Online navigation with neuromodulation-based Hebbian plasticity

Krubeal Danieli

December 1, 2024

Contents

1 Introduction

The ability to make decisions for long-term reward maximization is a fundamental aspect of cognition. The brain has evolved a complex web of interconnected regions that work together to express this behaviour under the constraints of biology. The Pre-Frontal Cortex (PFC) is considered a fundamental high-level region for the attention and cognitive control, in particular the medial PFC [1, 2]. Further, the orbitofrontal cortex (OFC) is thought to be involved in motivation and representation of the expected value of the actions, either positive or negative [3, 4, 5], and action selection in uncertain environments [6]. A relevant element for online executive functions is working memory, which is usually defined as the capacity to hold and manipulate information over short periods of time [7]; thus functionality has been associated with the dorsolateral PFC [8, 9, 10, 11]. However, it has also been suggested that the PFC is instead exercising a more top-down control over more sensory regions [12]. This cortical projectons have been proposed to target also the basal ganglia, which are thought to rely on first-order reward statistics, while the OFC is able to capture more complex contextual dynamics [13].

Regarding decision-making tasks, simple and well-studied ecological settings are foraging problems, *e.g.* food search, where an agent is set to choose between different options to maximize the expected reward. Depending on context, animals have been shown to exhibit different strategies. In this regard, *matching behaviour* is a well-known phenomenon in which the animal's choice behaviour is proportional to the reward probability of the options. This behaviour is thought to be the result of a trade-off between exploration and exploitation [14, 15], where the animal must balance the need to explore new alternatives with the need to exploit the best option found so far. Other options are *input matching*, where social cues are considered, and *probability matching*, where the animal's choice behaviour is proportional to the reward probability of the options [16, 17]. A popular formalization of such tasks in optimal decision theory is the "multi-armed bandit problem" (MABP) [18], where an agent is faced with a set of K possible actions, each one associated with an unknown reward probability distribution. The agent has to learn to choose the action that maximizes the expected reward by repeatedly selecting actions and observing the rewards obtained. This problem has been extensively studied in the context of reinforcement learning, and it is considered a fundamental building block for more complex tasks [14].

Extensive research has been conducted on the topic, and several algorithms have been proposed, such as Thompson sampling, ϵ-greedy, UCB1, VDBE, alongside convergence proofs for specific settings [19, 20, 21, 22, 23].

1.1 Related work

There exists various flavours of this problem, with the simplest having a stationary reward distribution, while the more challening ones have have *concept drift*, where the reward distribution changes over time. Over the years, several algorithms have been proposed, alongside with their theoretical guarantees. In this regard, Thompson sampling is a popular algorithm that has been shown to achieve near-optimal regret bounds in the stochastic setting [24], which a Bayesian approach the idea of maintaining a posterior distribution over the reward probabilities of the actions, and selecting actions according to the posterior distribution. Another popular algorithm is the Upper Confidence Bound (UCB) algorithm, which has been shown to achieve near-optimal regret bounds in the adversarial setting [25]. The algorithm is based on the idea of maintaining an upper confidence bound on the reward probabilities of the actions, and selecting actions according to the upper confidence bound.
Despite the success of these algorithms in solving the k-armed bandit problem, they lack biological plausibility. In contrast, the brain has evolved a complex network of interconnected regions that work together to solve this task. In particular, the dopamine-acetylcholine system has been shown to play a crucial role in learning and decision making [26].

In this work, we focus on a stochastic bandit problem, a more challenging variant of the original task endowed with *concept drift*, where the reward distribution changes over time [27, 28, 29]. We propose a biologically plausible model using spiking neural networks (SNN), obtaining good performance over an arbitrary number of bandit trials. Its architecture is composed of two parts. The first is a working memory component, whose scope is to maintain an active representation of the current stimulus, and it is inspired by the functionality of the dorsolateral PFC. Several previous studies have proposed bio-realistic models of WM [30], including focus on synaptic facilitation [31], random connectivity [32], excitation-inhibition balance (E-I) [33, 34], and echo state networks [35, 36]. Here in particular, we used a three-population SNN exploiting the E-I balance to maintain the stimulus trace, similarly to [37]. The second component is a one-layer hybrid network, and it is where a pre-defined architectural policy assign a subjective value to the available options *i.e.* the bandits. This component is modeled after the interaction between the OFC and the basal ganglia, in particular the striatum [16, 13], given the role of the frontal lobe in reversal learning [38], various network models of the PFC and BG for decision-making have been proposed, centered on the role of dopamine in the basal ganglia [39], spike-time dependent plasticity (STPD) [40], cortico-striatal interaction [41], and task switching processes [42, 43].

Regarding our model, the novelty relies in two main elements. The first is the decision making process itself, implemented as a dynamic population interaction between the neural traces active in working memory and the values represented in the hybrid network, similarly to bump attractor networks for perceptual computations [44, 45]. The second is learning, which is applied to the hybrid network, and it is based on a neuromodulated Hebbian-like synaptic plasticity rule with special weight-dependent kernels for long-term potentiation (LTP) and long-term depression (LTD). This choice is motivated by the asymmetry of the dopamine signal in supporting LTP or LTD [46, 47, 48, 49].

2 Methods

The section is organized as follows. First, we introduce a formalization of general problem setting, together with the variants considered in this work. Then, we outline the architecture of the our model and how it can be mapped to neurobiology. Finally, we describe the learning procedure, and showcase its dynamics in a simple example.

2.1 Binomial K-armed bandit problem

The standard formulation of the task is structured as a set of $\{1 \ldots K\}$ levers (or arms), with an associated reward distribution $\mathbf{p} = \{p_1, \ldots p_K\}$. At each iteration, the agent pulls a lever and collect a possible reward drawn as a Bernoulli variable $R \sim \mathcal{B}(\{0,1\}, p_k)$. The agent's objective is maximizing the total reward $\sum_t^T R_t$, after a certain number T of trials. Importantly, the agent is unaware of the true reward probability distribution, and thus has to make its decisions following a certain policy, usually denoted as ω. In the reinforcement learning literature, the policy is often defined as a distribution over the actions, here the levers K, given the current state, which in this case can be the history of past actions and rewards up to time $t \leq T$. Given the inherent stochasticity of the feedbacks from the environment, the definition of the policy is affected by the so-called exploration-exploitation trade-off, which here is phrased as the contrast between the option of the lever with the known highest expected reward versus the option to explore other levers, so to gather more information. A common approach is the $\epsilon-$greedy policy, where the choice to explore is

selected with a probability ϵ. Moreover, it is often preferable to have a more explorative behaviour early during the training, with the intent to have a good sample size for the empirical reward distribution, which can be later exploited for maximizing reward.

Another important concept in multi-armed bandit problems is the *regret*. Intuitively, it is defined as the deviation of the total reward obtained by the agent from the optimal reward that could have been obtained by always choosing the lever with the highest expected reward. Formally, the regret is defined as:

$$\rho = R^* - \sum_t^T R_t \tag{1}$$

where R^* is the reward obtained by always choosing the lever with the highest expected reward $R^* = T \max_k \{p_k\}$, and R_t is the empirical reward obtained up to time t by following policy ω as $R_t = \sum_{t=1}^T \omega_\theta(t)$. The regret is a measure of the performance of the agent, and it is often used to compare different algorithms. The goal of the agent is to minimize the regret, and thus maximize the total reward.

2.2 Model description

The model is constructed as a rate network of two populations of neurons M and V, the former representing the memory trace of the K available options (*i.e.* the bandits), and the latter the value of the options under the current policy. More formally, the model is defined by a set of coupled ordinary differential equations (ODEs). The first equation tracks the evolution of the neural activity $\mathbf{u}$ of population M, while the second tracks the activity $\mathbf{v}$ of the population V. The time constant τ is the same for both equations and it is set to 10ms.

$$\begin{aligned} \tau\dot{\mathbf{u}} &= -\mathbf{u} + \mathbf{W}^{VM}\mathbf{v} + \mathbf{I}_{\text{ext}} \\ \tau\dot{\mathbf{v}} &= -\mathbf{v} + \tilde{\mathbf{W}}^{MV}\mathbf{u} \end{aligned} \tag{2}$$

The external input $\mathbf{I}_{\text{ext}}$ is a constant input that is used to set the initial conditions of the neural activity $\mathbf{u}$.

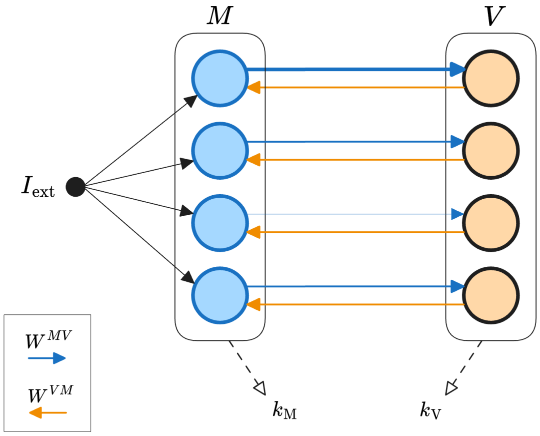

Figure 1: Model architecture - *The model is composed of a layer M (blue), receiving a feedfoward input I_{ext}, a layer V (orange), and connections $\boldsymbol{W}^{MV}$ and $\boldsymbol{W}^{VM}$. Additionally, two indexes k_M, k_V can be extracted from the layers and corresponds to the selection made by the two populations as $k_M = argmax_k\{\boldsymbol{u}\}$, $k_V = argmax_k\{\boldsymbol{v}\}$.*

Importantly, the two layers are not fully connected and the matrices are diagonal. Further, the weight matrix $\mathbf{W}^{VM}$ is simply the identity, while $\tilde{\mathbf{W}}^{MV}$ is a function of the actual weights $\Phi_v(\mathbf{W}^{MV})$ and it represents the contribution of the active options $\mathbf{u}$ to the value representation $\mathbf{v}$. The function Φ_v is defined as the sum of a generalized sigmoid and a Gaussian, whose shape is characterized by a bell curve smoothly settling to a constant value. See more in the appendinx 5.

2.2.1 Option selection

The decision-making process within a single round is structured in two distinct phases. Initially, the model receives a constant external input targeting all neurons in the memory population M equally. During this phase, $\mathbf{I}_{\text{ext}}$ works as an equilibrium value while the reciprocal interactions with population V push $\mathbf{u}$ to different values, depending on the current policy encoded in $\tilde{\mathbf{W}}^{MV}$. However, in the early rounds the weights $\mathbf{W}^{MV}$ are zero, and thus the contribution from V is null. After a fixed amount of time $\sim$ 5s, the second phase begins. Here, the external input is removed and the model is left to evolve autonomously, and since there are no recurrent connections in neither population the dynamics are entirely driven by their coupling. A selection k is sampled after another fixed amount of time $\sim$ 5s, and it is defined according to the following rule:

$$k = \begin{cases} \text{argmax}_k\{\mathbf{v}\} & \textit{if } \text{argmax}_k\{\mathbf{v}\} = \text{argmax}_k\{\mathbf{u}\} \\ \text{random}(K) & \textit{otherwise} \end{cases}$$

The selection rule is simple: if the value representation $\mathbf{v}$ is in agreement with the memory trace $\mathbf{u}$, then the option with the highest value is selected. Otherwise, a random option is chosen. This rule is a way to express the exploration-exploitation trade-off, and it is dependent on the current policy $\tilde{\mathbf{W}}^{MV}$.
Below 2.2.1, is reported the pseudo-code for algorithm behind the selection process.

Algorithm 1: Two-phases option selection process

```
Input: External input $\mathbf{I}_{\text{ext}}$, memory population $\mathbf{u}$, value population $\mathbf{v}$,
       policy weights $\tilde{\mathbf{W}}^{MV}$
Output: Selected action $k$
Phase 1: external input ;                                   // Duration: ~5s
Define constant $\mathbf{I}_{\text{ext}}$;
Update populations $\mathbf{u}, \mathbf{v}$ according to 2.2;
Phase 2: autonomous evolution ;                             // Duration: ~5s
Remove external input $\mathbf{I}_{\text{ext}}$;
Let system evolve through population coupling according to 2.2;
Selection process:;
$k_u \leftarrow \text{argmax}_k\{\mathbf{u}\}$;
$k_v \leftarrow \text{argmax}_k\{\mathbf{v}\}$;
if $k_u = k_v$ then
    $k \leftarrow k_v$ ;                                    // Exploitation
else
    $k \leftarrow \text{random}(K)$ ;                       // Exploration
end
return $k$
```

maybe place this directly in the results?

In figure 2 it is shown the history of selections over three trials. The initial rounds features higher variability. In particular, it can noted how the policy adopted by the model encounters period of exploration and successive settling over an explotative strategy, which can be reverted in case of a change in the environment's reward distribution.

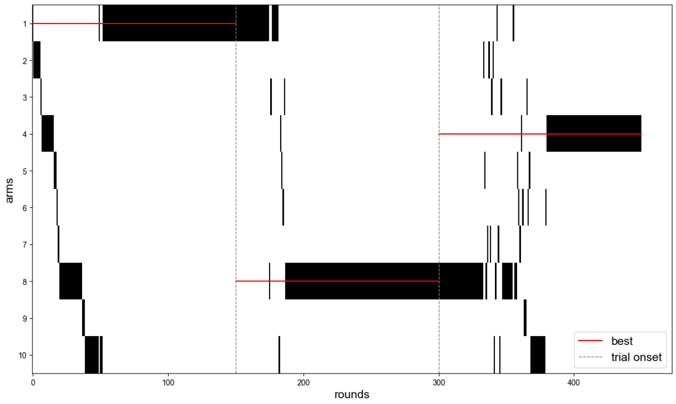

Figure 2: Selection evolution over rounds - *the x-axis represents the available arms, while the y-axis the number of rounds, with the dotted vertical lines indicating the start of a new trial with 150 rounds each. The model selections are the black vertical lines for an arm and a round. The red horizontal lines signal the arm with the highest reward probability, thus representing the best (and greediest) selection.*

2.3 Learning

Given a selected option k, the environment (set of bandits) samples and returns a reward $R \in [0, 1]$ with probability p_k. Then, the connections $\mathbf{W}^{MV}$ for the neuron corresponding to the option k are updated according to the following plasticity rule:

$$\Delta \mathbf{W}_k^{MV} = \tilde{\eta}_k \left(R \cdot W^+ - \mathbf{W}_k^{MV} \right) \tag{3}$$

Where W^+ is a constant value that sets the upper bound for the synaptic weights, and it is set to $W^+ = 5$, while $\tilde{\eta}_k$ is the learning rate for the option k determined by a function of the current weights $\mathbf{W}_k^{MV}$ and its shape is the same as Φ_v, but with different parameters.

2.4 Bio-inspired features

The model is inspired by the functioning of the prefrontal cortex (PFC) and its importance in decision-making processes. In particular, despite their marked simplicity, the two population M, V of the model can be related to the orbitofrontal cortex (OFC) and anterior cingulate cortex (ACC), respectively. More specifically, the OFC is known to be involved in the representation of the state different options and update their value with respect to rewarding outcomes and their history [50, 51]. The ACC has been associated to action values, and the dynamic interplay with OFC is observed to elicit transient pre-stimulus activation, which biases the decision towards the most valuable option [52, 53, 54]. In the model, the first layer represents the available options, while the learned connections with the second layer encode their values based on the recent reward history. Another similarity with this particular pre-frontal circuit is the realization of a choice as a sample of the network state after a period of autonomous neural activity, where the depth of the closest neural attractor depends on the strength and reliability of the highest option value [55, 56].

3 Results

The model has been tested in a series of benchmark environments, each with a different number of arms and reward distributions. The performance has been compared with the following algorithms: Random Baseline, Upper-Confidence Bound (UCB), Thompson Sampling, and Epsilon-Greedy. The results are summarized in table 3.

3.1 Game variants

The game environments considered in this work are non-stationary K-Armed Bandits with Binomial rewards. In particular, the agent is evaluated over a number T_{trials} of *trials*, each composed by an arbitrary number T_{rounds} of *rounds*; each *trial* is characterized by a different reward distribution $\mathbf{p} \sim \mathcal{U}(0,1)^K$ (although in practice the bounds have been set to $(0.1, 0.9)$ such that the distributions are less trivial). Our goal in this work is to investigate the performance of the agent in a non-stationary environment with Binomial reward distributions, meaning that its underlying distribution changes over time. We choose this setting as it resembles an ecological scenario in which an animal has to forage in a patchy environment, where the reward of a given patch can change over time. More specifically, we used four different variants:

Zero-steps distribution shift [KAB-0]: the reward distribution changes immediately at the end of a trial i to a new one $i + 1$ as $\omega_i \to \omega_{i+1}$.

Epsilon-steps distribution shift [KAB-ϵ]: the reward distribution ω changes gradually over rounds, tracked as time t, such that its shape tends towards a target distribution $\bar{\omega}_i$ as $\tau_\omega \dot{\omega}_t = \bar{\omega}_i - \omega_t$. Once distance is below a threshold ϵ as $|\bar{\omega}_i - \omega_t| < \epsilon$, the target distribution is changed to a new one $\bar{\omega}_i \to \bar{\omega}_{i+1}$.

Sinusoidal distribution shift [KAB-sin]: the reward distribution changes over rounds, with the probability of each arm following a sine wave with a specific frequency f_k and amplitude 1. At any given time t, the distribution is $\omega_t = \{\sin(2\pi f_k t) \text{ for } k = 1 \dots K\}$ and it is normalized as $\omega_t = \omega_t(\sum_k \omega_{t,i})^{-1}$ such that it sums to 1.

Partial sinusoidal distribution shift [KAB-sinP]: identical to the sinusoidal distribution shift, but only a subset of the arms changes sinusoidally while the rest is kept at a constant value.

3.2 Performance comparison

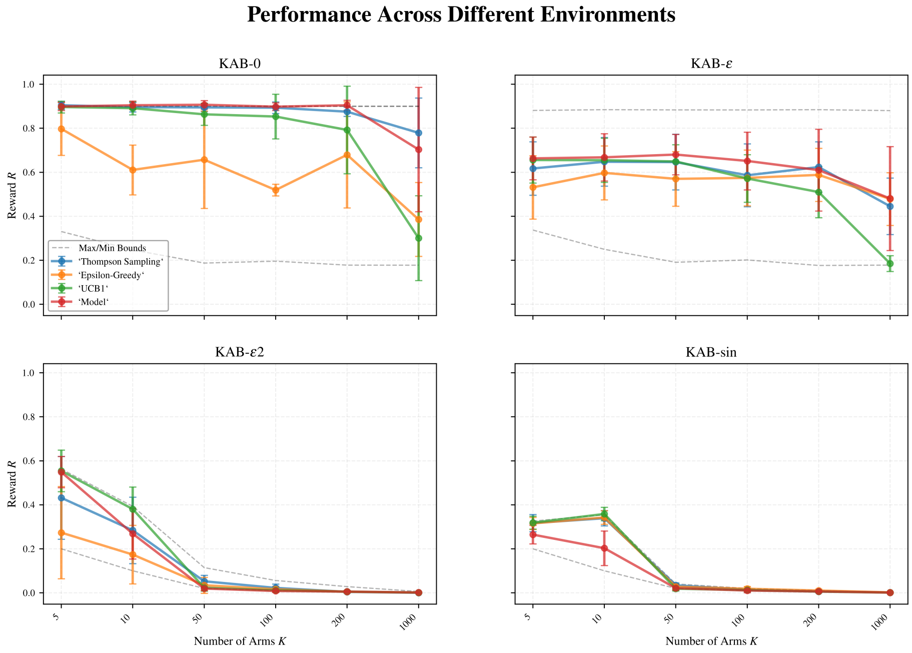

Figure 3: Performance comparison for different values of K and game variants *The model is compared with Thompson Sampling, Epsilon-Greedy, and UCB. The performance is measured as the average reward obtained by the agent over a number of trials.*

3.3 Decision-making dynamics

For a better understanding of the qualitative differences between the models, we analyzed the progress over the rounds and tracked the selected arms in the simplest case of zero-steps distribution shift. Additionally, in order to quantify the variability of the decision policy at a given time and highlight the particularity of each decision-making behaviour, we calculated the entropy of the distribution of chosen arms over a time window of 20 rounds as $H = -\sum_i^K p_i \log(p_i)$. In figure 4, it is plotted for each model the raster plot of selected arms together with its level of entropy. As expected, the over shape of the changes in the entropy over time are rather specific to each model. In particular, the UCB algorithm showed the highest variability, marked by a persistent exploratory behaviour throughout the trials despited converging to reward options. Thompson Sampling was able to reach most solutions, although with difficulty in adapting to new reward distributions leading to high entropy levels. $\epsilon-$Greedy also showed a good performance quite reliably, with the greedy strategy assuring low entropy for most of the rounds. Similar behaviour was observed for our model, which was able to reach the optimal policy and maintain it over time, with entropy peaking mostly at the beggining of the trials and being, on average, the lowest among all models. Indeed, the model dynamics make it particularly suited for the task of non-stationary K-armed bandits, as it is able to quickly adapt to new reward distributions and firmly maintain a greedy policy.

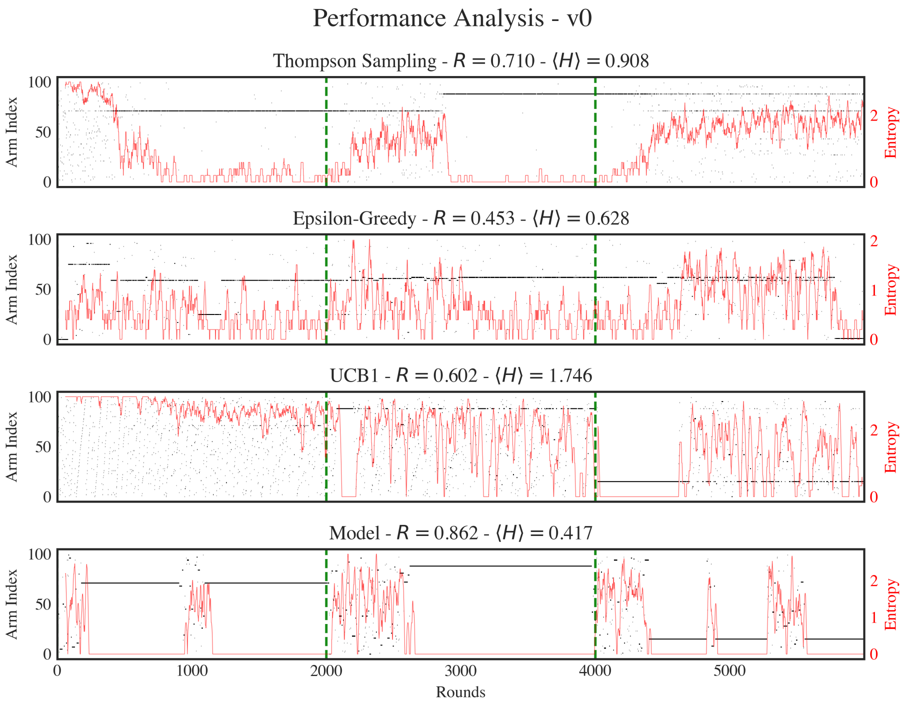

Figure 4: Decision-making dynamics for different models *The raster plot shows the selected arms over time (black dots), while the entropy plot shows the variability of the decision policy (red lines). It is also reported the total reward and average entropy over all trials.*

3.4 Robustness and parameter sensitivity

4 Discussion

In the context of human behaviour, it has been observed that the adopted policies vary considerably [57]. However, the subjects seems able to integrate environmental uncertainty and trial generalization in their strategy, and Bayesian algorithms are generally a good fit for the observed behaviour [58, 59].

In this work, we introduced a model based on rate neural networks that leverages multi-stable network manifolds and Hebbian-like synaptic plasticity to address the binomial K-armed bandit problem in a non-stationary environment. Our results prove the model's robustness and adaptability in dynamic settings where reward distributions change unpredictably.

The observed efficiency of our model can be attributed to the interplay of distinct structures at with different timescales [58, 48, 13], who emphasize the significant role of striato-orbitofrontal interactions in decision-making and learning.

Moreover, the dynamics of our model echo real-world decision-making processes observed in humans and other animals, where decision-making strategies shall adapt to changing environmental conditions [15]. This adaptability is replicated in the performance of our model, which effectively handles the zero-steps and epsilon-steps distribution shifts, there the probabilities change with within a variable number of steps.

Furthermore, the use of a Hebbian-like rule for synaptic updates in our model introduces a level of flexibility and responsiveness that is not commonly found in traditional reinforcement learning algorithms, which often rely on fixed learning rates or reward probabilities [14]. This approach may explain the superior performance of our model in environments where adaptability is critical [28].

In summary, the results of our study not only support the feasibility of using s for complex decision-making tasks but also highlight the potential of neuromodulatory systems to enhance the adaptability and efficiency of artificial neural networks. Future work could explore the integration of additional biological elements, such as the role of other neuromodulators like serotonin and acetylcholine, to further improve the model's performance and biological fidelity [60, 26].

This work was supported by the European Union (#add more).
The authors declare no competing interests.
The code is publicly available and can be found at `https://github.com/iKiru-hub/minBandit.git` (#change to Zenodo).

References

[1] Earl K. Miller and Jonathan D. Cohen. An Integrative Theory of Prefrontal Cortex Function. *Annual Review of Neuroscience*, 24(1):167–202, March 2001.

[2] Denis Sheynikhovich, Satoru Otani, Jing Bai, and Angelo Arleo. Long-term memory, synaptic plasticity and dopamine in rodent medial prefrontal cortex: Role in executive functions. *Frontiers in Behavioral Neuroscience*, 16, January 2023.

[3] J. O'Doherty, M. L. Kringelbach, E. T. Rolls, J. Hornak, and C. Andrews. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience*, 4(1):95–102, January 2001.

[4] Justin S. Riceberg and Matthew L. Shapiro. Reward Stability Determines the Contribution of Orbitofrontal Cortex to Adaptive Behavior. *Journal of Neuroscience*, 32(46):16402–16409, November 2012.

[5] Léon Tremblay and Wolfram Schultz. Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708, April 1999.

[6] Rebecca Elliott, Raymond J. Dolan, and Chris D. Frith. Dissociable Functions in the Medial and Lateral Orbitofrontal Cortex: Evidence from Human Neuroimaging Studies. *Cerebral Cortex*, 10(3):308–317, March 2000.

[7] Alan D. Baddeley and Graham Hitch. Working Memory. In *Psychology of Learning and Motivation*, volume 8, pages 47–89. Elsevier, 1974.

[8] Kimberlee D'Ardenne, Neir Eshel, Joseph Luka, Agatha Lenartowicz, Leigh E. Nystrom, and Jonathan D. Cohen. Role of prefrontal cortex and the midbrain dopamine system in working memory updating. *Proceedings of the National Academy of Sciences*, 109(49):19900–19909, December 2012.

[9] Jonathan D. Cohen, William M. Perlstein, Todd S. Braver, Leigh E. Nystrom, Douglas C. Noll, John Jonides, and Edward E. Smith. Temporal dynamics of brain activation during a working memory task. *Nature*, 386(6625):604–608, April 1997.

[10] Christos Constantinidis, Shintaro Funahashi, Daeyeol Lee, John D. Murray, Xue-Lian Qi, Min Wang, and Amy F. T. Arnsten. Persistent Spiking Activity Underlies Working Memory. *Journal of Neuroscience*, 38(32):7020–7028, August 2018.

[11] Joel Zylberberg and Ben W. Strowbridge. Mechanisms of Persistent Activity in Cortical Circuits: Possible Neural Substrates for Working Memory. *Annual Review of Neuroscience*, 40(Volume 40, 2017):603–627, July 2017.

[12] Antonio H. Lara and Jonathan D. Wallis. The Role of Prefrontal Cortex in Working Memory: A Mini Review. *Frontiers in Systems Neuroscience*, 9, December 2015.

[13] Michael J. Frank and Eric D. Claus. Anatomy of a decision: Striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, 113(2):300–326, April 2006.

[14] Richard S. Sutton and Andrew G. Barto. The Reinforcement Learning Problem. In *Reinforcement Learning: An Introduction*, pages 51–85. MIT Press, 1998.

[15] Yael Niv, Daphna Joel, Isaac Meilijson, and Eytan Ruppin. Evolution of Reinforcement Learning in Uncertain Environments: A Simple Explanation for Complex Foraging Behaviors. *International Society for Adaptive Behavior*, 2002.

[16] Bilal A. Bari and Jeremiah Y. Cohen. Dynamic decision making and value computations in medial frontal cortex. *International review of neurobiology*, 158:83–113, 2021.

[17] Alasdair I. Houston, Pete C. Trimmer, and John M. McNamara. Matching Behaviours and Rewards. *Trends in Cognitive Sciences*, 25(5):403–415, May 2021.

[18] Bruno B. Averbeck. Theory of Choice in Bandit, Information Sampling and Foraging Tasks. *PLoS Computational Biology*, 11(3):e1004164, March 2015.

[19] J. C. Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177, 1979.

[20] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Asymptotically Optimal Finite Time Analysis, July 2012.

[21] Yikun Ban, Jingrui He, and Curtiss B. Cook. Multi-facet Contextual Bandits: A Neural Network Perspective, June 2021.

[22] Michel Tokic. Adaptive ε-Greedy Exploration in Reinforcement Learning Based on Value Differences. In Rüdiger Dillmann, Jürgen Beyerer, Uwe D. Hanebeck, and Tanja Schultz, editors, *KI 2010: Advances in Artificial Intelligence*, volume 6359, pages 203–210. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[23] Michel Tokic and Günther Palm. Value-Difference Based Exploration: Adaptive Control between Epsilon-Greedy and Softmax. In Joscha Bach and Stefan Edelkamp, editors, *KI 2011: Advances in Artificial Intelligence*, volume 7006, pages 335–346. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[24] Shipra Agrawal and Navin Goyal. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 39.1–39.26. JMLR Workshop and Conference Proceedings, June 2012.

[25] Peter Auer and Nicolo Cesa-Bianchi. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 2002.

[26] Peter Dayan and Nathaniel D. Daw. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4):429–453, December 2008.

[27] Aurélien Garivier and Eric Moulines. On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems, May 2008.

[28] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[29] Emanuele Cavenaghi, Gabriele Sottocornola, Fabio Stella, and Markus Zanker. Non Stationary Multi-Armed Bandit: Empirical Evaluation of a New Concept Drift-Aware Algorithm. *Entropy*, 23(3):380, March 2021.

[30] Omri Barak and Misha Tsodyks. Working models of working memory. *Current Opinion in Neurobiology*, 25:20–24, April 2014.

[31] Omri Barak, Misha Tsodyks, and Ranulfo Romo. Neuronal Population Coding of Parametric Working Memory. *Journal of Neuroscience*, 30(28):9424–9430, July 2010.

[32] Flora Bouchacourt and Timothy J. Buschman. A Flexible Model of Working Memory. *Neuron*, 103(1):147–160.e8, July 2019.

[33] Nicolas Brunel and Xiao-Jing Wang. Effects of Neuromodulation in a Cortical Network Model of Object Working Memory Dominated by Recurrent Inhibition. *Journal of Computational Neuroscience*, 11(1):63–85, July 2001.

[34] Tim P. Vogels and L. F. Abbott. Gating multiple signals through detailed balance of excitation and inhibition in spiking networks. *Nature Neuroscience*, 12(4):483–491, April 2009.

[35] Razvan Pascanu and Herbert Jaeger. A neurodynamical model for working memory. *Neural Networks*, 24(2):199–207, March 2011.

[36] Georg Fette and Julian Eggert. Short Term Memory and Pattern Matching with Simple Echo State Networks. In Włodzisław Duch, Janusz Kacprzyk, Erkki Oja, and Sławomir Zadrożny, editors, *Artificial Neural Networks: Biological Inspirations – ICANN 2005*, pages 13–18, Berlin, Heidelberg, 2005. Springer.

[37] Yi Chen, Hanwen Liu, Kexin Shi, Malu Zhang, and Hong Qu. Spiking neural network with working memory can integrate and rectify spatiotemporal features. *Frontiers in Neuroscience*, 17, June 2023.

[38] Ramon Bartolo and Bruno B. Averbeck. Prefrontal Cortex Predicts State Switches during Reversal Learning. *Neuron*, 106(6):1044–1054.e4, June 2020.

[39] Chiara Baston and Mauro Ursino. A Biologically Inspired Computational Model of Basal Ganglia in Action Selection. *Computational Intelligence and Neuroscience*, 2015:e187417, November 2015.

[40] Ashwin Viswanathan Kannan, Goutam Mylavarapu, and Johnson P. Thomas. Unsupervised Spiking Neural Network Model of Prefrontal Cortex to study Task Switching with Synaptic deficiency, May 2023.

[41] M. J. Frank, B. Loughry, and R. C. O'Reilly. Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience*, 1(2):137–160, June 2001.

[42] Feifei Zhao, Yi Zeng, and Bo Xu. A Brain-Inspired Decision-Making Spiking Neural Network and Its Application in Unmanned Aerial Vehicle. *Frontiers in Neurorobotics*, 12, September 2018.

[43] Seth A. Herd, Randall C. O'Reilly, Tom E. Hazy, Christopher H. Chatham, Angela M. Brant, and Naomi P. Friedman. A neural network model of individual differences in task switching abilities. *Neuropsychologia*, 62:375–389, September 2014.

[44] Sam Carroll, Krešimir Josić, and Zachary P. Kilpatrick. Encoding certainty in bump attractors. *Journal of Computational Neuroscience*, 37(1):29–48, August 2014.

[45] Jose M. Esnaola-Acebes, Alex Roxin, and Klaus Wimmer. Bump attractor dynamics underlying stimulus integration in perceptual estimation tasks, March 2021.

[46] Wolfram Schultz, Peter Dayan, and P. Read Montague. A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599, March 1997.

[47] Philippe N. Tobler, Christopher D. Fiorillo, and Wolfram Schultz. Adaptive Coding of Reward Value by Dopamine Neurons. *Science*, 307(5715):1642–1645, March 2005.

[48] John N.J Reynolds and Jeffery R Wickens. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, 15(4-6):507–521, June 2002.

[49] Mojtaba Madadi Asl, Abdol-Hossein Vahabie, and Alireza Valizadeh. Dopaminergic Modulation of Synaptic Plasticity, Its Role in Neuropsychiatric Disorders, and Its Computational Modeling. *Basic and Clinical Neuroscience*, 10(1):1–12, 2019.

[50] Chung-Hay Luk and Jonathan D. Wallis. Choice Coding in Frontal Cortex during Stimulus-Guided or Action-Guided Decision-Making. *Journal of Neuroscience*, 33(5):1864–1871, January 2013.

[51] Steven W. Kennerley and Mark E. Walton. Decision Making and Reward in Frontal Cortex. *Behavioral Neuroscience*, 125(3):297–317, June 2011.

[52] Shintaro Funahashi. Prefrontal Contribution to Decision-Making under Free-Choice Conditions. *Frontiers in Neuroscience*, 11, July 2017.

[53] Encarni Marcos and Aldo Genovesio. Determining Monkey Free Choice Long before the Choice Is Made: The Principal Role of Prefrontal Neurons Involved in Both Decision and Motor Processes. *Frontiers in Neural Circuits*, 10, September 2016.

[54] Zuzanna Z. Balewski, Thomas W. Elston, Eric B. Knudsen, and Joni D. Wallis. Value dynamics affect choice preparation during decision-making. *Nature neuroscience*, 26(9):1575–1583, September 2023.

[55] Lars Bäckman, Lars Nyberg, Anna Soveri, Jarkko Johansson, Micael Andersson, Erika Dahlin, Anna S. Neely, Jere Virta, Matti Laine, and Juha O. Rinne. Effects of Working-Memory Training on Striatal Dopamine Release. *Science*, 333(6043):718–718, August 2011.

[56] Pierre Enel, Joni D Wallis, and Erin L Rich. Stable and dynamic representations of value in the prefrontal cortex. *eLife*, 9:e54313, July 2020.

[57] Mark Steyvers, Michael D. Lee, and Eric-Jan Wagenmakers. A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, 53(3):168–179, June 2009.

[58] Eric Schulz, Nicholas T. Franklin, and Samuel J. Gershman. Finding structure in multi-armed bandits. *Cognitive Psychology*, 119:101261, June 2020.

[59] Shunan Zhang and Angela J Yu. Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[60] Roshan Cools. Chemistry of the Adaptive Mind: Lessons from Dopamine. *Neuron*, 104(1):113–131, October 2019.

5 Appendix

5.1 Activation function

The function Φ. is defined by combining a generalized version of the sigmoid, namely with a gain $\beta \neq 1$ and offset $\alpha \neq 0$, and a Gaussian with mean μ and variance σ. Their contributions are weighted by as r and $1 - r$ ($r \in (0, 1)$) respectively.

$$\Phi_v(x) = r\left(1 + \exp^{-\beta(x-\alpha)}\right)^{-1} + (1-r)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The motivation behind this choice is to express a function that possesses a bounded region (depending on μ, σ) at a high/low peak (depeding on the value of γ_2), and a continuous transition to a constant value (depending on the steepness of the sigmoid β, shift α, and intensity γ_1).

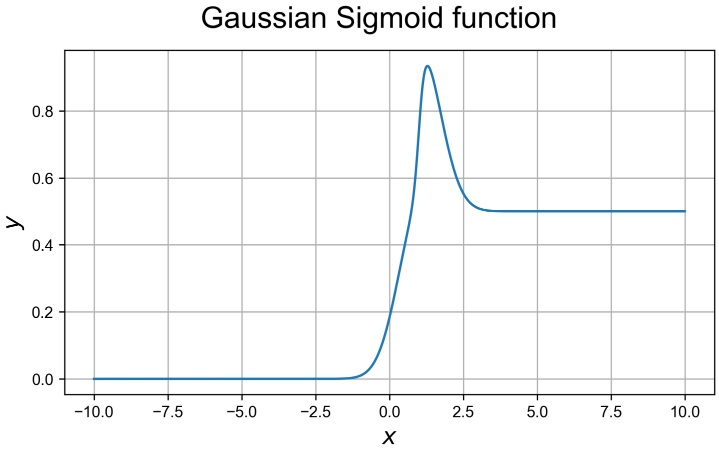

Figure 5: Activation function Φ_v - *Parameters* $\beta = 10$, $\alpha = 1$, $\mu = 1$, $\sigma = 1$, *and* $r = 0.5$.

5.2 Zero-steps distribution shift

In this first setting, as the end of a trial i the arm distribution changes immediately to a new one $i + 1$ as $\pi_i \rightarrow \pi_{i+1}$.
From figure 6 above, it is clear the ability of the model (in red) to reach almost always the optimal reward policy (*i.e.* the greediest) for all trials, even after the distribution shift. In comparison, the other algorithms start to struggle when the arms are more 100 and the distribution changes.

Next, it has been enquired how the model selection policy evolves over time and in comparison with the other algorithms, as visualized in figure 7.

The principal distinction is the model's strictly greedy behaviour once a good arm is found. Only in the case of a meaningful decrease in reward the exploration is resumed, in contrast with the other approaches in which occasional sub-optimal choices are made.

5.3 Epsilon-steps distribution shift

In the setting with a smooth distribution shift the difficulty of the problem is increased, especially since short-sighted greedy behaviours are easily sub-optimal. The model (always in red) is capable of reaching and maintaining a successful profile, even with many arms available.

5.4 Table of results

Table 1: Performance comparison for $K = 5$

Model	**KAB-0**	**KAB-ϵ**	**KAB-**sin
Optimal	0.900	0.881	0.563
Random	0.330	0.337	0.200
Thompson	0.905	0.617	0.317
ϵ-Greedy	0.797	0.531	0.315
UCB	0.897	0.656	0.319
Model	**0.899**	**0.663**	**0.265**

Table 2: Performance comparison for $K = 10$

Model	**KAB-0**	**KAB-ϵ**	**KAB-**sin
Optimal	0.900	0.885	0.355
Random	0.247	0.250	0.100
Thompson	0.896	0.648	0.339
ϵ-Greedy	0.611	0.597	0.343
UCB	0.891	0.655	0.358
Model	**0.905**	**0.668**	**0.203**

Table 3: Performance comparison for $K = 100$

Model	**KAB-0**	**KAB-ϵ**	**KAB-**sin
Optimal	0.900	0.883	0.020
Random	0.196	0.201	0.010
Thompson	0.894	0.586	0.013
ϵ-Greedy	0.519	0.574	0.018
UCB	0.853	0.572	0.012
Model	**0.898**	**0.651**	**0.010**

Table 4: Performance comparison for $K = 100$

Model	**KAB-0**	**KAB-ϵ**	**KAB-**sin
Optimal	0.900	0.885	0.010
Random	0.178	0.176	0.005
Thompson	0.875	0.624	0.006
ϵ-Greedy	0.679	0.588	0.010
UCB	0.792	0.510	0.006
Model	**0.905**	**0.610**	**0.006**

Table 5: Performance comparison for $K = 100$

Model	**KAB-0**	**KAB-ϵ**	**KAB-**sin
Optimal	0.900	0.880	0.002
Random	0.177	0.178	0.001
Thompson	0.779	0.445	0.001
ϵ-Greedy	0.386	0.478	0.002
UCB	0.301	0.185	0.001
Model	**0.703**	**0.480**	**0.001**

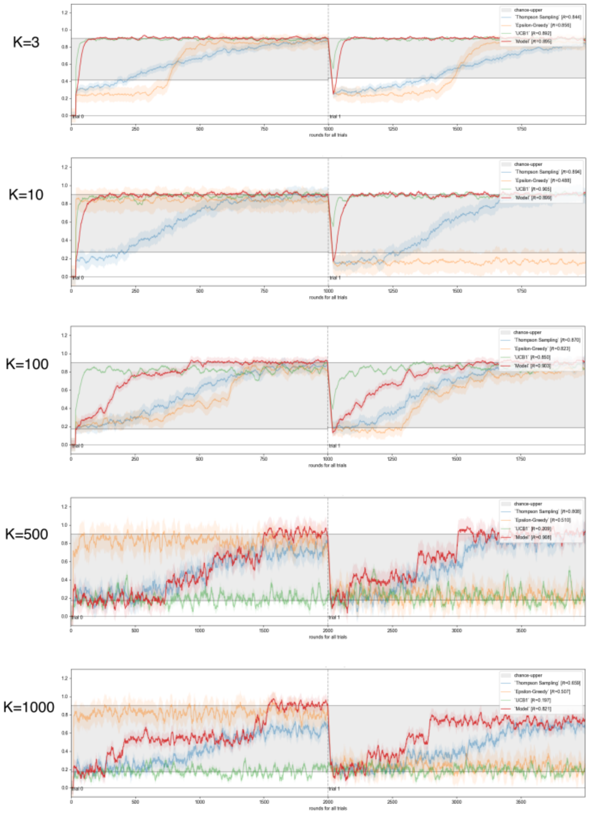

Figure 6: Performance with variable number of arms - *each plot is a simulation with K numbers of arms, the x-axis are rounds, the central vertical line signals the start of the second trial, the y-axis is the reward fraction. The shaded area is the reasonable reward range, where the lower bound is the chance level and the upper bound the best reward (following the optimal policy). The model performance is in red, while Upper-Confidence Bound green, Thompson Sampling blue, and Epsilon-Greedy orange.*

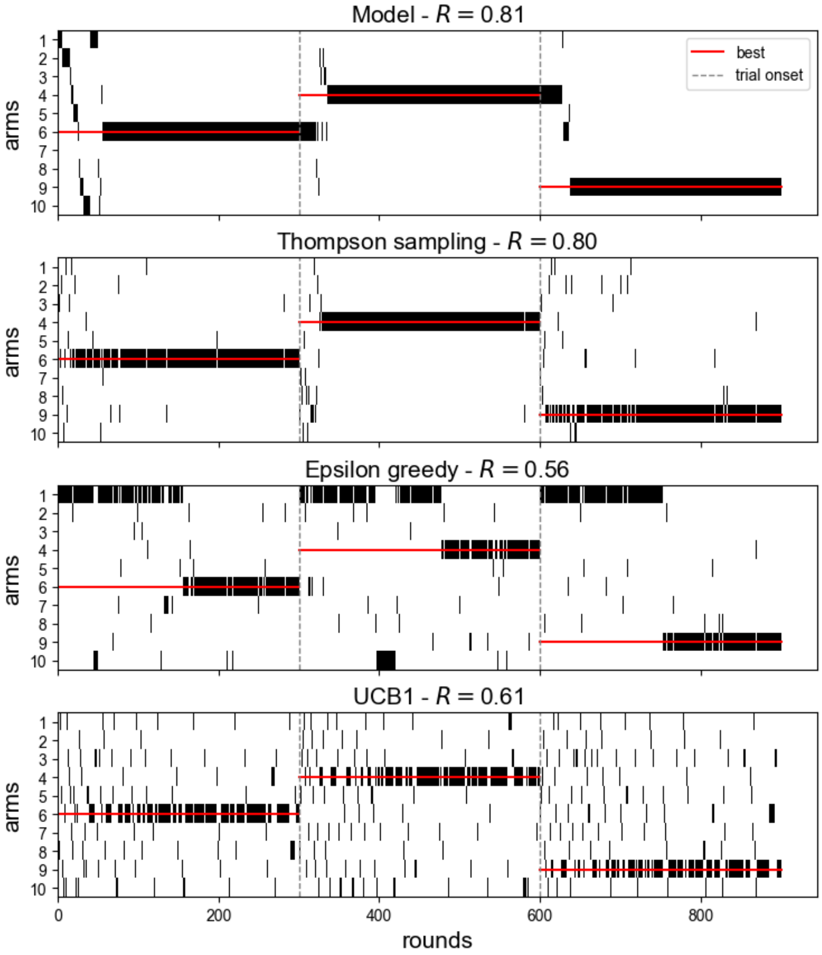

Figure 7: Selection evolution over rounds for multiple models - *the individual plots follow the same schema of 2, with the model name and reward per round fraction*

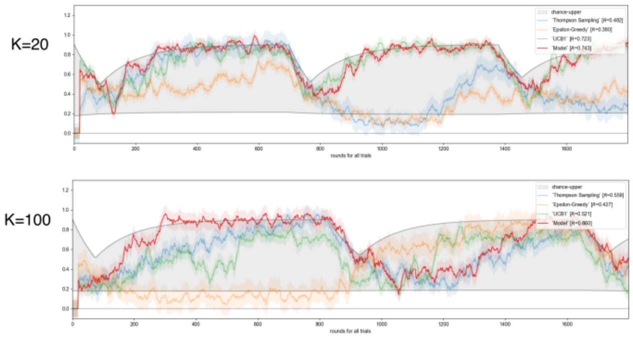

Figure 8: Performance with variable number of arms - *each plot is a simulation with K-numbers of arms, and the rest is also the same as before in 6.* Each trial has 3 rounds, meaning that every three steps the distribution change.