A bio-inspired minimal model for non-stationary K-armed bandits

Krubeal Danieli, Mikkel Elle Lepperød

December 2, 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]. These cortical projections 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].

Considering decision-making, simple and well-studied ecological settings are foraging tasks, such as food search. In this problems, the agent is usually asked to choose between different options to maximize the expected reward. In nature, animals have been shown to exhibit different strategies depending on context. *matching behaviour* is a well-known phenomenon in which the animal's decision patterns are proportional to the reward probability of the available options. This behaviour is thought to result from the trade-off between exploration and exploitation [14, 15]. This is a well known phenomenon in the reinforcement learning literature, in which the agent is faced with the dilemma of exploring new alternatives, potentially more rewarding, or exploiting known options, although possibly sup-optimal. Other behaviours dependant on contexts 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]. When considering more computational approaches in the study of choice behaviour, a popular formalization of such tasks is the *multi-armed bandit problem* (MABP) [18]. This settins is usually described in terms of a slot machine endowed with K distinct levers. During a round, the agent selects one of the levers and collects a reward R according to an unknown reward probability specific to the chosen lever. The goal is simply to maximizes the total reward after a given number steps, which is achieved by effectively updating a selection policy after each round. 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].

1.1 Related work

There exists various flavours of this problem, with the simplest having a stationary reward distributione.

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 [19, 20]. This approach relies on Bayesian optimization, where the goal is to maintain a posterior distribution over the reward probabilities of the actions, and selecting actions accordingly. Another popular algorithm is Upper Confidence Bound (UCB), which has been shown to achieve near-optimal regret bounds in the adversarial setting [21]. The approach is based on the idea of maintaining an upper limit on the reward probabilities of the actions, and selecting actions accordingly. Other successful algorithms are ϵ-greedy and VDBE [22, **?**, 23, 24, 25].

Despite the success of these algorithms in solving the k-armed bandit problem, they lack biological plausibility. The brain has evolved a complex network of interconnected regions that work together to solve this task.

In this work, we focuses on stochastic bandit problems, more challenging variant of the original task endowed with *concept drift*, where the reward distribution changes over time [26, 27, 28].

We propose a biologically plausible model using rate neurons, obtaining good performance even with settings with a large amount of arms. Its architecture is composed of two connected neuronal layers, both with as many neurons as the arm of the bandit task. The first layer is inspired by the functionality of the OFC, and its scope is to maintain an active representation of the arms weighted by the input from the second layer. The second, modeled after the ACC, is meant to represent the value of the arms, and its input connections are updated through a learning rule dependant on the reward history and current connectivity pattern.

Our models features two important aspects of brain decision making. Firstly, the decision making process itself is implemented as a dynamical interaction between neural populations, similarly to bump attractor networks for perceptual computations [29, 30]. The final choice of the arm is achieved by the agreement or disagreement between the two populations, and it depends on their underlying value representation [16, 17]. Secondly, plasticity is based on a non-associative learning rule, endowed with a non-linear kernel for the weight update quantity. Importantly, the specific values of a parametrization of the kernel has been optimized through an evolutionary algorithm. Behind this design choice there is our hypothesis that the scale of neural synaptic updates should vary according to its magnitudeand with a non-linear shape of its scaling function represent a better choice that a constant one. This considerations align with the idea that the brain is able to adapt its learning rate according to the context, and that the learning rate is not constant across neurons within the same network. This approach has been already adopted in several computational architectures, for instance using spiking neurons [31] and synaptic metaplasticity [32]. Lastly, there is experimental evidence that this function of adaptive plasticity might be covered by dopamine [33]. Indeed, it is role in prediction error and reward signaling is well established [34], together with its importance in modulating processes in the PFC [35, 36, 8].

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 $\{1 \dots K\}$ levers (or arms), with an associated reward distribution $\mathbf{p} = \{p_1, \dots p_K\}$. 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, 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 $\mathbf{u}$.

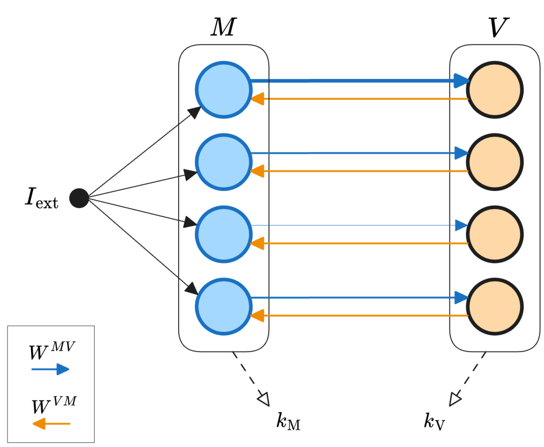

Figure 1: Model architecture - *The model is composed of a layer M (blue), receiving a feedfoward input I_{ext}, a layer V (orange), and connections $\boldsymbol{W}^{MV}$ and $\boldsymbol{W}^{VM}$. Additionally, two indexes k_M, k_V can be extracted from the layers and corresponds to the selection made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

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.

Algorithm 1: Two-phases option selection process

Input: External input $\mathbf{I}_{\text{ext}}$, population $\mathbf{u}$, population $\mathbf{v}$, 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

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 [37, 38]. 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 [39, 40, 41]. 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 [42, 43].

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.

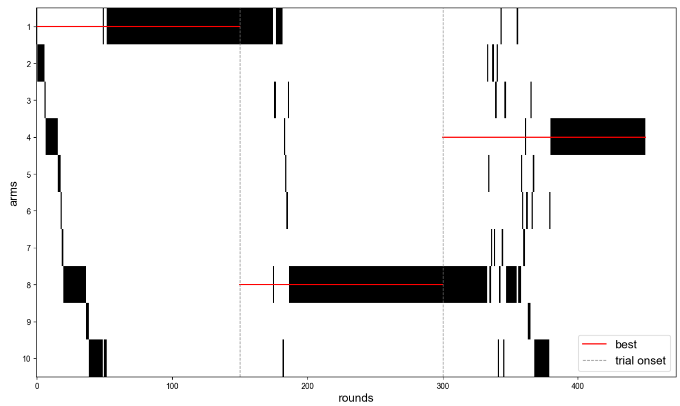

Figure 2: SELECTION EVOLUTION OVER ROUNDS - *the x-axis represents the available arms, while the y-axis the number of rounds, with the dotted vertical lines indicating the start of a new trial with 150 rounds each. The model selections are the black vertical lines for an arm and a round. The red horizontal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

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 an environment with food (reward) is distributed over a set of fixed locations, but whose occurrence probability 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 $\mathbf{p}_i \rightarrow \mathbf{p}_{i+1}$.

Epsilon-steps distribution shift [KAB-ϵ]: the reward distribution $\mathbf{p}$ changes gradually over rounds, tracked as time t, such that its shape tends towards a target distribution $\mathbf{q}_i$ as $\tau_p \dot{\mathbf{p}}_t = \mathbf{q}_i - \mathbf{p}_t$. Here, $\dot{\mathbf{p}}$ is the time derivative of the distribution and τ_p is its time constant. Once distance is below a threshold ϵ as $|\mathbf{q}_i - \mathbf{p}_t| < \epsilon$, the target distribution is changed to a new one $\mathbf{q}_i \rightarrow \mathbf{q}_{i+1}$.

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 $\mathbf{p}_t = \{\sin(2\pi f_k t)$ for $k = 1 \ldots K\}$ and it is normalized as $\mathbf{p}_t = \mathbf{p}_t(\sum_k \mathbf{p}_{t,i})^{-1}$ 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

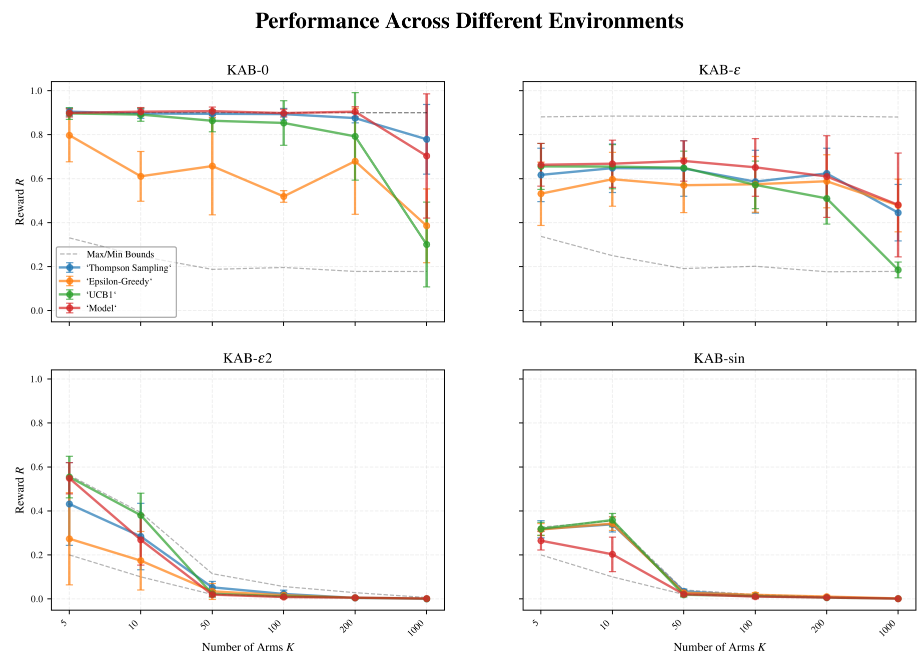

Figure 3: PERFORMANCE COMPARISON FOR DIFFERENT VALUES OF K AND GAME VARIANTS *The model is compared with Thompson Sampling, Epsilon-Greedy, and UCB. The performance is measured as the average reward obtained by the agent over a number of trials.*

3.3 Decision-making dynamics

3.3.1 Entropy analysis

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.

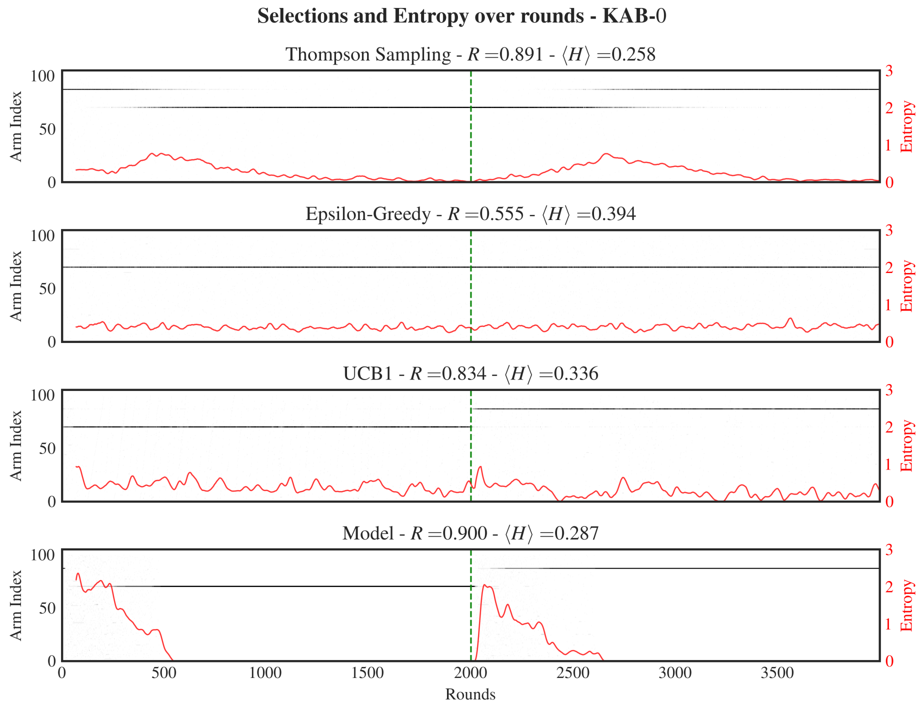

Figure 4: Decision-making dynamics for different models *Each plot display the results from one model obtained from average over 20 iterations. The raster plots (black dots) show arm selected at each round. The red lines represent the entropy level, calculated from the distribution of selections over the preceding 20 rounds; the line is then smoothed with a 30-steps moving average. In the plot titles, the total reward and average entropy over all trials are also reported.*

3.3.2 Weight update dynamics

Next, we analyzed the weight update dynamics of the model over the rounds. In figure 5, we plotted the evolution of the weights for each arm over time, averaged over 20 simulations and smoothed over 30 rounds. The results show that the model is able to quickly adapt to new reward distributions. It is also able to maintain the optimal policy over time, with the weights remaining approximately stable. The update quantity ΔW_k^{MV} changes sign according to the collected reward, with its magnitude being higher at the beginning of the trials. Initially, the sign is mostly positive (potentiation) since the weights start at zero, and after some uncertainty a consistently preferred arm emerges. However, when the reward distribution switches a regular series of sub-optimal choices is made, leading to zero reward. This causes an accumulation of weight updates with negative sign (depression), eventually bringing the value of the preferred arm to drop. In the meantime, other options are probed until another the choices converge to another arm, promoted by a trail of positive weight updates.

This behaviour is consistent with the low entropy levels observed in the previous analysis.

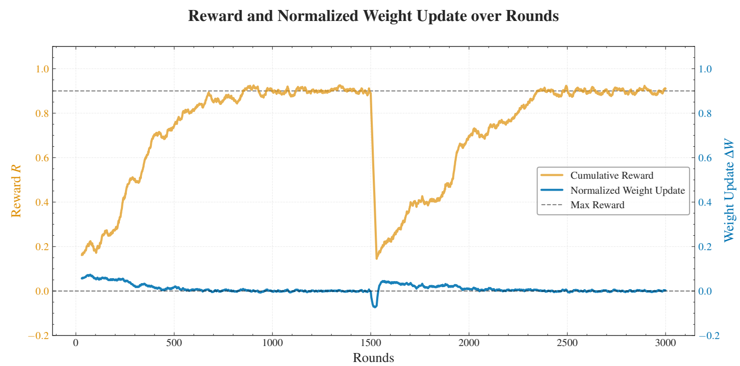

Figure 5: WEIGHT UPDATE DEVELOPEMENT FOR THE MODEL *The plot displays the weight update quantity* ΔW_k^{MV} *for each round (blue line), smoothed as a 20-steps moving average. It is also reported the average reward in a window of 30 rounds (golden line). The results have been obtained averaging over 20 iterations.*

3.3.3 Robustness

Then, we sought to investigate the robustness of the model. This was accomplished by evaluating the performances in a stationary setting with $K = 10$ and increasing levels of entropy in the reward distribution. For the details of calculation of the reward distribution see the appendix 5.2.

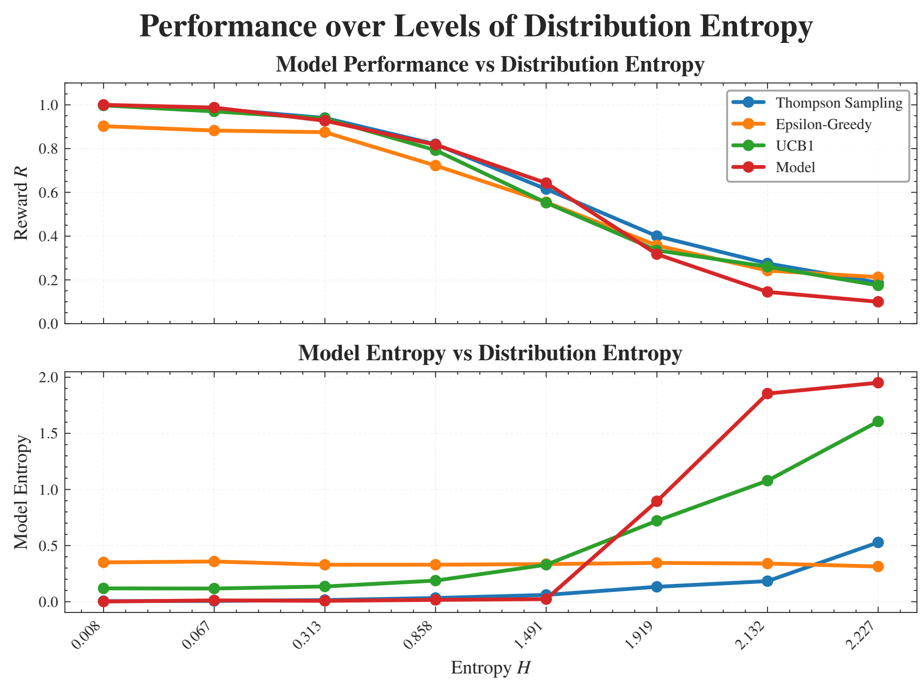

Figure 6: Entropy analysis for the model in a stationary setting
Top plot: *average reward obtained by each model for for increasing levels of entropy in the reward distribution.* Bottom plot: *average entropy of the selections (as calculated in 3.3.1) for increasing levels of entropy in the reward distribution.*

4 Discussion

The process of making decision in uncertain settings is a remarkable aspect of cognition. For instance, such behaviour is implemented in animals during foraging and matching behaviour. In the context of humans, it has been observed that the pool of adopted policies vary considerably [44]. 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 choices [45, 46]. A useful formalization of such tasks is the multi-armed bandit problem, which has been extensively studied in the context of reinforcement learning [14]. Although several algorithms have been proposed to solve the problem with robust theoretical guarentees, there is a general lack of biological plausibility of the architecture and dynamics.

In this work, we introduced a rate neural networks to address the binomial K-armed bandit problem in a non-stationary environment. The results obtained from our model show that it is able to effectively adapt to changing reward distributions and maintain a greedy policy over time, on par with the standard algorithms.

The observed efficiency of our model can be attributed to that attractor dynamics and a non-associative synaptic plasticity rule.

These architectural choices were inspired by real-world decision-making processes observed in humans and other animals.

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] 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.

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

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

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

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

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

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

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

[31] Jeffrey B. Inglis, Vivian V. Valentin, and F. Gregory Ashby. Modulation of Dopamine for Adaptive Learning: A Neurocomputational Model. *Computational brain & behavior*, 4(1):34–52, March 2021.

[32] Kiyohito Iigaya. Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system. *eLife*, 5:e18073, August 2016.

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

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

[35] Danila Di Domenico and Lisa Mapelli. Dopaminergic Modulation of Prefrontal Cortex Inhibition. *Biomedicines*, 11(5):1276, May 2023.

[36] Sweyta Lohani, Adria K. Martig, Karl Deisseroth, Ilana B. Witten, and Bita Moghaddam. Dopamine Modulation of Prefrontal Cortex Activity Is Manifold and Operates at Multiple Temporal and Spatial Scales. *Cell Reports*, 27(1):99–114.e6, April 2019.

[37] 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.

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

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

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

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

[46] 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.

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).

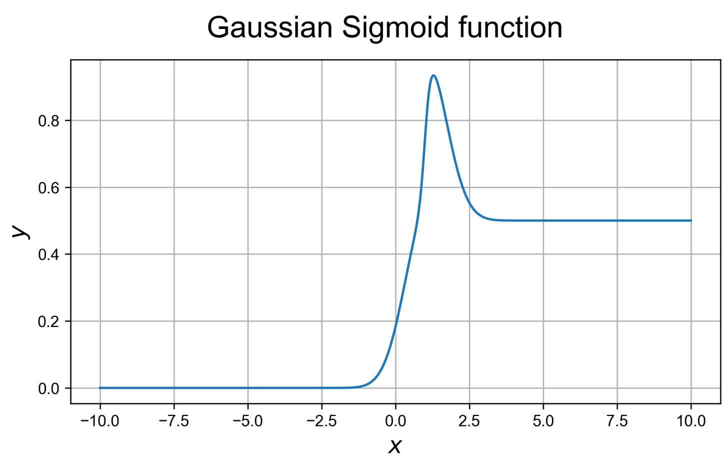

Figure 7: Activation function Φ_v - *Parameters* $\beta = 10$, $\alpha = 1$, $\mu = 1$, $\sigma = 1$, *and* $r = 0.5$.

5.2 Reward distribution entropy

The calculation of a set of N reward probability distribution $\mathbf{p}_i$ for $i \dots N$ for K values with a progressively decreasing levels of entropy $\mathbf{h}_i$ for $i \dots N$ has been obtained by the following algorithm:

Algorithm 2: Reward Probability Distribution Generation

Input: Number of distributions N, dimension K
Output: Set of probability distributions $\mathbf{p}_i$ with decreasing entropy
Initial Setup: Define set $B = \{1.5^x \mid x = 1, \ldots, 7\}$;
for $i \leftarrow 1$ *to* N **do**
  $\mathbf{z} \leftarrow \text{RandomVector}(0,1)^K$;
  $j \leftarrow \text{RandomIndex}(K)$;
  $\mathbf{z}_j \leftarrow 1$;
  $\beta_i \leftarrow$ Sample index=i from (B) ; `// Sample temperature from` B
  $\mathbf{p}_i \leftarrow \frac{\exp(\beta_i \mathbf{z})}{\sum_j \exp(\beta_i \mathbf{z}_j)}$; `// Softmax with temperature`
end
return $\mathbf{p}_i$

5.3 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**