Modeling the Role of Striatum in Stochastic Multi Context Tasks

Sabyasachi Shivkumar[1], V. Srinivasa Chakravarthy[1] and Nicolas P. Rougier[2]

[1] Computational Neuroscience Lab, Bhupat and Jyoti Mehta School of Biosciences, Department of Biotechnology, Indian Institute of Technology Madras, Chennai, India

[2] INRIA Bordeaux Sud-Ouest, Institut des Maladies Neurodégénératives, Université de Bordeaux, Bordeaux, France

Email: Nicolas.Rougier@inria.fr

Abstract

Decision making tasks in changing environments with probabilistic reward schemes present various challenges to the agents performing the task. These agents must use the experience gained in the past trials to characterize the environment which guides their actions. We present two models to predict an agent's behavior in these tasks - a theoretical model which defines a Bayes optimal solution to the problem under realistic task conditions. The second is a computational model of the basal ganglia which presents a neural mechanism to solve the same. Both the models are shown to reproduce results in behavioral experiments and are compared to each other. This comparison allows us to characterize the theoretical model as a bound on the neural model and the neural model as a biologically plausible implementation of the theoretical model. Furthermore, we predict the performance of the agents in various stochastic regimes which could be tested in future studies.

Introduction

Uncertainty is a common problem faced by animals and humans alike in their day to day decision making. This uncertainty can be grouped into either expected or unexpected uncertainty based on the nature of the variability (Yu and Dayan). For example, a predator pouncing on the prey has a general estimate of the environmental variables like the speed of the prey, wind speed, air drag etc. This presents a known risk of failure to catch the prey. However, there are some factors like wind speed whose distributions themselves change based on other factors. In such a case, the speed of a predator when it is in the direction of the wind may not be safe in the case the wind is against it. Such parameters represent the context

which the animal must infer to adapt its behaviour. The first case falls under expected uncertainty which characterizes the variability in the different parameters of the environmental model constructed by the agent (since we use agents to model animals performing reward based tasks, we use the terms animal and agent interchangeably). Another common example of this type is seen when there is a stochastic reward while the agent is performing a reward based learning task. Standard reinforcement learning models have been used to tackle problems with expected uncertainty (Kaelbling, Littman et al. 1996, Sutton and Barto 1998). The second case of variability in the predator-prey example falls under unexpected uncertainty where we observe a consistent difference in the observations of the environment as compared to the predictions based on the agent's internal model. This could occur for example when there is a change in the environment (non-stationary environment). Specialised reinforcement learning models like modular reinforcement learning (Doya, Samejima et al. 2002) identify the context of the environment and are successful in tackling such tasks. In this work, we study reward based tasks which involve both expected and unexpected uncertainty arising due to change in the context.

Earlier experiments have studied animal behaviour in stochastic tasks (Schultz 2004). T-Maze experiments are a common paradigm for studying such tasks where the animal has to choose between one of the two arms of the maze and gets a reward upon traversing the chosen arm (Brunswik 1939, Graybiel 2005). Another interesting task to study decision making with stochastic rewards is the shape selection task where the animal has to choose amongst several shapes (each associated with a probability of reward) displayed on a screen (Pasquereau, Nadjar et al. 2007). Experiments involving non-stationary environments often have a cue indicating change in environment. However, some tasks like the serial reversal task have a reward distribution that varies with the environmental context. In these tasks, the animal has to figure out a change in context by a trial and error method (Brunswik 1939). While stochastic and non-stationary tasks have been well studied separately, tasks involving both stochasticity and changing environments are relatively unexplored and are the focus of this work.

A lot of results tend to identify Basal Ganglia (BG) as a key player in reward based learning tasks and model it as a Reinforcement Learning (RL) engine (Joel, Niv et al. 2002, Chakravarthy, Joseph et al. 2010). Furthermore, striatum, which is a major component of the BG, has a rich microcircuitry consisting of central structures called striosomes, and

matrisomes surrounding the striosomes (Graybiel, Flaherty et al. 1991). The striatum is believed to form representations of state and action space used for performing RL tasks (Charpier and Deniau 1997). Specifically, the striosomes and matrisomes are believed to map the state space (Wilson, Takahashi et al. 2014), and the action space (Flaherty and Graybiel 1994) respectively, based on their differential cortical projections. In addition, the striatum has reciprocal projections to both the Ventral Tegmental Area (VTA) and the Substantia Nigra pars compacta (SNc). It receives reward prediction error information from these midbrain nuclei and uses it to map the developed representations to state (Granger 2006) and action values (Seo, Lee et al. 2012) which are used for action selection. The striatum has also been hypothesized to perform context dependent tasks by mapping different contexts to different striatal modules (Amemori, Gibb et al. 2011, Shivkumar, Muralidharan et al. 2017).

In this article, we focus on stochastic and multi-context tasks (formally defined in *Methods*) and develop both theoretical and biologically plausible models to solve them. After formalizing the task description, we derive a model performing full Bayesian inference on the same. To compare the model performance to the animals performing these tasks (Brunswik 1939, Lloyd and Leslie 2013), we introduce some realistic task constraints to develop the theoretical model which does Bayesian inference in an iterative fashion (see *Methods*). Following this, we present a biologically plausible model of the striatum which is a variant to the one in the basal ganglia model developed to solve context dependent tasks (Shivkumar, Muralidharan et al. 2017). This model uses a layered Self Organizing Map (Kohonen 1998) architecture to model the striosomes and matrisomes as Strio-SOM and Matri-SOM where a single Strio-SOM neuron projects to a neighbourhood of the surrounding Matri-SOM neurons. The Strio-SOM and the Matri-SOM activity are mapped to compute state and action values respectively and used for action selection. This striatal model is extended to a multi-module based architecture to deal with multiple context paradigms. The biological plausibility imposes on the model limitations such as finite memory which is also incorporated into the theoretical model. Thus, the theoretical model sets a bound on the expected performance for a probabilistic context dependent task. We show that the neural model is very close to this bound for low values of stochasticity in the reward distribution.

Methods

Stochastic Multi Context Task

A stochastic multi context tasks is an extension of the standard task used in a RL setting. In this section, we introduce the various task settings and parameters and define the notation used in the rest of the paper. In a standard task, the agent is in a state ***s*** and can take action ***a***. Upon taking an action ***a***, the agent goes to a state ***s'*** and is given a reward *r*. The reward *r* is obtained from the reward distribution function $R: S \times A \mapsto \mathcal{R}$ with $r = R(s, a)$ where S and A are state and action spaces of dimensions dim(***s***) and dim(***a***) respectively and $\mathcal{R}$ is the reward space which is a subset of $\mathbb{R}$ (dim(**x**) denotes the dimension of the vector **x**). The goal of the agent in such tasks is to optimize its decisions with respect to the obtained reward.

This problem becomes harder when the environment is not stationary and the reward distribution changes based on which context the environment is present in. Mathematically, this means that the reward distribution function is redefined as $R: S \times A \times C \mapsto \mathcal{R}$ and $r = R(s, a, c)$ where C is the context space of dimension dim(***c***) and ***c*** is the context in which the agent is present. The problem is harder in this case since the agent must identify the context in which it is present and then choose the action accordingly. This class of tasks are termed as multi-context tasks. The problem of identifying a change in context has been studied in the change detection theory (Hartland, Baskiotis et al. 2007). Given infinite memory, the Page-Hinkley statistics (Hinkley 1970) can be shown to give the minimum expected time before detecting a change in context for rewards given that the rewards come from the exponential family of distributions (Lorden 1971, Hartland, Gelly et al. 2006).

The rewards as defined above are not deterministic in general. The multi-context tasks defined above are a special case of the general multi-context problems which have stochastic rewards. Mathematically, R is a probability distribution over $\mathcal{R}$ and *r* is a sample drawn from this distribution. While individually having multiple contexts or stochasticity is reasonably solvable, together they make the problem highly non-trivial. This class of problems belongs to stochastic multi context problems. Such problems can be viewed as an extension of contextual bandits (Langford and Zhang 2008) where the context information is not given to the agent.

Bayesian Model Formulation

We defined the stochastic multi-context problem in the previous section. In this section, we present an algorithm to solve the problem. We consider a simpler version of the problem but the discussions can be extended to harder tasks. We consider a single state so that the reward only depends on the context and the action chosen. We look at a setting where there are two possible actions, a_1 and a_2 and two contexts c_1 and c_2. Let a_1 be the optimal action in c_1 and a_2 in c_2. Also we restrict $\mathcal{R}$ to have 2 values- $R_{success}$ and $R_{failure}$. Since there are two possible actions and contexts, we define a reward distribution matrix as follows

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$$

where r_{ij} is the probability of getting a reward $R_{success}$ while taking action a_j in context c_i. We get $R_{failure}$ with a probability (1- r_{ij}) while taking action a_j in context c_i. With the help of this, we define the reward distribution function as

$$R(c_i, a_j) = \begin{cases} R_{success} & \text{with probability } r_{ij} \\ R_{failure} & \text{with probability } 1 - r_{ij} \end{cases}$$

Having formulated the problem, we notice that solving the problem can be reduced to the estimation of the current context since we know the optimal action in each context. Assuming we choose action ***a*** and get a reward r, using Bayes Theorem, we have

$$P(c = c_1 \mid a, r) = \frac{P(a, r \mid c = c_1) P(c = c_1)}{P(a, r \mid c = c_1) P(c = c_1) + P(a, r \mid c = c_2) P(c = c_2)} \qquad \textbf{Eq. 1}$$

$$P(c = c_2 \mid a, r) = \frac{P(a, r \mid c = c_2) P(c = c_2)}{P(a, r \mid c = c_1) P(c = c_1) + P(a, r \mid c = c_2) P(c = c_2)} \qquad \textbf{Eq. 2}$$

Assuming we do not have any initial knowledge of the current context, we have $P(c = c_1) = P(c = c_2) = 0.5$. Also $P(c = c_2 \mid a, r) = 1 - P(c = c_1 \mid a, r)$. Hence, we only need to track Eq. 1 which can be reduced to,

$$P(c=c_1 \mid a,r)=\frac{P(a,r \mid c=c_1)}{P(a,r \mid c=c_1)+P(a,r \mid c=c_2)} \qquad \textbf{Eq. 3}$$

We can now extend this to multiple trials by keeping track of the history of action selection and rewards obtained. At the ith trial, let the action chosen be a^i and the reward obtained be r^i. We get at the nth trial

$$P((c^n=c_1),...,(c^1=c_1) \mid (a^n,r^n),...(a^1,r^1))=$$

$$\frac{P((a^n,r^n),...(a^1,r^1) \mid (c^n=c_1),...,(c^1=c_1))}{P((a^n,r^n),...(a^1,r^1) \mid (c^n=c_1),...,(c^1=c_1))+P((a^n,r^n),...(a^1,r^1) \mid (c^n=c_2),...,(c^1=c_2))} \qquad \textbf{Eq. 4}$$

and correspondingly for context 2 as well. Due to independence of trials, Eq. 4 can be simplified as

$$P((c^n=c_1),...,(c^1=c_1) \mid (a^n,r^n),...(a^1,r^1))=$$

$$\frac{\prod_{i=1}^{n} P(a^i,r^i \mid c^i=c_1)}{\prod_{i=1}^{n} P(a^i,r^i \mid c^i=c_1)+\prod_{i=1}^{n} P(a^i,r^i \mid c^i=c_2)} \qquad \textbf{Eq. 5}$$

Instead of keeping the full history since beginning, we can consider a sliding history for a particular window length h, making Eq. 5,

$$P((c^n=c_1),...,(c^1=c_1) \mid (a^n,r^n),...(a^1,r^1))=$$

$$\frac{\prod_{i=n-h+1}^{n} P(a^i, r^i \mid c^i = c_1)}{\prod_{i=n-h+1}^{n} P(a^i, r^i \mid c^i = c_1) + \prod_{i=n-h+1}^{n} P(a^i, r^i \mid c^i = c_2)} \quad \text{Eq. 6}$$

These terms can be read from the reward distribution function. However, the reward distribution function is not accessible to the agent and this makes this model unrealistic. We thus need to estimate these terms which gives rise to the proposed theoretical model.

Theoretical Model

The Bayesian model developed in the previous section seems to solve the problem of estimating the context in which the agent is present. However it uses $P(a^i, r^i \mid c^i = c_1)$ which is not available to the agent. Thus, the next best option is to estimate the context the agent is in and then choose the actions accordingly. We denote the context estimated by the agent using $\hat{c}$. Following the same steps as above we get the expression for the estimated context as

$$P((\hat{c}^n = \hat{c}_1), \ldots, (\hat{c}^1 = \hat{c}_1) \mid (a^n, r^n), \ldots (a^1, r^1)) =$$

$$\frac{\prod_{i=n-h+1}^{n} P(a^i, r^i \mid \hat{c}^i = \hat{c}_1)}{\prod_{i=n-h+1}^{n} P(a^i, r^i \mid \hat{c}^i = \hat{c}_1) + \prod_{i=n-h+1}^{n} P(a^i, r^i \mid \hat{c}^i = \hat{c}_2)} \quad \text{Eq. 7}$$

Now we can get values for the terms in Eq. 7 since the agent knows which context it estimated it was in when taking the action. Using the information from the preceding trials, we can estimate the probability as,

$$P(a^i, r^i \mid \hat{c}^i = \hat{c}_1) = \frac{N((a^i, r^i) \mid \hat{c} = \hat{c}_1)}{N(\hat{c} = \hat{c}_1)} \quad \text{Eq. 8}$$

where $N((a^i, r^i) \mid \hat{c} = \hat{c}_1)$ is the number of times the agent chose a^i when it estimated the context as $\hat{c}_1$ and got the reward r^i and $N(\hat{c} = \hat{c}_1)$ is the number of times the agent estimated its context as $\hat{c}_1$. This expression was derived so that agent can estimate the context it is in by looking at the term $P((\hat{c}^n = \hat{c}_1), ..., (\hat{c}^1 = \hat{c}_1) \mid (a^n, r^n), ...(a^1, r^1))$.

But to calculate this, we need terms that imply that the agent has to estimate the context and choose actions accordingly. There is thus an inherent circularity in the problem. To break this circularity, we can solve the problem iteratively. We try to estimate the reward distribution function at trial number *t* and denote this as $\hat{\mathbf{R}}^t$. In addition, we keep track of another matrix $\hat{\mathbf{N}}^t$ which has the number of times the agent chose a particular action in a particular estimated context. The two matrices are as follows

$$\hat{\mathbf{R}}^t = \begin{bmatrix} \hat{r}^t_{11} & \hat{r}^t_{12} \\ \hat{r}^t_{21} & \hat{r}^t_{22} \end{bmatrix}$$

where $\hat{r}^t_{ij}$ represents the estimated probability of getting a reward $R_{success}$ when choosing action a_j in estimated context $\hat{c}_i$ at trial *t*.

$$\hat{\mathbf{N}}^t = \begin{bmatrix} \hat{n}^t_{11} & \hat{n}^t_{12} \\ \hat{n}^t_{21} & \hat{n}^t_{22} \end{bmatrix}$$

where $\hat{n}^t_{ij}$ represents the number of times the agent chose action a_j in estimated context $\hat{c}_i$ at trial *t*. For ease of notation, we also define $L^i_k = P(a^i, r^i \mid \hat{c}^i = \hat{c}_k)$ and k varies from 1 to 2. With this Eq. 7 becomes

$$P((\hat{c}^n = \hat{c}_1), ..., (\hat{c}^1 = \hat{c}_1) \mid (a^n, r^n), ...(a^1, r^1)) = \frac{\prod_{i=n-h+1}^{n} L^i_1}{\prod_{i=n-h+1}^{n} L^i_1 + \prod_{i=n-h+1}^{n} L^i_2} \quad \text{Eq. 9}$$

Since the reward probabilities are equally likely at the beginning of the trial, we have

$$\hat{\mathbf{R}}^0 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\hat{\mathbf{N}}^0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$P(\hat{c}^0 = \hat{c}_1) = P(\hat{c}^0 = \hat{c}_2) = 0.5$$

In trial *t*, the agent estimates its current context ($\hat{c}_i$) based on its estimate in the previous trial and chooses the action (a_j) as given in Eq. 10 and Eq. 11 respectively.

$$i = \arg\max_{k \in \{1,2\}} P(\hat{c}^{t-1} = \hat{c}_k) \qquad \textbf{Eq. 10}$$

$$j = \begin{cases} \arg\max_{k \in \{1,2\}} \hat{r}_{ik} & \text{with probability 1-}\epsilon \\ 1 + b & \text{with probability } \epsilon \end{cases} \qquad \textbf{Eq. 11}$$

where ϵ denoted the probability of exploration and $b \sim Ber(0.5)$, where *Ber*(p) denotes a number 0 or 1 drawn with a probability (1-p) and p respectively. The exploration ensures that all the actions are sampled in the initial trials.

Based on the choice of $\hat{c}_i$ and a_j, the agent can update the values of $\hat{\mathbf{R}}^t$ and $\hat{\mathbf{N}}^t$ as given in Eq. 12 and Eq. 13 respectively.

$$\hat{n}_{ij}^t = \hat{n}_{ij}^{t-1} + 1 \qquad \textbf{Eq. 12}$$

$$\hat{r}_{ij}^t = \frac{(\hat{n}_{ij}^{t-1} * \hat{r}_{ij}^{t-1} + r^t)}{\hat{n}_{ij}^t} \qquad \textbf{Eq. 13}$$

where r^t denotes the reward obtained at trial t.

Since $\hat{r}_{ij}^t$ represents the estimated probability of getting a reward $R_{success}$ when choosing action a_j in estimated context $\hat{c}_i$ at trial *t*, 1-$\hat{r}_{ij}^t$ represents the estimated probability of getting a reward $R_{failure}$. Thus L_i^t is given in

$$L_i^t = \begin{cases} \hat{r}_{ij}^t & r^t = R_{success} \\ 1 - \hat{r}_{ij}^t & r^t = R_{failure} \end{cases} \qquad \textbf{Eq. 14}$$

Substituting values of Eq. 14 in Eq. 9, we can get the estimates of the context in trial *t* as given in,

Eq. 15

$$P(\hat{c}^t = \hat{c}_1) = \frac{\prod_{f=t-h+1}^{t} L_1^f}{\prod_{f=t-h+1}^{t} L_1^f + \prod_{f=t-h+1}^{t} L_2^f}$$

Eq. 10 to Eq. 15 can be used to formulate an algorithm for the agent to solve a stochastic multi context task as shown in Fig. 1

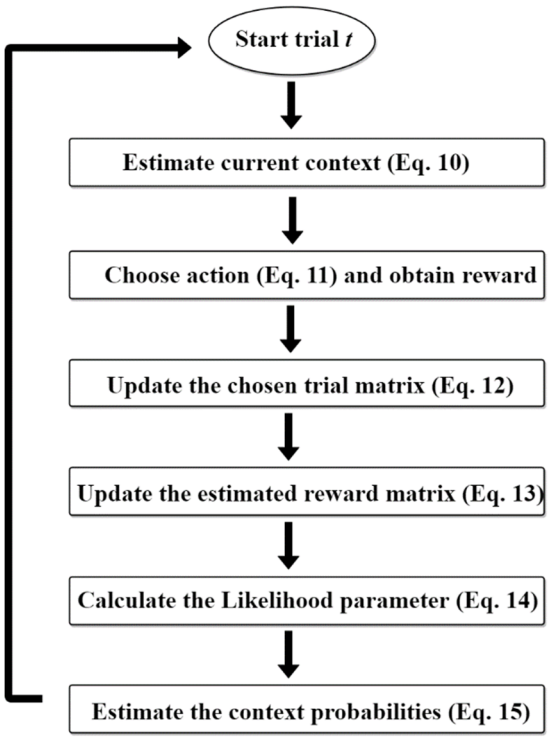

Fig. 1 Flowchart depicting steps to solve a stochastic multi context task.

Stochastic Reward Based Task Learning in Striatum

We proposed a theoretical model in the last section to solve stochastic multi context tasks. In this section we develop a biologically plausible model of the striatum for these tasks. This model is derived from an existing model of the basal ganglia proposed to solve multi-context problems (Shivkumar, Muralidharan et al. 2017). The center-surround structures seen in the striatum (Fig. 2A) are modeled using a layered SOM model. In a layered SOM model, each neuron in the center SOM layer projects to a secondary SOM layer.

The center layer in the striatal model is the Strio-SOM, which maps the state space and is believed to model the striosomes. The neurons in the Strio-SOM project to the Matri-SOM which maps the action space and is believed to model the matrisomes Fig. 2B.

Given m_1 x n_1 neurons in the Strio-SOM and m_2 x n_2 neurons in the Matri-SOM, the weights of the Strio-SOM(W^S) have dimension m_1 x n_1 x dim(**s**) where ***s*** is the state vector. Similarly, for an action vector ***a*** the weights of all the Matri-SOMs (W^M) are of dimension m_1 x n_1 x m_2 x n_2 x dim(***a***) as each neuron in the Strio-SOM projects to a Matri-SOM.

For a state input ***s***, the activity for a neuron *n* in the Strio-SOM is given in Eq. 16.

$$X^S_{[n]} = \exp\left(\frac{-\|W^S_{[n]} - s\|_2^2}{\sigma_S^2}\right) \qquad \textbf{Eq. 16}$$

where [n] represents the spatial location of the neuron n and σ_S controls the spread of the neuron activity. The complete activity of the Strio-SOM (X^S) is the combination of individual activity of all the neurons. The neuron with the highest activity ("winner") for a state **s** is denoted by n_s^*.

Similarly, for an action input ***a*** corresponding to a state ***s***, the activity for a neuron *n* in the Matri-SOM is given in Eq. 17.

$$X^M_{[n_s^*][n]} = \exp\left(\frac{-\|W^M_{[n_s^*][n]} - a\|_2^2}{\sigma_M^2}\right) \qquad \textbf{Eq. 17}$$

where σ_M controls the spread of the neuron activity. The complete activity of the Matri-SOM corresponding to neuron n_s^* ($X^M_{[n_s^*]}$) is the combination of individual activities of all the neurons in the Matri-SOM corresponding to n_s^*. The neuron with the highest activity ("winner") for an action **a** in a state **s** is denoted as $n_{s,a}^*$.

The weight of a neuron *n* in the Strio-SOM for a state input **s** is updated according to Eq. 18

$$W^S_{[n]} \leftarrow W^S_{[n]} + \eta_S.\exp\left(\frac{-\|[n]-[n_s^*]\|_2^2}{\sigma_S^2}\right).(s - W^S_{[n]}) \qquad \textbf{Eq. 18}$$

The weight of neuron n in the Matri-SOM for an action input **a** in a state **s** is updated according to Eq. 19.

$$W^{M}_{[n_s^*][n]} \leftarrow W^{M}_{[n_s^*][n]} + \eta_M . \exp\left(\frac{-\|[n]-[n_{s,a}^*]\|_2^2}{\sigma_M^2}\right).(a - W^{M}_{[n_s^*][n]}) \qquad \text{Eq. 19}$$

These representations can be used to evaluate the states and actions and guide the decision making process. The schematic of our striatal model to solve stochastic RL tasks is given in Fig. 2C.

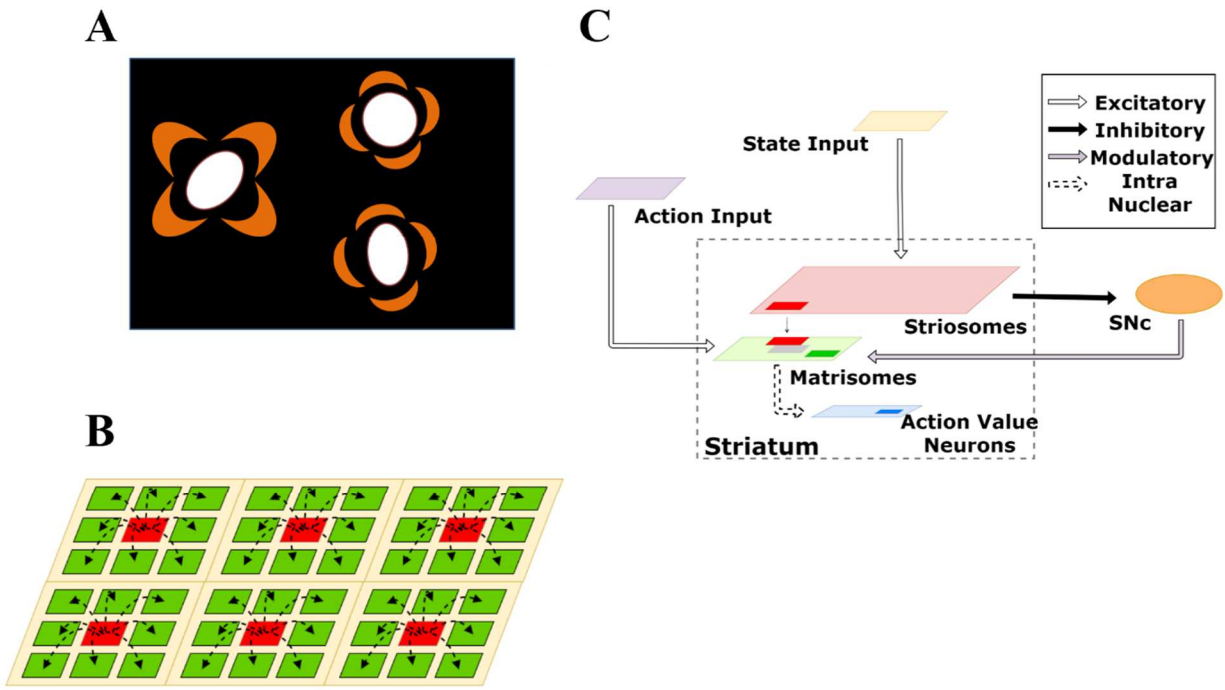

Fig. 2 **A)** Schematic of the centre-surround mapping of striosomes and matrisomes in the striatum. The white centre represents the striosomes and the surround orange represents the matrisomes. **B)** Schematic of the layered SOM architecture where each neuron in the Strio-SOM (Red) projects to the neurons in the Matri-SOM (Green) **C)** Schematic diagram of the Striatum model where the arrows indicate the connections and their types.

Let the agent performing the task be in state ***s***. The striosome activity gives us the representation of the state in the striatum. This activity is modeled by the Strio-SOM as given in Eq. 16. Thus the activity is of dimension m_1 x n_1.

This activity of the Strio-SOM projects to the SNc and represents the value for the state **s** in our model (Eq. 20). The Striatal-SNc ($W^{Str \rightarrow SNc}$) are trained using the signal from SNc which is representative of Temporal Difference (TD) error (δ) (Eq. 21). The TD error is calculated as $\delta = r + \gamma V(s') - V(s)$ where ***s'*** is the new state after taking action **a**, r is the reward obtained and γ is the discount factor.

$$V(s) = \sum_{\forall n} W^{Str \rightarrow SNc}{}_{[n]} X^{S}{}_{[n]} \qquad \textbf{Eq. 20}$$

$$\Delta W^{Str \rightarrow SNc}{}_{[n]} = \eta^{Str \rightarrow SNc} \delta X^{S}{}_{[n]} \qquad \textbf{Eq. 21}$$

where V(s) represents the value for state **s**, $\eta^{Str \rightarrow SNc}$ is the learning rate for $W^{Str \rightarrow SNc}$.

The actions that can be performed in a state ***s*** are represented by the matrisome activity surrounding the striosome neuron for that state. This is given by the activity of the Matri-SOM corresponding to the neuron with the highest activity in the Strio-SOM (n_s^*) in our model. The activity of a Matri-SOM neuron for an action ***a*** is given in Eq. 17 and is of dimension m_2 x n_2.

The Matri-SOM activity x for action ***a*** is projected to the action value neurons as given in Eq. 22. If n_a is the action value neuron for the action ***a***, $X^{Q}{}_{[n_a]}$ corresponds to the action value for the action in the state ***s*** in our model. These connections are also trained using TD error as above and the update equation is given in Eq. 23

$$X^{Q}{}_{[n_a]} = \sum_{\forall n} W^{Str(X_m) \rightarrow Str(Q)}{}_{[n_s^*][n]} X^{M}{}_{[n_s^*][n]} \qquad \textbf{Eq. 22}$$

$$\Delta W^{Str(X_m) \rightarrow Str(Q)}{}_{[n_s^*][n]} = \eta^{Str(X_m) \rightarrow Str(Q)} \delta X^{M}{}_{[n_s^*][n]} \qquad \textbf{Eq. 23}$$

where X^Q represents the activity of the action value neurons, $\eta^{Str(X_m) \rightarrow Str(Q)}$ is the learning rate for $W^{Str(X_m) \rightarrow Str(Q)}$.

The activity of the action value neurons are used for action selection by using a softmax policy in our model (Eq. 24). We believe that this is carried out by the dynamics of the STN-GPe oscillations with the striatal action value neurons projecting to the GPe. This is further elaborated in the 'Discussion' section.

$$P(a \mid s) = \frac{\exp(\beta X^{Q}{}_{[n_a]})}{\sum_{a' \in \mathcal{A}} \exp(\beta . X^{Q}{}_{[n_{a'}]})} \quad \text{Eq. 24}$$

where β is the inverse temperature and $\mathcal{A}$ denotes the action set for the agent.

Exploiting the Striatal Modularity for solving context dependent tasks

The modular nature of the striatal anatomy has been proposed to be responsible for solving context dependent tasks using a modular RL framework (Shivkumar, Muralidharan et al. 2017). In this method, the agent allocates separate modules to separate contexts. Each of the modules has its own copy of the environment in a particular context, represented by an environment feature signal (ρ). This copy is used to generate a responsibility signal, denoted by λ, which indicates how close the current context is to the one represented by the module. Thus by identifying the module with the highest responsibility signal we can follow the policy developed in that module to solve the problem in an efficient manner. We can extend the model described above to incorporate the modular RL framework. The schematic for the extended model is given in Fig. 3.

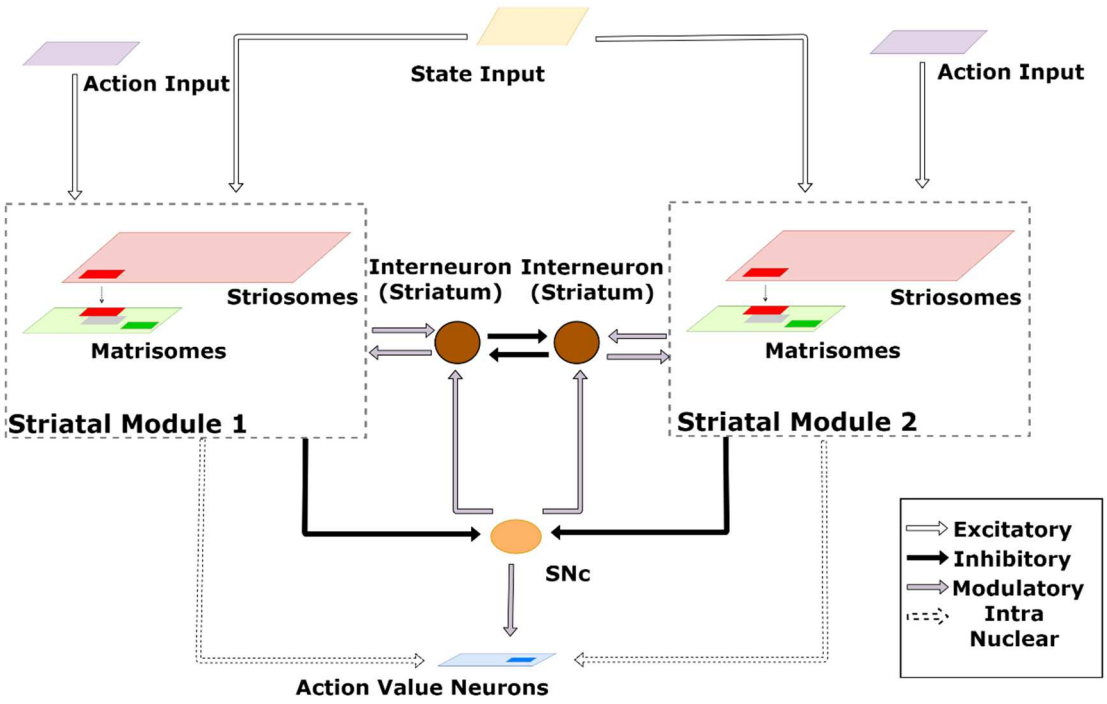

Fig. 3 A schematic of the extended model to handle modular RL tasks showing the case with two striatal modules. The state representations of the two modules are used to calculate their respective responsibilities which are then used by the striatal interneurons to choose the appropriate module.

We believe that context selection happens at the level of the striatum and the context modulated activity is projected to the action value neurons. For clarity, we have expanded the intra-nuclear activity of the striatum in the model schematic (Fig. 3). Supposing there are K modules denoted by M_1, M_2 …, M_K. We now define the weights and activities in the previous sections for each module and denote $\{M_i\}$ with each term associated with module M_i. Thus, for a module *m*, the following variables undergo a change in notation: $X^S \rightarrow X^{S,\{m\}}$ (Eq. 16), $X^M \rightarrow X^{M,\{m\}}$ (Eq. 17), $W^S \rightarrow W^{S,\{m\}}$ (Eq. 18), $W^M \rightarrow W^{M,\{m\}}$ (Eq. 19), $V(s) \rightarrow V^{\{m\}}(s)$ (Eq. 20), $W^{Str\rightarrow SNc} \rightarrow W^{Str\rightarrow SNc,\{m\}}$ (Eq. 21), $W^{Str(X_m)\rightarrow Str(Q)} \rightarrow W^{Str(X_m)\rightarrow Str(Q),\{m\}}$ (Eq. 23).

We propose that in addition to the value of the state **s**, the activity of the Strio-SOM also projects to the SNc to represent the environment feature signal ($\rho^{\{m\}}$). The weights of these projections are denoted as $W_\rho^{Str\rightarrow SNc,\{m\}}$ and are trained using the signal from SNc which is representative of context prediction error (δ^*). The corresponding equations are given in Eq. 25 and Eq. 26. The context prediction error is calculated as $\delta^* = r - \rho^{\{m\}}(s)$

$$\rho^{\{m\}}(s) = \sum_{\forall n} W_{\rho}^{Str \to SNc,\{m\}}{}_{[n]} X^{S,\{m\}}{}_{[n]} \qquad \text{Eq. 25}$$

$$\Delta W_{\rho}^{Str \to SNc,\{m\}}{}_{[n]} = \eta_{\rho}^{Str \to SNc} \delta^{*} X^{S,\{m\}}{}_{[n]} \qquad \text{Eq. 26}$$

The responsibility signal for each module is denoted by $\lambda^{\{m\}}$ for module m. In a given state **s**, the module with the highest λ is chosen for deciding the action in that state. Biologically, we believe that this selection of the appropriate module for the context is guided by the striatal interneurons (Sullivan, Chen et al. 2008). Let the winning module in the state **s** be denoted by m^{*}. The winning module projects to the action value neurons (Eq. 27) following which the processing is the same as in the previous section.

$$X^{Q}{}_{[n_a]} = \sum_{\forall n} W^{Str(X_m) \to Str(Q),\{m^*\}}{}_{[n_s^*][n]} X^{M,\{m^*\}}{}_{[n_s^*][n]} \qquad \text{Eq. 27}$$

The dynamics of the responsibility signal is given in Eq. 28

$$\dot{\lambda} = -\lambda - \alpha_{\lambda} (\delta^{*})^{2} \qquad \text{Eq. 28}$$

where α_{λ} controls the influence of context prediction error on the responsibility signal and δ^{*} is the context prediction error.

Results

Performance of theoretical model on T-Maze tasks

The study of context dependent stochastic tasks is a reasonably underexplored area owing to the complexity of decision making involved in these tasks. However, some of the earlier results (Lloyd and Leslie 2013) make some predictions which we aim to replicate with our model.

The task performed by the agent is a T-maze task (Olton 1979) where the agent has to choose one of the arms in a maze. Upon choosing the arm, the agent gets a reward R_{max} with a given probability ($P_{success}$) and a reward R_{min} with a given probability ($P_{failure}$). The task can be extended to a context-dependent problem by reversing the reward distributions with trials.

We study the performance with changing R_{max}/R_{min} and $P_{success}/P_{failure}$. Animals tend to choose rewards which have a higher magnitude and greater rewards lead to faster convergence (Fig. 4A). Similarly, with the same magnitude, animals tend to prefer distributions which reward with a higher probability (Fig. 4C). These effects are captured by our model as shown in Fig. 4B and Fig. 4D respectively. The figures show the ratio of the correct choices by the agent in 50 trials averaged over 50 sessions. The value of exploration factor, ϵ (Eq. 11) was set as 0.1 and the window length, h (Eq. 7) was chosen as 5.

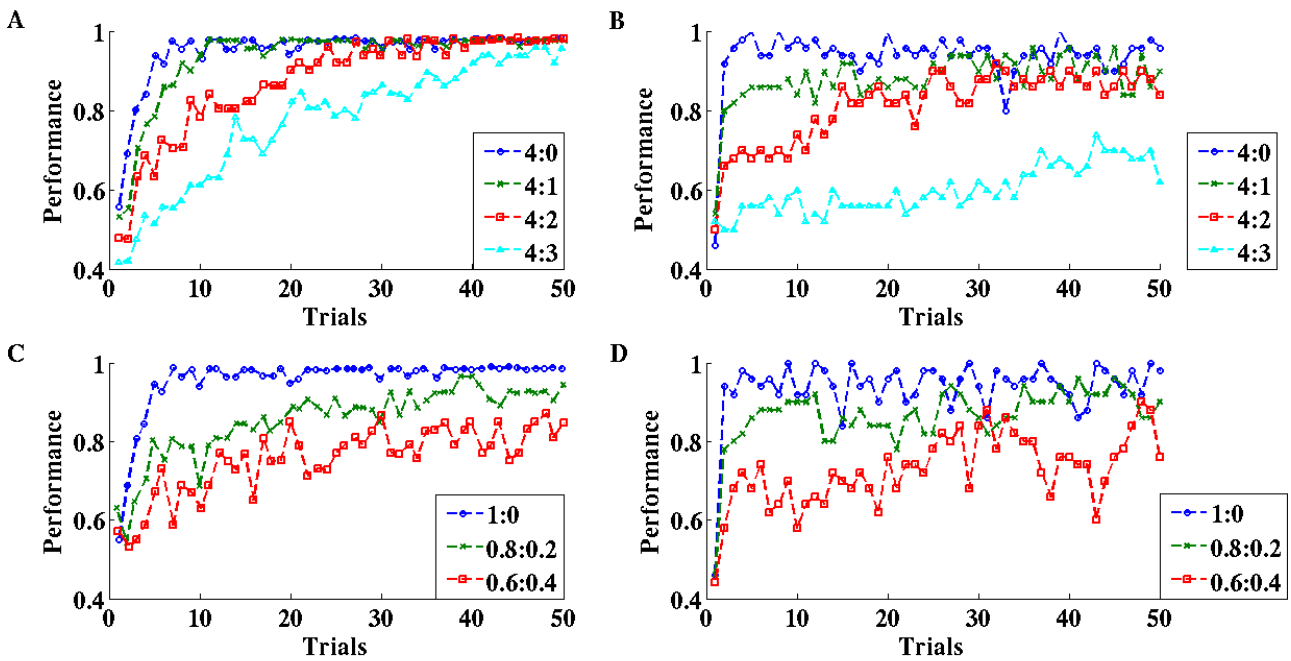

Fig. 4 **A)** Demonstration of change in performance with varying reward magnitudes (Figure adapted from (Lloyd and Leslie 2013)) where the legend indicates the ratio of the magnitude of rewards in the two arms. **B)** Performance of our model on the varying reward magnitude task **C)** Demonstration of change in performance with varying reward probabilities (Figure adapted from (Lloyd and Leslie 2013)) where the legend indicate the ratio of the probability of getting a reward in the two arms. **D)** Performance of our model on the varying reward probability task.

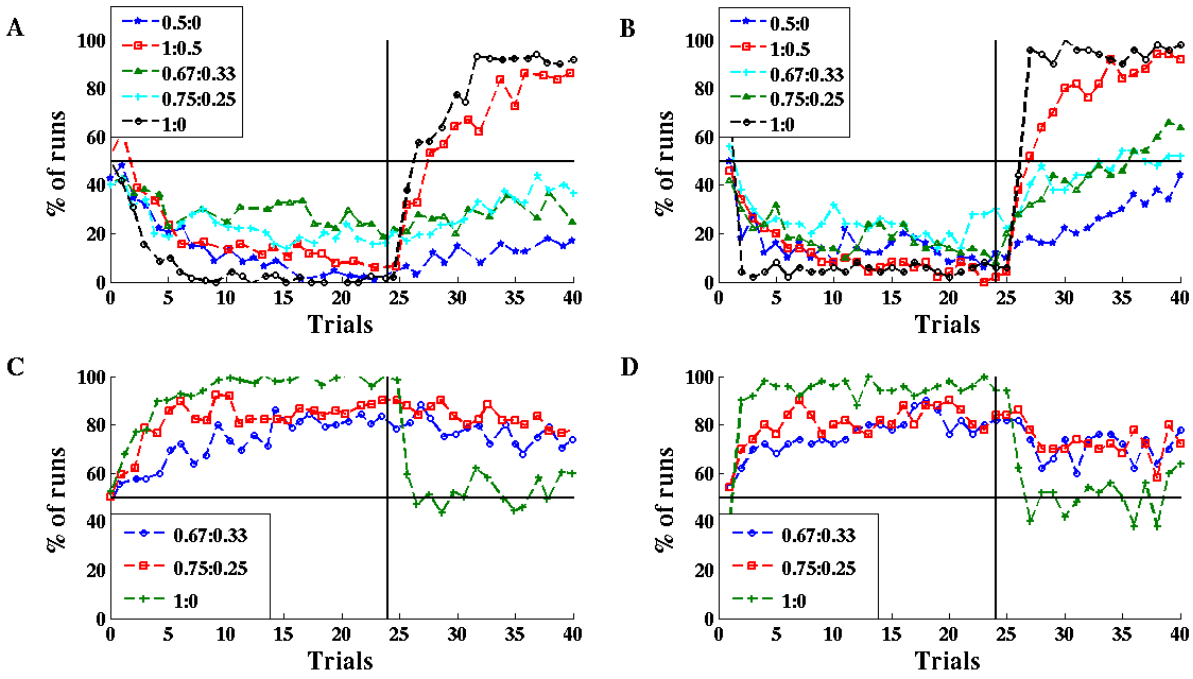

Fig. 5 **A)** Percentage of trials where the animal chooses the arm which is non-profitable for the first 24 trials and becomes profitable following that. (Figure adapted from (Lloyd and Leslie 2013)). **B)** Performance of the model on the task described in **A**. We see that the model shows similar trends where the definite reward tasks show faster reversal learning. **C)** Percentage of trials where the animal chooses the arm which was rewarding before 24 trials following which both arms are not rewarded (Figure adapted from (Lloyd and Leslie 2013)). **D)** Performance of the model on the task described in **C** where the model shows similar trends. The unlearning for the deterministic reward condition is faster than the stochastic reward conditions.

Experimental evidence (Brunswik 1939) shows that partial reinforcement and stochastic rewards have a significant effect on reversal learning. We consider a task where the animal is trained on a T-maze with different reward probabilities for 24 trials and then the rewarding probabilities are reversed. We look at the percentage of the trials where the animal chooses the arm which is unprofitable at first and becomes profitable after the reversal. We can observe that the model results (Fig. 5B) show similar trends to earlier results (Fig. 5A). The tasks with deterministic rewards showed quicker reversal as compared to probabilistic rewards that showed slower policy modulation by the agent.

Stochastic reward distributions also have an effect on extinction (Miltenberger 2011) of a learned policy. To test this, we consider a task where the animal on a T-maze for 24 trials as

above. However, the rewards for both arms are set as 0 following the 24 trials and the rate of unlearning is studied. We observe that definite rewarding tasks show faster extinction as compared to the tasks with stochastic rewards (Fig. 5C) which is captured by the model (Fig. 5D).

Solving Stochastic Reward Based Tasks using the Striatum Model

In this section, we demonstrate that the proposed model of striatum model is capable of solving stochastic reward based tasks. We consider a cue based decision making task where the animal has to choose one of the cues displayed on the screen. This task was first described in (Pasquereau, Nadjar et al. 2007) and a schematic of the task is given in Fig. 6A. The animal is presented with two cues in each trial at two locations (Fig. 6A). Each shape is associated with a different probability of reward. The agent has to choose one of the shapes and gets a reward according to the associated probability.

We show that our striatal model is able to solve this task. We consider a 4 dimensional state vector, where each dimension is 1 if the shape is shown and 0 otherwise. The action vector is also 4 dimensional with each dimension denoting the action that is chosen by the agent. The various parameters of the model are given in Table 2.

Table 2: Parameter values for cue based decision making task

Parameter	Value	Parameter	Value
Strio-SOM Dimension (m_1xn_1)	**3x2**	**Matri-SOM Dimension (m_2xn_2)**	**3x3**
σ_S	**0.01**	$\boldsymbol{\sigma_M}$	**0.1**
η_S	**0.4**	$\boldsymbol{\eta_M}$	**0.4**
γ	**0.95**	$\boldsymbol{\eta^{Str\rightarrow SNc}}$	**0.05**
$\boldsymbol{\eta^{Str(Xm)\rightarrow Str(Q)}}$	**$5x10^{-4}$**	**B**	**50**
α_λ	**0.8**	$\boldsymbol{\eta_\rho^{Str\rightarrow SNc}}$	**0.1**

The agent (model) is pre-trained where it is given various state and action inputs. We show that the representational maps developed have a center-surround structure (Fig. 6C) when we

view the activity corresponding to all the actions for a particular state. The ratio of correct choices chosen in 200 trials averaged over 25 sessions in given in Fig. 6B. Thus, we can see that the agent is able to solve stochastic reward based tasks. Experimental evidence shows that that the percentage of times the agent chooses the arm with reward probability P1, when the ratio of the reward probabilities is P1/(P1+P2), follows a sigmoid activity with center at 0.5 which is well captured by the model (Fig. 6D).

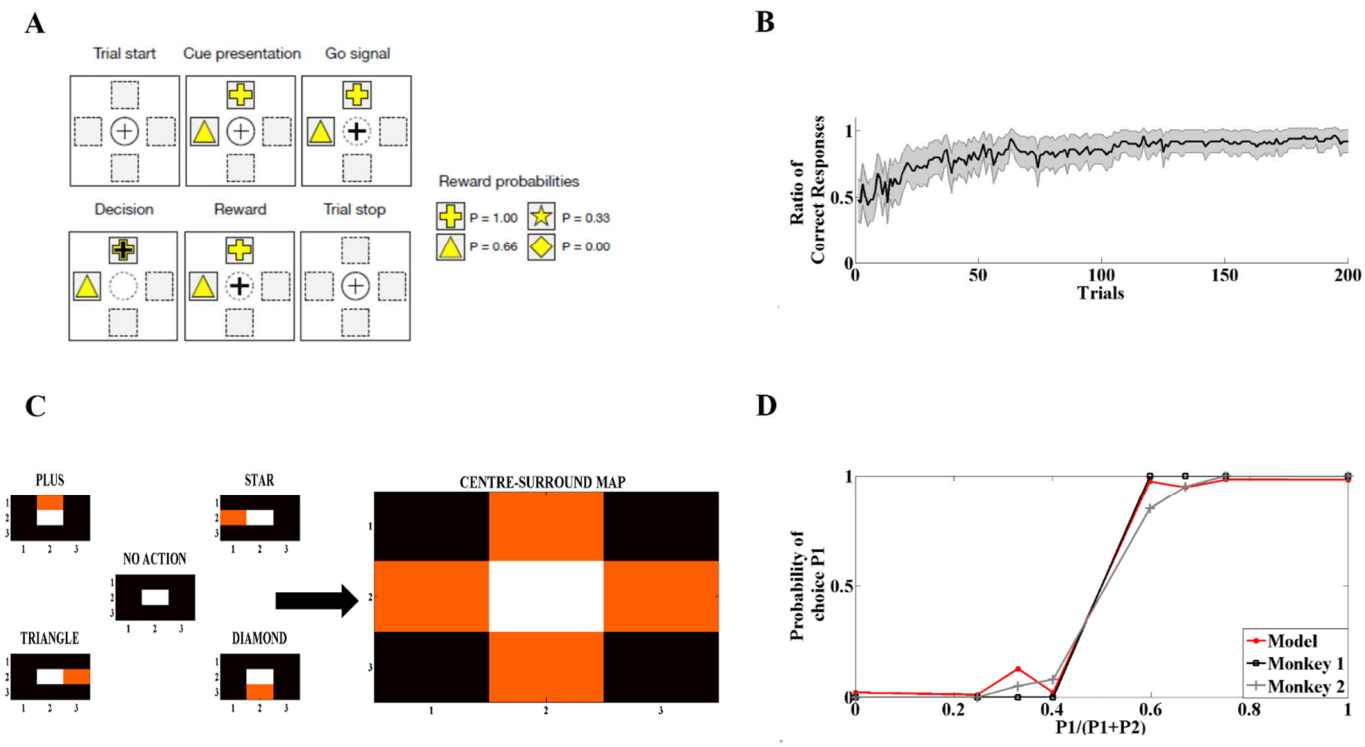

Fig. 6 **A)** Schematic of the cue based decision making task where the agent has to choose between the two shapes shown in the screen and each shape has a different probability of reward associated with it. **B)** Percentage of correct responses averaged over 25 sessions for 200 trials. **C)** Mapping of the action inputs forms a center-surround structure when we view the combined activity of the Matri-SOM for all action inputs **D)** Ratio of choosing response 1 with associated probability P1 w.r.t to the sum P1+P2. The model follows a similar trend to the experimental plot adapted from (Pasquereau, Nadjar et al. 2007)

Comparing the Theoretical and Neural model

We have introduced both a theoretical model capable of solving stochastic multi-context tasks and a neural network model which provides a biologically plausible mechanism for the same task. Since there are no available experiments dealing with these tasks (to the best of our knowledge), we shall use the theoretical model to understand the performances of the neural model. In that regard, we use a stochastic two arm bandit task which was the underlying problem in both the tasks described beforehand. The reward distributions is reversed after 500 trials and the performance of the agent is characterized by averaging performances over 25 sessions. We also observe the performances for different values of ε which represents the probability of reward for the non-profitable arm.

Fig. 7A demonstrates the probability of context 1 estimated by the theoretical model whereas Fig. 7B gives the estimation by the neural network model. We observe that the theoretical model is able to identify the context even for larger values of ε. However, the neural network model is mostly able to identify the context for small values of ε but fails for larger values. A similar trend can be seen in Fig. 8A and Fig. 8B where we measure the percentage of correct choices by the agent. We observe that the theoretical model is able to learn faster upon context reversal for all values of ε but the neural model needs to relearn for higher values of ε.

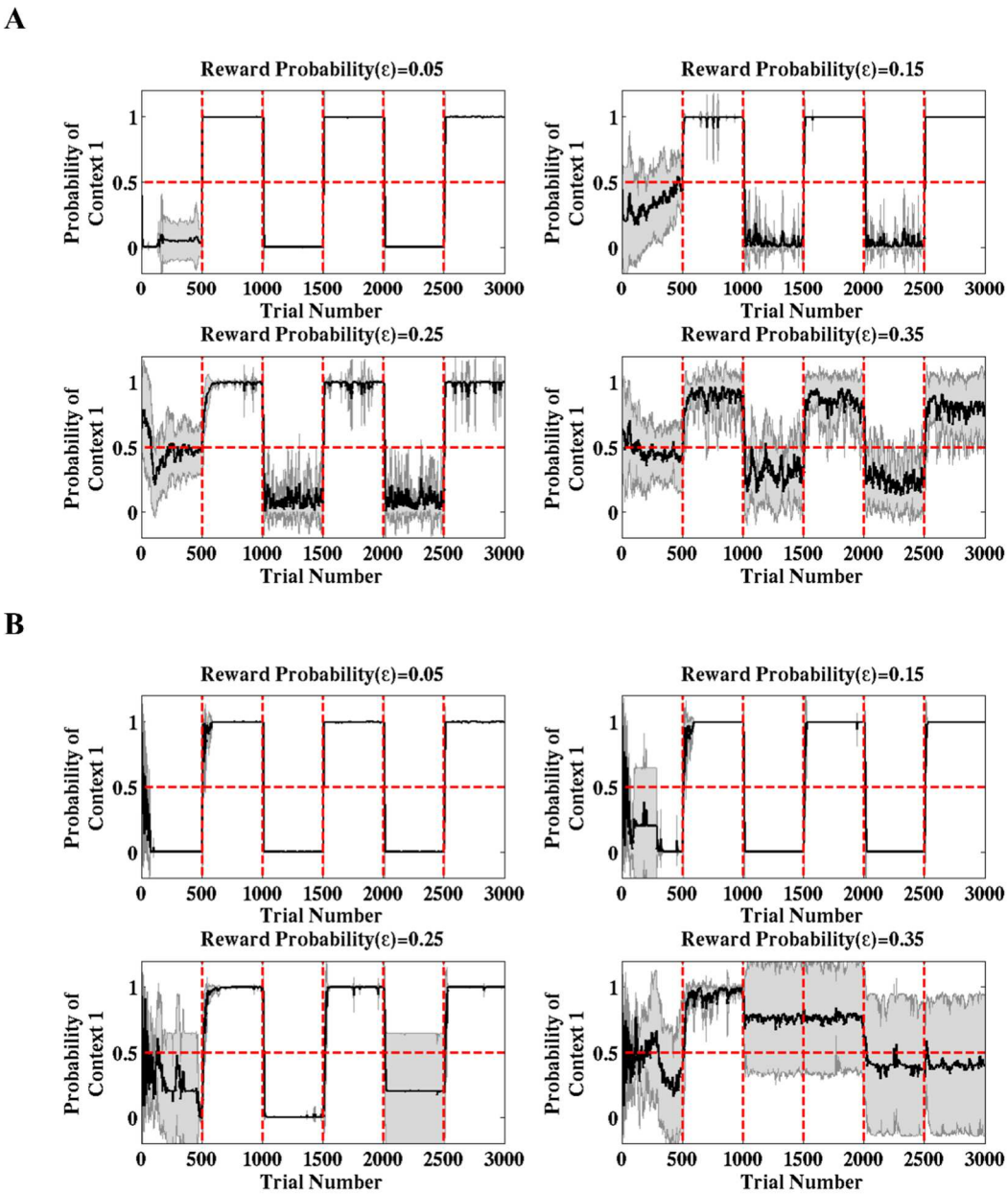

Fig. 7 **A)** Probability of context 1 as estimated by the theoretical model. The vertical red dotted lines indicate the trials where the context changes. The solid black line shows the mean estimate of the probability of context 1 across multiple sessions and the shaded grey region represents the standard error associated with the estimate. **B)** Probability of context 1 as estimated by the neural model. Similar to **A** where the red lines indicate context change, black line indicates the estimate of the probability of context 1 and the grey line the standard error.

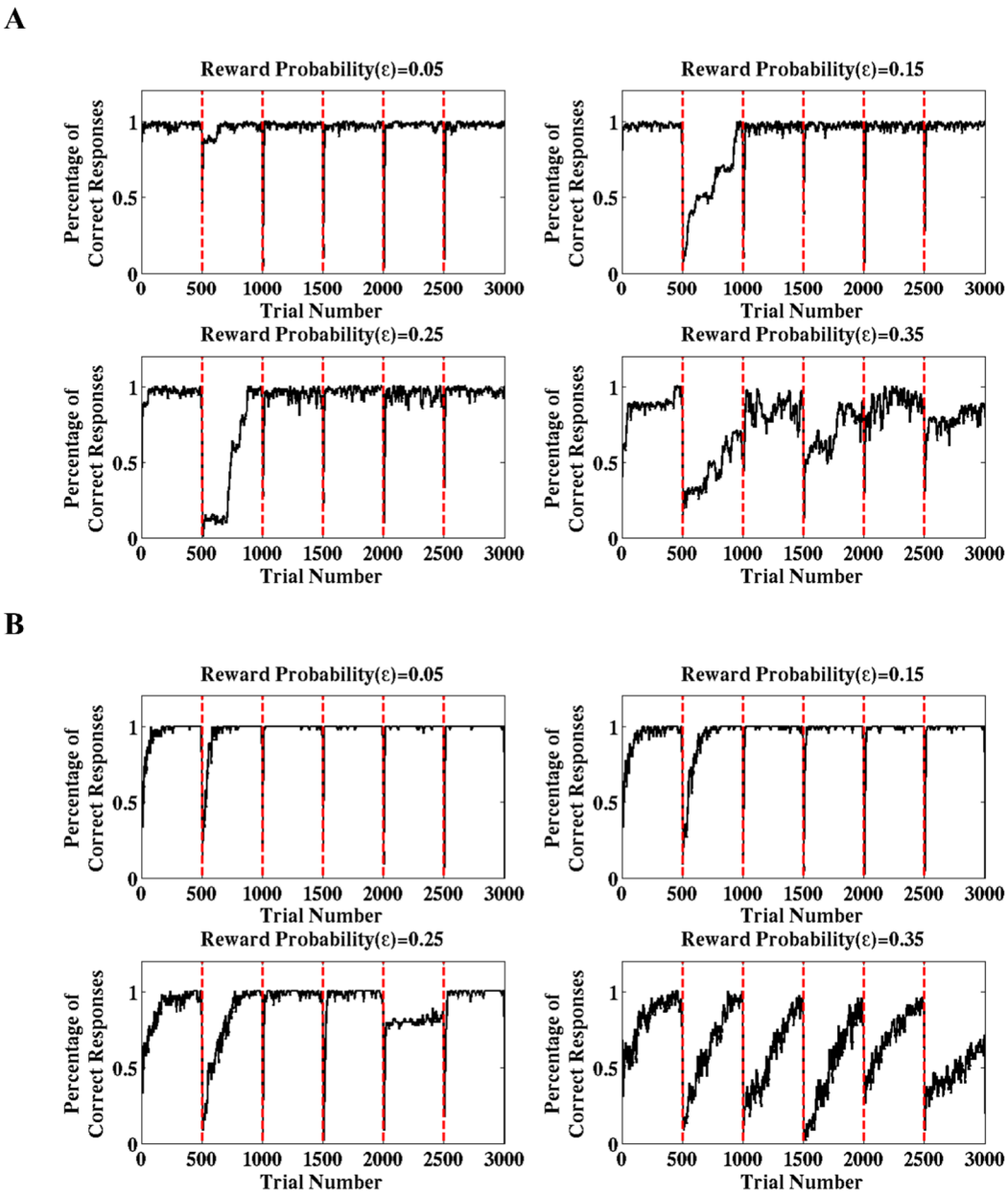

Fig. 8 **A)** Percentage of correct responses by the theoretical model. The vertical red lines indicate the trials where the context changes and the black line denotes the ratio of the correct responses averaged across multiple sessions. **B)** Percentage of correct responses by the neural model. The vertical red lines indicate the trials where the context changes and the black line denotes the ratio of the correct responses averaged across multiple sessions.

From the experimental results, we can conclude that the neural model is able to follow the theoretical model only for low values of ε and behaves like a single context agent for larger values. This can be further seen in Fig. 9 which shows that the neural model performances lie between the theoretical optimal and a single context model and could be the biological mechanism used for solving stochastic multi context tasks.

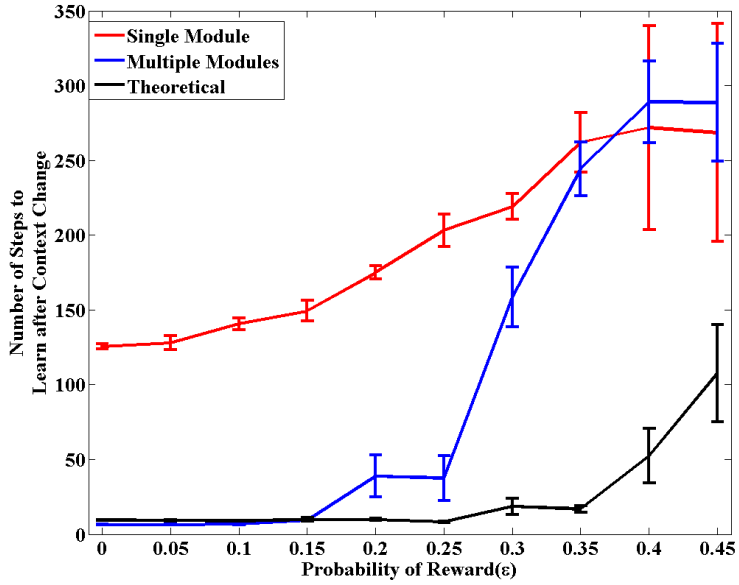

Fig. 9 Schematic of the extended model to handle modular RL tasks showing the case with two striatal modules. The state representations of the two modules are used to calculate their respective responsibilities which are then used by the striatal interneurons to choose the appropriate module.

Discussion

We have presented a theoretical model to solve stochastic multi-context tasks. This is also accompanied by a biologically plausible computational model of the striatum which also attempts to tackle these problems.

Adapting to changing contexts

The problem of identifying a change in context in the environment based on the rewards obtained in the previous trials has been extensively studied in the field of change detection (Basseville and Nikiforov). Given the past history of reward samples upon taking a particular action, Page Hinkley (PH) statistics (Hinkley 1970) is a popular method for testing the hypothesis that a change in context has occurred (Hartland, Gelly et al. 2006, Hartland, Baskiotis et al. 2007). Under the constraint that the rewards come from the exponential family of distributions, PH statistics guarantee minimal expected time before change

detection (Lorden 1971). Our model uses similar ideas of accumulation of mean of rewards in the past trial to predict change in contexts but uses limited memory as a realistic biological constraint. In addition, the model predicts the probability of context change in each trial as opposed to only predicting the occurrence of context change. The model uses information about all the actions in the limited history as opposed to traditional change detection algorithms which assume that the rewards in the history were generated from a single action.

Theoretical Model as a Constrained Version of the Full Bayesian Model

The inherent complexity of stochastic context-dependent problems motivated a Bayesian approach to solve these problems. While the full Bayesian model attempted to give the best possible bound for these tasks, the theoretical model aimed to give a characterization of the expected performance under some realistic constraints such as the ones encountered by an animal solving these tasks. One of the key constraints is the assumption of a limited history. Since the animal has finite memory, it can use information from only a small and recent history to guide its decision making (Todd, Niv et al. 2009). Exploration in action selection is a facet of RL and is also observed in earlier studies (Doya 2008). This is also captured as a constraint in our theoretical model (Eq. 11).

Striatal Microanatomy and Contextual Learning

Our striatal model is derived from a computational model of the basal ganglia proposed for handling context dependent tasks (Shivkumar, Muralidharan et al. 2017). The model is based on the assumption that the striosomes map the state space and the matrisomes map the action space. This is supported from earlier results that the striosomes receive input from the orbitofrontal cortex (Eblen and Graybiel 1995) known for coding reward related states (Wilson, Takahashi et al. 2014). Anatomical studies also show that striosome medium spiny neurons (MSNs) project directly to SNc (Lanciego, Luquin et al. 2012) which could compute state values as in our model.

Evidence suggests that similar to how projections from the striosomes code for state value, projections from the matrisomes code for action value (Doya 2002). Experimental results show the existence of such neurons in the striatum which code specifically for action value (Samejima, Ueda et al. 2005). This is well captured in our model as the Matri-SOM projects to action value neurons in out striatal model.

Action selection is done using the softmax policy (Eq. 24) following the action value computation in the striatum. This policy uses a parameter β which controls the exploration of the agent. We believe that this could be the role of STN, GPe and GPi before action selection is done at the level of the thalamus. This is supported by earlier results which suggest that the underlying stochasticity in the soft-max rule could be achieved indirectly by the chaotic dynamics of the STN-GPe loop (Kalva, Rengaswamy et al. 2012).

Comparing the Theoretical and the Neural Model

The two models proposed in our work were developed and validated independently from each other. However, they share some common features and we can observe that the performance of the neural model falls between the performance of the theoretical model and the neural model with a single module (Fig. 9).

The theoretical model acts as a lower bound to the performance of the neural model for the given stochasticity in the problem. The neural model is also able to achieve performance on par with the theoretical model for low values of ϵ but fails to do so for larger ϵ where it becomes similar to a single module system. Thus, we predict that our neural model can explain behavior in stochastic multi context tasks for $\epsilon < 0.3$. This also allows us to bound performance of the animal performing such tasks in highly stochastic conditions which is challenging from an experimental perspective owing to the large number of trials required.

Another feature of our theoretical model is that it is a very simple model with no assumptions on the reward or the context distributions. However, despite its simplistic formulation, the model is quite powerful and can capture all the previous results reasonably well. The modular arrangement of identifying context and using it for task selection is very similar to the proposed striatal model. Thus, the striatal model could be a biologically plausible neural implementation of the theoretical model.

Acknowledgements

We would like to thank Vignesh Muralidharan for aiding in the development of the neural model.

References

Amemori, K.-i., L. G. Gibb and A. M. Graybiel (2011). "Shifting responsibly: the importance of striatal modularity to reinforcement learning in uncertain environments." Frontiers in human neuroscience **5**: 47.
Basseville, M. and I. V. Nikiforov Detection of abrupt changes: theory and application.
Brunswik, E. (1939). "Probability as a determiner of rat behavior." Journal of Experimental Psychology **25**(2): 175.
Chakravarthy, V. S., D. Joseph and R. S. Bapi (2010). "What do the basal ganglia do? A modeling perspective." Biological cybernetics **103**(3): 237-253.
Charpier, S. and J. Deniau (1997). "In vivo activity-dependent plasticity at cortico-striatal connections: evidence for physiological long-term potentiation." Proceedings of the National Academy of Sciences **94**(13): 7036-7040.
Doya, K. (2002). "Metalearning and neuromodulation." Neural Networks **15**(4): 495-506.
Doya, K. (2008). "Modulators of decision making." Nature neuroscience **11**(4): 410-416.
Doya, K., K. Samejima, K.-i. Katagiri and M. Kawato (2002). "Multiple model-based reinforcement learning." Neural computation **14**(6): 1347-1369.
Eblen, F. and A. M. Graybiel (1995). "Highly restricted origin of prefrontal cortical inputs to striosomes in the macaque monkey." Journal of neuroscience **15**(9): 5999-6013.
Flaherty, A. and A. M. Graybiel (1994). "Input-output organization of the sensorimotor striatum in the squirrel monkey." Journal of Neuroscience **14**(2): 599-610.
Granger, R. (2006). "Engines of the brain: The computational instruction set of human cognition." AI Magazine **27**(2): 15.
Graybiel, A., A. Flaherty and J.-M. Gimenez-Amaya (1991). Striosomes and matrisomes. The basal ganglia III, Springer**:** 3-12.
Graybiel, A. M. (2005). "The basal ganglia: learning new tricks and loving it." Current opinion in neurobiology **15**(6): 638-644.
Hartland, C., N. Baskiotis, S. Gelly, M. Sebag and O. Teytaud (2007). "Change point detection and meta-bandits for online learning in dynamic environments." CAp: 237-250.
Hartland, C., S. Gelly, N. Baskiotis, O. Teytaud and M. Sebag (2006). "Multi-armed bandit, dynamic environments and meta-bandits."
Hinkley, D. V. (1970). "Inference about the change-point in a sequence of random variables." Biometrika: 1-17.
Joel, D., Y. Niv and E. Ruppin (2002). "Actor–critic models of the basal ganglia: New anatomical and computational perspectives." Neural networks **15**(4): 535-547.
Kaelbling, L. P., M. L. Littman and A. W. Moore (1996). "Reinforcement learning: A survey." Journal of artificial intelligence research **4**: 237-285.
Kalva, S. K., M. Rengaswamy, V. S. Chakravarthy and N. Gupte (2012). "On the neural substrates for exploratory dynamics in basal ganglia: a model." Neural Networks **32**: 65-73.
Kohonen, T. (1998). "The self-organizing map." Neurocomputing **21**(1): 1-6.
Lanciego, J. L., N. Luquin and J. A. Obeso (2012). "Functional neuroanatomy of the basal ganglia." Cold Spring Harbor perspectives in medicine **2**(12): a009621.
Langford, J. and T. Zhang (2008). The epoch-greedy algorithm for multi-armed bandits with side information. Advances in neural information processing systems.
Lloyd, K. and D. S. Leslie (2013). "Context-dependent decision-making: a simple Bayesian model." Journal of The Royal Society Interface **10**(82): 20130069.
Lorden, G. (1971). "Procedures for reacting to a change in distribution." The Annals of Mathematical Statistics: 1897-1908.
Miltenberger, R. G. (2011). Behavior modification: Principles and procedures, Cengage Learning.
Olton, D. S. (1979). "Mazes, maps, and memory." American psychologist **34**(7): 583.

Pasquereau, B., A. Nadjar, D. Arkadir, E. Bezard, M. Goillandeau, B. Bioulac, C. E. Gross and T. Boraud (2007). "Shaping of motor responses by incentive values through the basal ganglia." Journal of Neuroscience **27**(5): 1176-1183.
Samejima, K., Y. Ueda, K. Doya and M. Kimura (2005). "Representation of action-specific reward values in the striatum." Science **310**(5752): 1337-1340.
Schultz, W. (2004). "Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology." Current opinion in neurobiology **14**(2): 139-147.
Seo, M., E. Lee and B. B. Averbeck (2012). "Action selection and action value in frontal-striatal circuits." Neuron **74**(5): 947-960.
Shivkumar, S., V. Muralidharan and V. S. Chakravarthy (2017). "A Biologically Plausible Architecture of the Striatum to Solve Context-Dependent Reinforcement Learning Tasks." Frontiers in neural circuits **11**.
Sullivan, M. A., H. Chen and H. Morikawa (2008). "Recurrent inhibitory network among striatal cholinergic interneurons." Journal of neuroscience **28**(35): 8682-8690.
Sutton, R. S. and A. G. Barto (1998). Reinforcement learning: An introduction, MIT press Cambridge.
Todd, M. T., Y. Niv and J. D. Cohen (2009). Learning to use working memory in partially observable environments through dopaminergic reinforcement. Advances in neural information processing systems.
Wilson, R. C., Y. K. Takahashi, G. Schoenbaum and Y. Niv (2014). "Orbitofrontal cortex as a cognitive map of task space." Neuron **81**(2): 267-279.
Yu, A. and P. Dayan "Expected and unexpected uncertainty: ACh and NE in the neocortex."