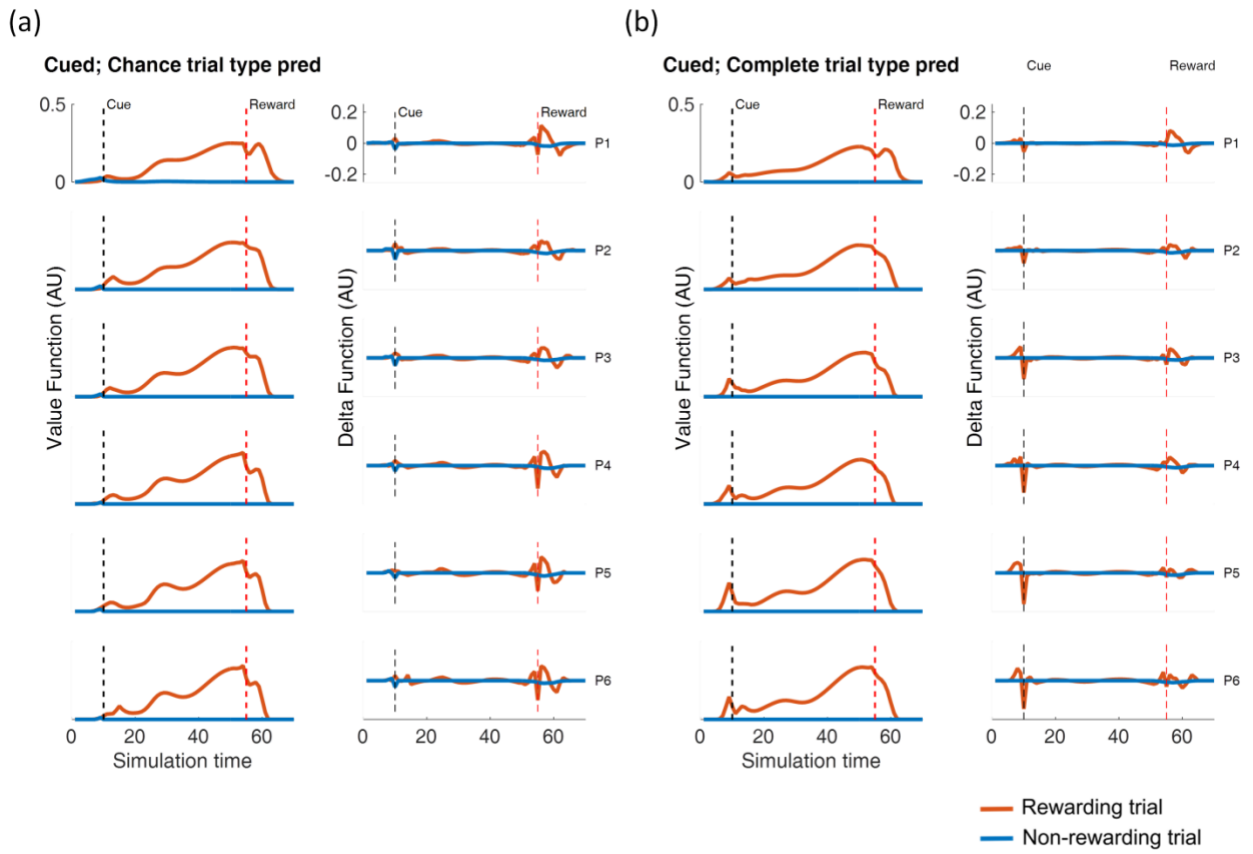

Figure 6: Predictions of the MSTD model in cued trials with varying levels of trial value predictability. (a) Cued trials with chance trial value predictability (b) Cued trials with complete trial value predictability. For both (a) and (b) - The first column displays the average value function across R and NR trials in each of the six parts of a single session. Each consecutive part (Px where x = 1:6) consisted of successive 1/6th trials of all R and NR trials. The second column captures the temporal difference error (TD error) function across R and NR trials in each of the six parts of a single session. The black and red vertical dotted lines indicated the time of cue and the average time of reward in all subplots. The time to reward delivery from the cue and the total trial time was centered at 70 and 55 respectively with a spread ranging from time - 2 to time + 2 time steps.

Relearning the value function post cue reversal is encoded in M1.

Monkeys Alex and Bond performed a cue reversal observation CCT with complete and chance TVP respectively. The monkeys performed two sessions with a color cue - reward association that they had learned. Cue-reward association reversal was initiated at the beginning of session 3. This means that the previously reward predicting cues were now no-reward predicting cues starting in session 3, and no-reward cues not predicted reward. Post cue reversal the monkeys performed another two or three sessions. These sessions were all performed in the same day. In Supp.fig.7(a-b), we show the normalized smoothed difference in the mean activity of example units between R and NR trials (normalized mean R activity - mean NR activity) in false colors. The mean R/NR activity was calculated across a moving average window of 20 trials moving by 1 trial. The example units encoded the expected reward value at the beginning of the trial before

cue reversal in session 3. The relearning of the cue-reward association resulted in the difference in R-NR activity to move towards the reward delivery/withholding period of the trial. Relearning of the new cue-reward association resulted in the activity of these example units again being able to accurately predict reward at the beginning of the trial (Supp.fig.7 a-b). We also performed simulations of the cue-reversal task with the MSTD model. Cue-reward association was reversed in the middle of the simulation session (session = 360 trials). The model parameter used were as stated in earlier sections. We show that the predictions made by the MSTD model are similar to what was observed in M1. The value function in the MSTD model also performed relearning following the cue reversal (Supp.fig.7(c) P4-6) in a manner similar to that observed in M1 units.

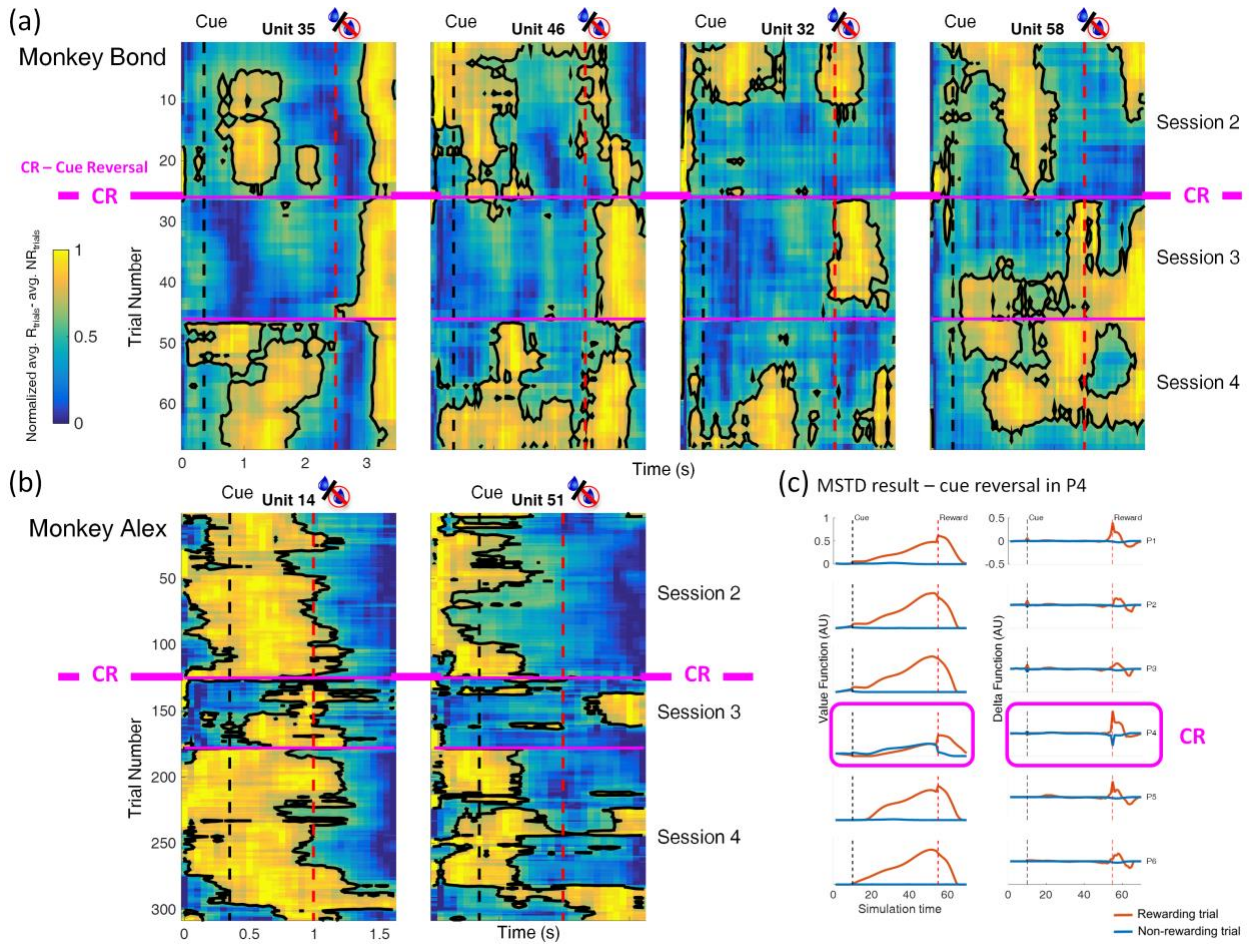

Figure 7 : Relearning of the value function following cue reversal is encoded in M1 and reproduced in an MSTD simulation. (a) and (b) Normalized smoothed difference in the mean activity of example units between R and NR trials (normalized mean R activity - mean NR activity) in false colors. The mean R/NR activity was calculated across a moving window of 20 trials moving by 1 trial. (c) Cue-reversal simulation results on the MSTD model. The first column displays the average value function across R and NR trials in each of the six parts of a single session. Each consecutive part (Px where x = 1:6) consisted of successive 1/6th trials of all R and NR trials. The second column captures the temporal difference error (TD error) function across R and NR trials in each of the six parts of a single session. CR in all subplots stands for cue reversal representing the time/window in which the cue-reward association was reversed.

M1 units signal omission of predicted reward.

In order to determine if M1 has clear reward prediction error signals, we used reward catch trails in 10% of the trials in a set of sessions. In these catch trial sessions, a conditioned stimulus was used to signal rewarding trails as in our cued tasks, however, in 10% of the cued rewarding trials, we withheld the expected reward. The catch trials were dispersed randomly throughout the session. Such catch trials should introduced a discrepancy between the expected reward and the actual reward received. This discrepancy between the expected and actual reward is called the reward prediction error (RPE) signal and is considered one of the main components of TDRL. Given that M1 neurons encode the expected reward value, one might also expect it to modulate its activity to an error in its expected reward. Ramakrishnan et. al. [1] have recently reported that the primary motor cortex encodes a reward prediction error (RPE) signal. Here we support the finding and extend it further to state that we see two unit subpopulations that modulated their activity in an opposite manner in response to a reward omission, with one population increasing activity and the other decreasing activity (see Supp.fig.8).

Monkey Soph and PG performed a cued GF task manually. The task consisted of rewarding and punishment cues such that it resulted in one of four possible outcomes; for successfully completed trials - no reward or reward, for unsuccessfully completed trial - no reward and possibly punishment in the form of a timeout period. The cue at the beginning of the trial informed the monkeys of the outcome contingency in a given trial. Ten percent of trials were catch trials, where the expected reward based on the cue were not delivered. In this analysis, we considered regular trials as those cued rewarding, completed successfully, and with reward delivered, and catch trials as those cued rewarding, completed successfully, but with no reward delivered.

M1 units were sensitive to the omission of the predicted reward (Supp.fig. 8). There were two types of units, those that significantly elevated their activity (Supp.fig. 8(a) and first column of Supp.fig. 8(b)) and those that significantly depressed their activity in response to a reward omission (second column of Supp.fig. 8(b)). The percentage of units from the total population that significantly increased or decreased their activity in response to a reward omission were 9% and 17% in monkey PG (total units = 82) and 81% and 1% in monkey Soph (total units = 79) respectively. Two subpopulations of units similar to what we have reported here in M1 were also reported in posterior cingulate cortex (CGp) by McCoy et. al. [2].

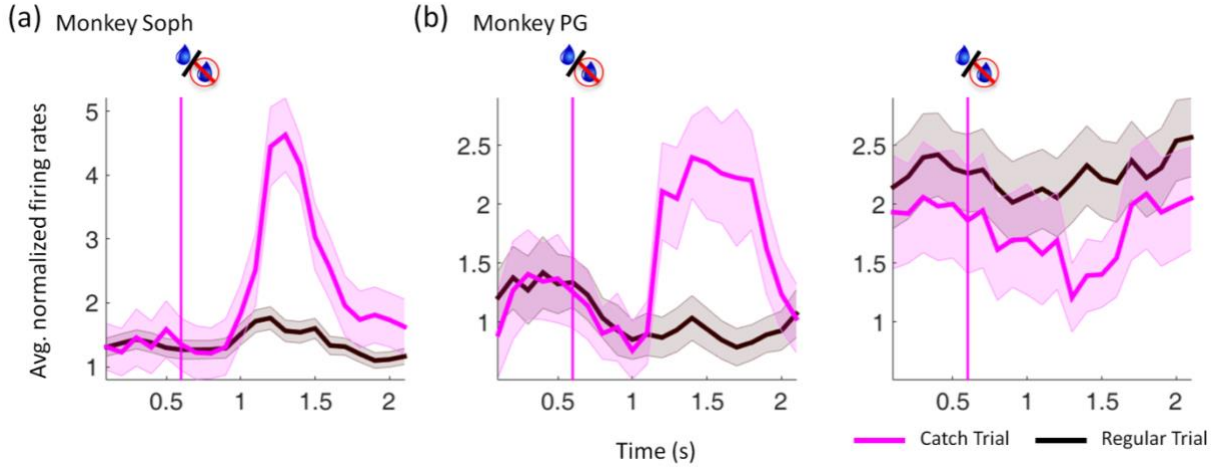

Figure 8: Two Distinct Populations of Reward Prediction Error in M1. (a) and (b) show the PSTHs (mean and standard error) of the M1 subpopulation aligned to the reward delivery/omission for monkeys Soph and PG in regular and catch trials. The pink vertical line displays the usual time of reward delivery in R trials. The average activity across M1 subpopulation between regular and catch trials were significantly different (ttest2, $p<0.05$). Binning of the data was done at 100ms. (a) Average activity of M1 units that significantly increase their firing rate in response to reward omission. (ttest2, $p<0.05$) (b) Average activity of M1 units that significantly increase (1st col) and decrease (2nd col) their firing rate in response to reward omission. (ttest2, $p<0.05$) We observed similarities in the M1 response to a reward omission as reported by us (a-b) and reported by Ramakrishnan et. al.[1] and the similarity between M1 (a-b) and posterior Cingulate Gyrus[2] in terms of containing two subpopulations of units that either increased or decreased their activity to a reward omission in comparison with a regularly rewarded trial.

References:

1. Ramakrishnan, A. *et al.* Cortical neurons multiplex reward-related signals along with sensory and motor information. *Proc National Acad Sci* **114,** E4841–E4850 (2017).
2. McCoy, A. N., Crowley, J. C., Haghighian, G., Dean, H. L. & Platt, M. L. Saccade reward signals in posterior cingulate cortex. *Neuron* **40,** 1031–40 (2003).