# A biologically plausible decision-making model based on interacting cortical columns

E. Baspinar[1], G. Cecchini[2,3], M. DePass[3], M. Andujar[4],
P. Pani[4], S. Ferraina[4], R. Moreno-Bote[3,5], I.Cos[2,5], A. Destexhe[1]

[1] CNRS, NeuroPSI, Saclay, France
[2]Facultat de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, Catalonia, Spain
[3]Center for Brain and Cognition, DTIC, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain
[4]Department of Physiology and Pharmacology, Sapienza University of Rome, Rome, Italy
[5]Serra-Hunter Fellow Programme, Barcelona, Catalonia, Spain

## Abstract

We present a new AdEx mean-field framework to model two networks of excitatory and inhibitory neurons, representing two cortical columns, and interconnected with excitatory connections contacting both Regularly Spiking (excitatory) and Fast Spiking (inhibitory) cells. This connection scheme is biophysically plausible since it is based on intercolumnar excitation and intracolumnar excitation-inhibition. This configuration introduces bicolumnar competition, sufficient for choosing between two alternatives. Each column represents a pool of neurons voting for one of two choices indicated by two stimuli presented on a monitor in human and macaque experiments. The task also requires maximizing the cumulative reward over each episode, which consists of a certain number of trials. The cumulative reward depends on the coherency between choices of the participant/model and preset strategy in the experiment. We endow the model with a reward-driven learning mechanism allowing to capture the implemented strategy, as well as to model individual exploratory behavior. We compare the simulation results to the behavioral data obtained from the human and macaque experiments in terms of performance and reaction time. This model provides a biophysical ground for simpler phenomenological models proposed for similar decision-making tasks and can be applied to neurophysiological data obtained from the macaque brain. Finally, it can be embedded in whole-brain simulators, such as The Virtual Brain (TVB), to study decision-making in terms of large scale brain dynamics.

## 1 Introduction

Decision making is crucial for survival in many species. It refers to selecting between given alternatives by taking into account the consequences of each choice. To produce a cognitive behavior, the brain employs several decision-making mechanisms which encode and interpret sensory stimuli, weight evidence to select between choice alternatives, and finally, generates oriented motor actions.

There have been proposed many experiment procedures by neuroscientists to unveil the neural mechanisms underlying decision-making processes. On the side of psychophysics, behavioral data is sampled while the participant performs decision-making tasks, such as distinguishing between different stimuli and reaching to one of them based on a certain strategy [1, 2]. On the neurophysiology side, electrophysiological measurements provide single neuron or neuronal population activity in specific areas of the brain. This can be used to study the link between those areas, functions, and the behavioral variables in the performed tasks [3–7].

The prefrontal cortex has been traditionally considered as a key area involved in decision-making in the brain [8–12]. In addition to the prefrontal cortex, new studies have identified other cortical and subcortical structures participating in decision-making processes [13–16]. Those regions can be considered based on two functional categories: benefit and cost.

Our everyday routine is based on analyzing, judging and evaluating the advantages and disadvantages of the choices under different circumstances. During all those processes, the brain areas responsible for benefit

and cost estimation are activated [17–19]. Consequently, a decision is made either to perform a task or to avoid it. We will take into account these two mechanisms as the main ingredients training our model to identify the good decision.

Several neural population models have been introduced so far for similar decision-making mechanisms [20–22]. In [23], a recurrent network model of object working memory was proposed. Thereafter another network model with similar architecture was proposed for a visual discrimination task in [24]. This final architecture was adapted to a two-choice perceptual task, where the effect of memory based experience on decisions was included [2]. All those previous models are based on connectivity architectures which are not in accordance with the biological framework, in which long range cortico-cortical connections originate from excitatory pyramidal cells [25, 26], and make the synapses onto other excitatory pyramidial cells as well as onto the inhibitory interneuron cells [27, 28]. To tackle this point, we consider the model topology presented in [2] as the starting point of our model architecture, and modify it such that the aforementioned connectivity observed in the biological framework is taken into account.

Reliable neural responses to stimuli are generally observed at the population level in neurophysiological experiments. In other words, the information processing and the produced response do not correspond to a single neuron but rather to a population activity. Neural response is obtained as averaged time-integrated activity of individual dynamics of the neurons in interaction within the population. A neuronal population can be modeled as a stochastic network system, which consists of a number of stochastic differential equations. Averaged network behavior models the population behavior. This requires high computational power and it is challenging to apply analytical approaches since the network is high dimensional [29]. Another approach is to consider the averaged network behavior at the coarse-grained continuum limit where the number of neurons in the network is assumed to be infinity. In this way, the asymptotic limit of the network can be written in terms of the probability distribution of the state variables appearing in a single neuron in the network. This asymptotic limit is the so-called *mean-field limit* [30–32].

Adaptive Exponential (AdEx) mean-field framework [33, 34] approximates closely neuronal population behavior modeled by AdEx network. In the case of cerebral cortex, AdEx networks are used to model two cell types: Regular Spiking (RS) neurons, displaying spike-frequency adaptation as observed in excitatory pyramidal neurons, and Fast Spiking (FS) neurons, with no adaptation, as observed in inhibitory interneurons. AdEx mean-field models are low dimensional, simpler and easier to analyze compared to AdEx networks, yet they approximate closely the network dynamics, motivating our choice of model.

We present a novel biophysically inspired AdEx mean-field model for decision-making with a reward mechanism by starting from the frameworks presented in [2, 24]. The novelties of the model are in terms of both connectivity and structure. The model reproduces the behavioral results obtained from the experiments conducted on humans and macaques. The implementation of the model is provided in Python as a Jupyter notebook [35].

There are several challenges both at mathematical and numerical levels. Firstly, the model dimension is increased and the intercolumnar connections introduce the derivatives of cross correlations between population firing rates of different columns. This demands higher memory and computational power. Secondly, an adequate plasticity mechanism is to be designed so that the model can learn the preset decision making strategy of the task. Finally, the increased number of parameters makes it difficult to fit them to the experimental behavior data.

In Section 2, we explain the experiment protocols applied on human and macaque participants. In Section 3, we present our model. This is followed in Section 4 by the noise sources of the model. We provide in Section 5 some explanations regarding the quantification of the behavioral data and the effects of some parameters on the model behavior. We present the simulation results and their comparison to the biological data in Section 6. Finally, we conclude by summarizing the novelties and pointing towards the future perspectives in Section 7.

# 2 Experiment setup

## 2.1 Human experiment

Decisions are made by taking into account immediate and longer term consequences. In many cases, making decision with a consideration of stronger benefits in the long term is important for survival. This requires

a control over future planning to identify the optimal decision making strategy. To study the behavioral background of such mechanisms in humans, we designed an experiment with two versions named Horizon 0 and Horizon 1. Each experiment is composed of $K$ episodes. Here $K$ denotes the total number of episodes in the experiment. Each episode is composed of 1 or 2 trials in Horizon 0 and Horizon 1, respectively. Behavioral results are recorded in terms of several measures. Here we use performance and reaction times as measures. The experiments were conducted at Universitat de Barcelona, Facultat de Matemàtiques i Informàtica. The experiments were performed by 28 participants. We apply our model to one participant as an example, however the model can be applied to all participants. The participant which we considered is a 20-year-old healthy female. We provide a summary of the experiment. For details, we refer to [36] .

In Horizon 0, one episode has one single trial. In Horizon 1, one episode is composed of two successive trials, which are named as Trial 1 and Trial 2. In Horizon 1, as illustrated in Figure 1a, the participant is told to choose one of the stimuli in each trial. Each stimulus is a partially filled vertical bar. The percentage of the filled part is different for each bar. We increase the filled parts of both bars with the same amount at the end of Trial 1 if the choice of the participant is in coherence with the preset strategy in the task. Otherwise, we decrease the filled parts with the same amount. This amount is called gain. It is generated randomly at the beginning of the experiment and kept fixed throughout the whole experiment. The preset strategy in Horizon 1 experiments is to choose the smaller stimulus in Trial 1 and the larger stimulus in Trial 2. This boils down to choosing the larger stimulus in each episode in Horizon 0.

Cumulative reward is the sum of the filled parts of the chosen stimuli throughout the trials of one episode. The goal of the participants is to maximize the cumulative reward in every episode. This requires: (i) to identify the correspondence of reward to the made choices, (ii) to capture the preset strategy to optimize the made choices in accordance with the strategy.

The amounts of the filled parts of the bars are $M \pm d/2$. Here $M$ is randomly generated from a uniform distribution at the beginning of each episode. Five different values of $d$ are used in the experiments. Each one of them is used for $K/5$ episodes of the experiment. In each episode, $d$ is fixed. Here $M$ denotes the mean value of the stimuli, and $d$ determines the perceptual difficulty to distinguish between two stimuli. In Horizon 1, if the participant chooses the small stimulus in Trial 1, the gain $G$ is added to both stimuli and $M + G \pm d/2$ become the filled quantities of the two stimuli in Trial 2. Otherwise, the gain $G$ is subtracted from the stimuli and the filled quantities become $M - G \pm d/2$ in Trial 2. The same procedure is applied in Trial 2 but now the expected choice by the preset strategy is the large stimulus. Moreover, the participant does not see at the end of Trial 2 any added or subtracted gain since there is no Trial 3. This forces the participant to learn the consequence of the episode instead of learning a sequence of choices, because there is no feedback at the end of the episode. See Figure 1a for an example scenario of a Horizon 1 episode and Figure 2 for some relevant example experiment results.

## 2.2 Macaque experiment

The mechanism behind resisting immediate rewards in consideration of larger rewards in the long term was investigated in macaque. For this, a control over immediate reward gain is essential to identify the optimal decision making strategy. We designed a similar task to the human experiment so as to study the behavioral background of such control mechanisms in macaques. One 16-year-old, healthy, male macaque weighting 9.5–10.5 kg was trained for performing a task composed of two successive trials combined in one episode. This accounts to Horizon 1 complexity. The trials are named as Trial 1 and Trial 2. The amount of reward (water drops) received in Trial 2 depends on the choice made in Trial 1 by the participant. The behavioral data was recorded in two forms from the human experiment: performance and reaction times. The macaque experiments were conducted in Sapienza Università di Roma, Dipartimento di Fisiologia e Farmacologia.

A dataset of black and white, 180x180 pixels stimuli were shown to the participant in each trial. Six stimuli (peripheral targets; PTs) were extracted randomly from *Microsoft PowerPoint* shapes library and adjusted such that its black/white ratio remains constant during the experiment. Two of those six randomly extracted PTs were shown in Trial 1 and they were attributed to different rewards: PT 1 – 3 drops, PT 2 – 2 drops. The remaining four stimuli were separated into two groups of two PTs, Group 1 with PT 3 and PT 4, and Group 2 with PT 5 and PT 6. Each of those four PTs corresponds to a different reward: PT 3 – 1 drop, PT 4 – 0 drop, PT 5 – 4 drops and PT 6 – 6 drops. The group which is shown in Trial 2 was determined by the choice made in Trial 1. More precisely, if the participant chooses PT 1 in Trial 1, Group 1 will be shown

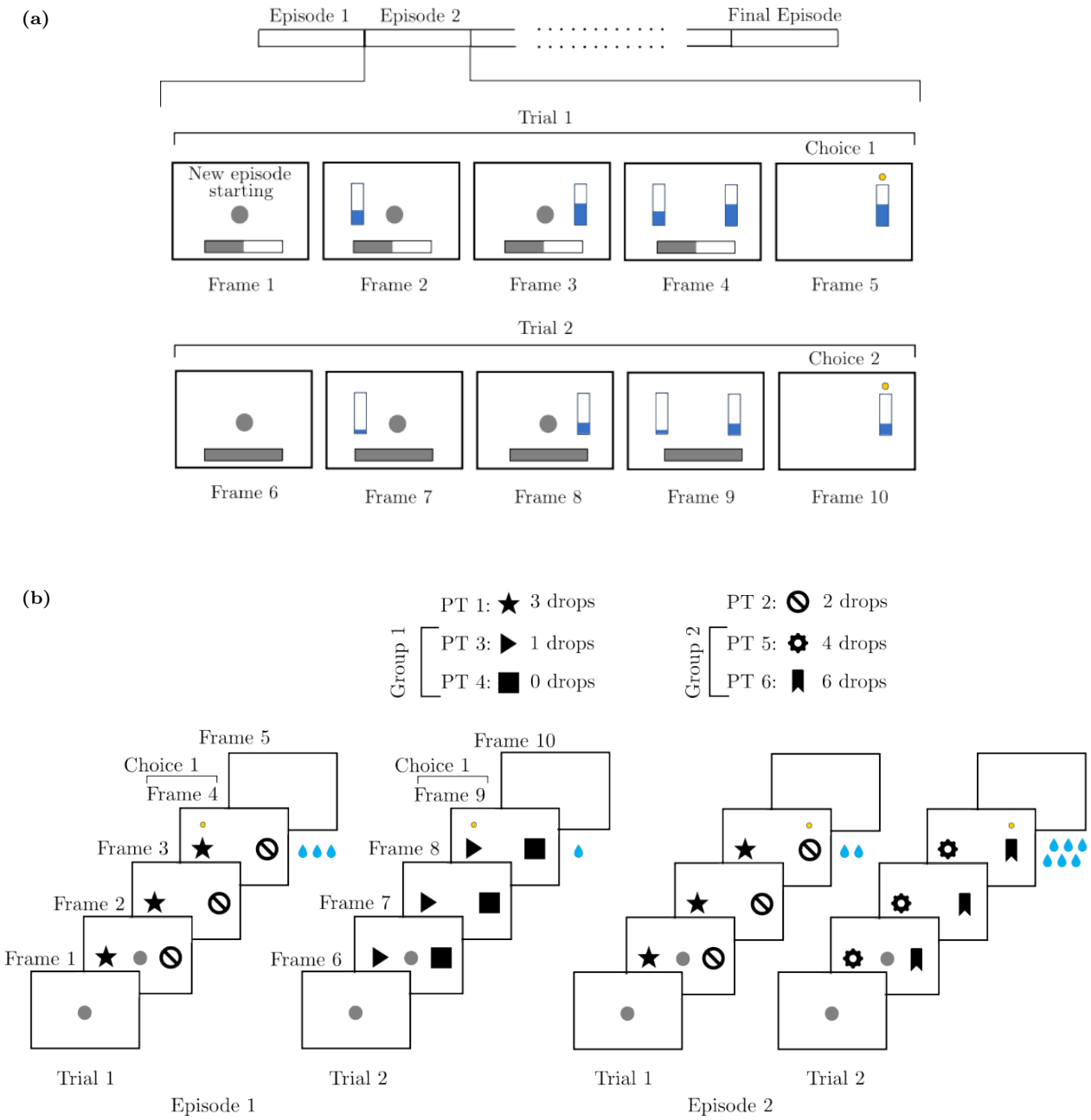


**Figure 1:** Example scenarios from human and macaque experiments. **(a)** An episode of the Horizon 1 human experiment: A central target (CT) appears in Frame 1 to indicate the beginning of the episode. A progress bar at the bottom indicates in which trial the participant is. The stimuli are shown separately in Frames 2 and 3. Then the stimuli are shown together in Frame 4. The participant moves the pointer on the larger stimulus as highlighted by the yellow dot in Frame 5. The end of the episode is indicated in Frame 6 with a dot at the center. The stimuli after the subtracted gain are shown separately in Frames 7 and 8. Finally, the updated stimuli are shown together in Frame 9 and the participant chooses the larger one as highlighted in Frame 10. **(b)** Two episodes of the macaque experiment. A CT appears in Frame 1 to indicate the beginning of the episode. The participant touches the CT and holds its finger on it to initiate the episode. Then the peripheral targets (PTs) are shown together on the both sides of the CT in Frame 2. The CT disappears (Go signal) as the participant starts to move its hand to choose one of those PTs as shown in Frame 3. The chosen PT (PT 1) is highlighted by the yellow dot in Frame 4. The corresponding reward (3 water drops) is provided and in Frame 5 a blank screen is shown to indicate the end of Trial 1. The same procedure is followed in Trial 2 but now with the PTs of Group 1. In the second scenario, the cumulative reward is always higher since PT 2 is chosen in Trial 1 and hence the Group 2 PTs, which provide a higher reward, are shown in Trial 2. The nomenclature and corresponding rewards to the PTs are given at the top.

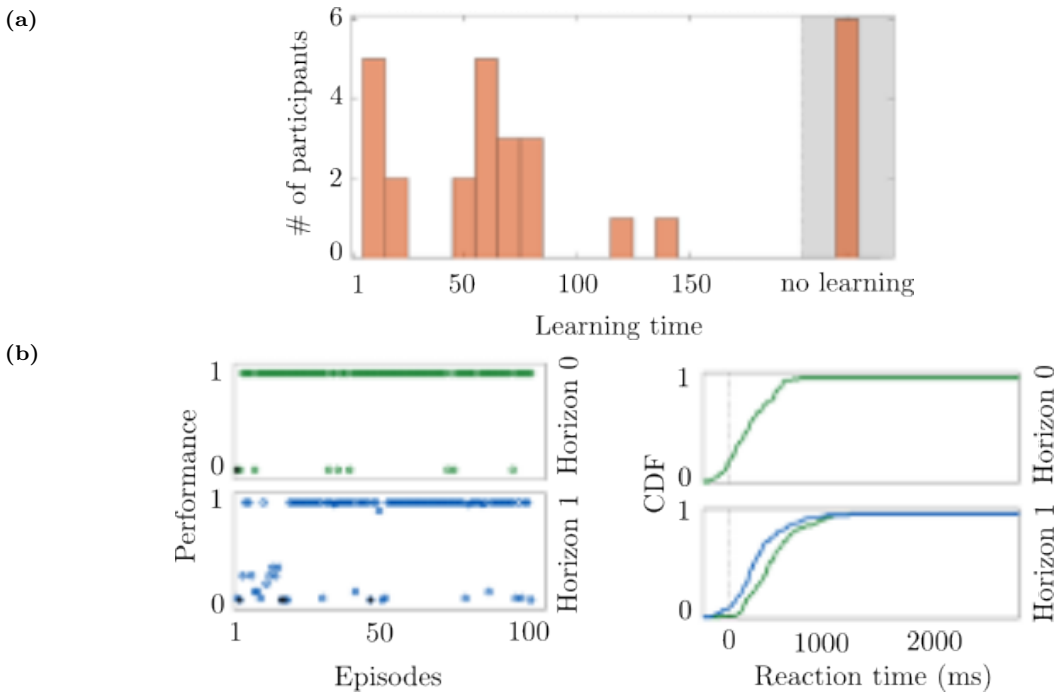


**Figure 2:** Some experiment results from Cecchini et al. [36]. **(a)** The histogram of the learning times obtained from 28 human subjects. The learning time is in terms of episode numbers. The long learning time is classified as no learning. **(b)** The performance of one of the participants, and the cumulative distribution function (CDF) of the reaction times of the same participant. The results are presented separately for Horizon 0 and Horizon 1, where in the latter the CDFs of Trial 1 and Trial 2 are provided in green and blue, respectively.

in Trial 2, otherwise Group 2 will be shown in Trial 2; see Figure 1b.

The same protocol was applied for each trial. As illustrated in Figure 1b, at the beginning of the trial, the central target (CT), which was located at the center of a 17 inch touchscreen monitor (LCD, $800 \times 600$), was shown to indicate the beginning of the episode. To initialize the episode, it was required that the participant touches the CT and holds its finger on the CT for $550 \pm 50$ ms. Once this requirement is fulfilled, Stimuli 1 and 2 are shown as the PTs appearing on the left and right hand sides of the CT. As the participant starts to move his hand after an additional holding time ($\approx 400 - 600$ ms), the CT disappears. This is considered as the Go signal initiating the reaction time, which is tracked for a maximum duration of 2000 ms. If a choice is made between two PTs within this 2000 ms, the reward is delivered to the participant after an additional holding time of 600 ms on the chosen PT. Once this protocol for Trial 1 is completed successfully by the participant, Trial 2 begins after the inter-trial interval of 1200 ms, and it follows the same protocol. An error which appears at any stage of the protocol of any of the two trials results in an interruption of the episode. A new episode is initialized after 2000 ms, which is the inter-episode interval.

Cumulative reward is the sum of the water drops at the end of an episode. The goal of the macaque is to maximize the cumulative reward throughout each episode. This accounts to: (i) to identify the reward values and the correspondence between reward value and stimulus, (ii) making the optimal sequence of decision, that is, selecting PT2 in Trial 1 and PT6 in Trial 2.

There are two major differences in comparison to the human task. Firstly, the stimuli are partially filled bars in the human task, whereas they are some figures chosen randomly from the Powerpoint shapes library in the macaque task. Secondly, in the human task, the higher and lower rewards are provided as the addition of the gain to the filled parts and the subtraction of the gain from the filled parts of the stimuli, respectively. In the monkey task, the reward is provided as a certain number of water drops to the macaque.

# 3 Model

## 3.1 Basic module

Basic module produces the dynamics of two columns competing with each other. Each column is modeled as a pair, called pool, of excitatory and inhibitory populations. Pools $A$ and $B$ vote for Stimulus $A$ represented with $v_A$ and Stimulus $B$ denoted by $v_B$, respectively. We will denote the excitatory populations by $e_A, e_B$ and the inhibitory populations by $i_A, i_B$, where the subindices mark the corresponding pool. There are only excitatory connections across the pools. Each cross-pool excitatory connection targets both the excitatory and the inhibitory populations of the other pool. The excitatory and inhibitory populations within each pool are fully connected with recurrent and cross-population connections; see Figure 3.

In each pool, there exist $N_e$ excitatory and $N_i$ inhibitory neurons accounting all together to $N_{\text{tot}} = N_e + N_i$ neurons. The ratio $N_i/N_e$ is 0.2. We denote the internal excitatory and inhibitory connection weights by $w_+$ and $w_-$, respectively. We employ $w_c^e$ and $w_c^i$ for the intercolumnar connections and the supindices denote the target populations.

Base drive $v_{\text{AI}}$ keeps the excitatory populations in the asynchronous irregular (AI) state. We fix $v_{\text{AI}}$ to 5 Hz. The terms $\lambda^A_{w_r}(v_A, v_B)$ and $\lambda^B_{w_r}(v_A, v_B)$ represent the external inputs with a bias introduced by a regulatory module modeling plasticity as explained in the following section. The biased inputs are fed to Pools $A$ and $B$ through the function $\lambda$ simultaneously.

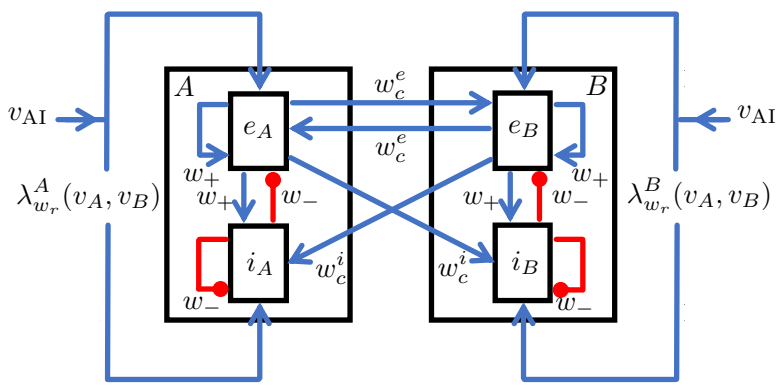


**Figure 3:** Basic module with two pools of excitatory and inhibitory populations

Our model consists of 18 state variables: $v_\alpha$, $C_{\alpha\beta}$ and $W_\alpha$ where $\alpha, \beta \in \{e_A, i_A, e_B, i_B\}$. Here $v_\alpha$ is the firing rate of the population $\alpha$ and $C_{\alpha\beta}$ denotes the cross correlation between $v_\alpha$ and $v_\beta$. The variable $W_\alpha$ denotes the slow adaptation for the population $\alpha$. Inhibitory neurons are known to have no adaptation, therefore $W_{i_A}(t) = W_{i_B}(t) = 0$ for all $t \in [0, \infty)$ with $t$ denoting the time variable. Finally, correlation is symmetric, hence $C_{\alpha\beta} = C_{\beta\alpha}$ for all $\alpha, \beta$.

We assume that the derivatives of the adaptation variables with respect to the firing rates $v_\alpha$ are 0. This is due to the fact that the adaptation variables $W_\alpha$ evolve on a slow time scale and its change with respect to the firing rates $v_\alpha$ evolving on a fast time scale is negligible.

The model equations read as:

$$\begin{aligned}
T\,\partial_t v_\alpha =& (F_\alpha - v_\alpha) + \frac{1}{2} C_{\xi\eta} \partial_{\xi\eta} F_\alpha + \sigma\,\omega_\alpha \\
T\,\partial_t C_{\alpha\beta} =& \delta_{\alpha\beta} A^{-1}_{\alpha\alpha} + (F_\alpha - v_\alpha)(F_\beta - v_\beta) + C_{\beta\xi}\partial_\xi F_\alpha + C_{\alpha\xi}\partial_\xi F_\eta - 2C_{\alpha\beta} \\
\partial_t W_\alpha =& -\frac{W_\alpha}{\tau_w} + \Big(\delta_{\alpha e_A} + \delta_{\alpha e_B}\Big)\Big(b v_\alpha + a\Big(\mu_V(v_{e_A}, v_{e_B}, W_\alpha) - E_L\Big)\Big),
\end{aligned} \tag{1}$$

where

$$\begin{aligned}
F_{e_A} &= F_{e_A}(\tilde{v}_{e_A}, \tilde{v}_{i_A}, W_{e_A}), \qquad & F_{i_A} &= F_{e_A}(\tilde{v}_{e_A}, \tilde{v}_{i_A}, W_{i_A}), \\
F_{e_B} &= F_{e_B}(\tilde{v}_{e_B}, \tilde{v}_{i_B}, W_{e_B}), \qquad & F_{i_B} &= F_{e_A}(\tilde{v}_{e_B}, \tilde{v}_{i_B}, W_{i_B}),
\end{aligned} \tag{2}$$

are the population transfer functions with subindices indicating the corresponding population. Here $\omega_\alpha = w_\alpha(t)$ is a white Gaussian noise:

$$\mathbb{E}[\omega_\alpha(t)] = 0, \qquad \mathbb{E}[\omega_\alpha(t)\omega_\beta(t')] = \delta_{\alpha\beta}\delta_{tt'}, \quad \text{for all } t, t' \geq 0.$$

In (1), $\sigma > 0$ denotes the noise intensity. The function $A_{\alpha\beta}$ is defined as follows [34]:

$$A_{\alpha\beta} = \delta_{\alpha\beta}\frac{N_\alpha}{F_\alpha(1/T - F_\beta)},$$

with $N_\alpha$ denoting the number of neurons in the population $\alpha$. The term $T$ is the time scale parameter both for the firing rate and for the cross correlation variables appearing in (1) and it should be chosen properly not to violate Markovian assumption [34]. We use the same term $\mu_V$ as given in [34, Section 2.3.1]. We use the notation given by $\partial_\alpha = \frac{\partial}{\partial v_\alpha}$ and $\partial_{\alpha\beta} = \frac{\partial^2}{\partial v_\alpha \partial_{v_\beta}}$ for the partial derivatives. Here $\delta$ is the Dirac delta function, $E_L$ is a constant representing the reverse leakage potential, and finally $\tau_w$ is a time scale-like parameter for $W_\alpha$. We express the probability of the intercolumnar connectivity with $p_c \geq 0$. Finally, we write the regulated inputs $\tilde{v}_\alpha$ of the transfer functions given in (2) as follows:

$$\begin{aligned}
\tilde{v}_{e_A}(t) &= v_{e_A}(t) + v_{\mathrm{AI}} + \lambda^A_{w_r}(v_A(t), v_B(t)) \\
&\qquad + w^e_c\Big(v_{e_B}(t) + v_{\mathrm{AI}} + \lambda^B_{w_r}(v_A(t), v_B(t), t)\Big)p_c N_{e_B}, \\
\tilde{v}_{i_A}(t) &= v_{i_A}(t) + \lambda^A_{w_r}(v_A(t), v_B(t)) \\
&\qquad + w^i_c\Big(v_{e_B}(t) + v_{\mathrm{AI}} + \lambda^B_{w_r}(v_A(t), v_B(t), t)\Big)p_c N_{e_B}, \\
\tilde{v}_{e_B}(t) &= v_{e_B}(t) + v_{\mathrm{AI}} + \lambda^B_{w_r}(v_A(t), v_B(t)) \\
&\qquad + w^e_c\Big(v_{e_A}(t) + v_{\mathrm{AI}} + \lambda^A_{w_r}(v_A(t), v_B(t))\Big)p_c N_{e_A}, \\
\tilde{v}_{i_B}(t) &= v_{i_B}(t) + \lambda^B_{w_r}(v_A(t), v_B(t)) \\
&\qquad + w^i_c\Big(v_{e_A}(t) + v_{\mathrm{AI}} + \lambda^A_{w_r}(v_A(t), v_B(t), t)\Big)p_c N_{e_B}.
\end{aligned} \tag{3}$$

Here the time dependency is explicitly denoted and the terms with no explicit time variable are constants.

## 3.2 Regulatory module

Regulatory module introduces a bias by weighting the stimuli in each pool. The pool voting for the promoted stimulus is more likely to have a higher excitatory firing rate than the other pool. Consequently, the pool with the higher excitatory firing rate wins the bicolumnar competition and makes the decision; see Figures 5 and 12. The instant at which the decision is made is the reaction time. It is the instant at which the difference between the excitatory firing rates exceeds a prefixed threshold (see Appendix).

The regulatory module was adapted from [36] and it evolves during the $i$th trial of the $E$th episode via

$$\begin{cases} \tau_\psi \frac{d\psi^E_i(t)}{dt} = -4\psi^E_i(t)\Big(\psi^E_i(t) - 1\Big)\Big(\psi^E_i(t) - 1/2\Big) + \frac{\sigma}{(c_0 t)^2}\zeta_i(t), \qquad t \in (0, t_F], \\ \psi^E_i(0) = \phi^{E-1}_i, \end{cases} \tag{4}$$

where $\tau_\psi$ is the time scale parameter and $\psi$ is the output of the regulatory pool. Here $t_F > 0$ is the final time of the trial and it is the same for every trial. Here $\zeta_i = \zeta_i(t)$ is a white Gaussian noise whose intensity level is a scaled version of $\sigma > 0$, and $c_0 > 0$ is a constant. The noise term $\zeta_i$ introduces a strong stochastic behavior initially, then it decays in time. This noise models mainly the exploratory behavior of the participant.

The evolution process given by (4) is reinitialized at the beginning of each trial of the $E$th episode from the initial condition fixed to the value determined by the reward $\phi^E_i$ corresponding to the same trial but of the $(E-1)$th episode. In this way, the reward mechanism provides for each trial a feedback to the regulatory module at the end of the $(E-1)$th episode. This feedback determines towards which decision the regulatory module will introduce the bias to the basic module in the $E$th episode.

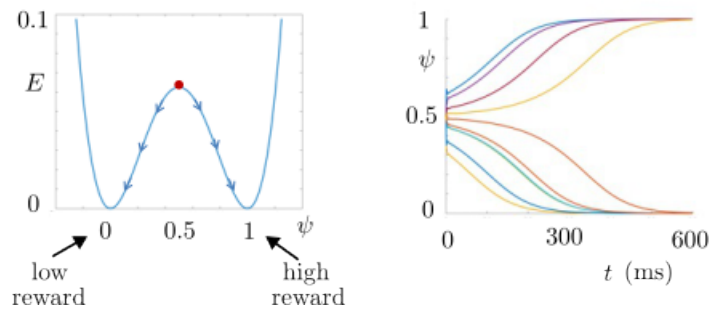


**Figure 4:** Regulatory module. Left: Energy functional of $\psi$. The neutral initial condition is $\psi(0) = 0.5$. The high and low rewards introduce a bias making $\psi(0) > 0.5$ and $\psi(0) < 0.5$ in the next episode, respectively. This results in $\psi = 1$ and $\psi = 0$ at the end of the corresponding trial of the next episode. The former introduces the bias for making the same decision as in the previous episode and the latter for changing the decision. Right: Time evolution of $\psi$ starting from different initial conditions. At the beginning, a small noise is introduced for modeling whatsover which might perturb the regulatory process, such as the exploratory behavior or perceptual difficulties.

The dynamics of the regulatory module is transmitted to Pool $A$ and Pool $B$ via

$$\begin{aligned} \lambda^A_{w_r}(v_A(t), v_B(t), t) =& \psi(t)\, v_A(t) + (1 - \psi(t))\, v_B(t), \\ \lambda^B_{w_r}(v_A(t), v_B(t), t) =& \psi(t)\, v_B(t) + (1 - \psi(t))\, v_A(t). \end{aligned} \tag{5}$$

At each trial, both stimuli are shown to the participant and they have an excitatory effect on all the populations. The decision depends on the evolution of the function $\psi$, which converges to either 1 or 0. The convergence to 1 pushes the model to choose the bigger stimulus. The convergence to 0 pushes the model to switch to choose the smaller stimulus.

The regulatory module can be thought of as a gating mechanism arranging motor-plan flexibility for the response to the stimuli (or the inputs evoked by the stimuli). This mechanism might provide an explanation to the observations showing that relevant sensory inputs, distractors, and direct perturbations of decision-making circuits affect the behavior more strongly when they are introduced in the early phase of the trial [37–42].

## 3.3 Reward mechanism

Reward mechanism endows the model with online learning. It allows the model to learn the preset strategy maximizing the cumulative reward in each episode. Once the strategy is learned, the system makes the decisions in almost complete coherence with the strategy.

The reward mechanism is updated through a discrete evolution where the temporal variable is the episode number. In other words, the reward function value remains constant during a trial and it is updated at the end of the trial. The updated value is fed as the initial condition to the regulatory function $\psi$ in the trial corresponding to the episode coming after; see (4). We use the notation $M_i^E$ to denote the mean value of the stimuli corresponding to the $i$th trial of the $E$th episode. At the end of the episode, e.g., with $i = 2$ in Horizon 1, $M_3^E$ refers to the mean value updated by the gain. We write $N$ to express the number of trials in one episode. Then we write the evolution for the reward function $\phi$ as follows:

$$\begin{cases} \phi_i^{E+1} = \phi_i^E + k\,(M_{i+1}^E - M_i^E)(2\psi_i^E(t_F) - 1)(\phi_i^E - 1)^2(\phi_i^E)^2, \\ \phi_i^1 = C, \quad \text{for all} \quad i \in \{1, \ldots, N\}, \end{cases} \tag{6}$$

with $C$ denoting a constant, which is fixed to 0.5 in our framework. This system initiates the reward mechanism for each trial separately, yet the trials are not independent due to the coupling effect arising from $(M_{i+1}^E - M_i^E)$ factor in (6).

Finally, the only difference between the frameworks which we use in human and macaque simulations appears in (6). In the case of macaque, we assign to each stimulus, the number of water drops corresponding

to the reward of the stimulus. This allows us to translate each symbol to a numerical value, and a value associated to the corresponding reward. We replace the first line of (6) with

$$\phi_i^{E+1} = \phi_i^E + k\, F_i\, (2\psi_i^E(t_F) - 1)(\phi_i^E - 1)^2(\phi_i^E)^2, \tag{7}$$

where

$$F_i^E := (\delta_{i-1}(\text{other}_1^E - \text{choice}_1^E) + \delta_{i-2}(\text{choice}_2^E - \text{other}_2^E)), \tag{8}$$

with $i$ and $\delta$ denoting the trial number and Dirac delta, respectively. Here $\text{choice}_i^E$ and $\text{other}_i^E$ denote the numbers of drops given as rewards for the chosen and the other stimuli in the $i$th trial of the $E$th episode, respectively.

### 3.4 External stimuli

One of the most common choices to model the external stimuli is Heaviside function in mean-field models. However, since Heaviside functions have discontinuity due to the discrete jump, our biophysical model undergoes a transient phase which might obscure the instant marking the reaction time. For this reason, we performed simulations by using sufficiently sharp sigmoid function as external stimulus; see Figure 12a and Figure 13 in Appendix.

### 3.5 Complete model with the reward mechanism

We show in Figure 5 the results of our model simulation with 6 episodes, each composed of 2 trials. The reward mechanism is initiated from 0.5 for both trials in the first episode. We use the same parameters given in (17)-(19). The parameter $d$ is kept constant for all episodes.

The preset strategy requires choosing the smaller stimulus (the bar which is filled less) in the first trial and the larger stimulus (the bar which is filled more) in the second trial at each episode. We observe in Figure 5 that the model explores the strategy in Episode 1. Its decision is completely random since $\psi_i^0(0) = 0.5$ for all $i \in \{0, 1, \ldots, N-1\}$. We observe that in Trial 1, there is no winner of the competition. In Episode 1, the reward mechanism introduces a bias in the regulatory function $\psi$ by feeding the reward to $\psi$ as the initial condition at the beginning of each trial. The same procedure is repeated for every episode. We observe in Figure 5 that after Episode 1, the model learns the strategy and starting from Episode 2, it makes the choice in accordance with the strategy. The speed of learning depends on the parameter $k$ in (6). In Figure 5, we chose $k = 0.1$ and we fixed the reward gain $G$ to 0.75. The gain is added to the stimuli if the correct decision according to the preset strategy is made, otherwise, it is subtracted from the stimuli; see Figure 1a.

## 4 Noise sources

We model distortion effects as an Ornstein-Uhlenbeck (OU) process which can be sampled directly from a Gaussian distribution once the initial condition of the OU process is chosen properly. This is similar to the previous framework [34], however with one difference: We choose the initial condition of the OU process as a zero mean Gaussian, where its variance is scaled by the convergence rate of the OU process as given below. This allows to avoid explicit simulation of the OU process.

Formally, the noise in mean-field models is a stochastic process evolving independently of the rest of the state variables. For a white Gaussian noise, this evolution can be written as an OU process of the following type:

$$\begin{aligned} d\omega_\alpha(t) =& -\theta_\alpha \omega_\alpha(t)dt + \sigma dW_\alpha(t), \\ \omega_\alpha(0) =& \Omega, \end{aligned} \tag{9}$$

where $W_\alpha(t)$ is a standard Brownian motion, $\theta_\alpha$ is a positive constant associated to the convergence rate of the process to its long term mean, and $\Omega$ is the initial condition, which is a centered Gaussian noise with variance $\sigma_0^2 > 0$. This is a linear stochastic differential equation and its unique solution [43] is

$$\omega_\alpha(t) = \mathrm{e}^{-\theta_\alpha t}\, \Omega + \sigma \int_0^t \mathrm{e}^{-\alpha(t-s)}\, dW_\alpha(s), \tag{10}$$

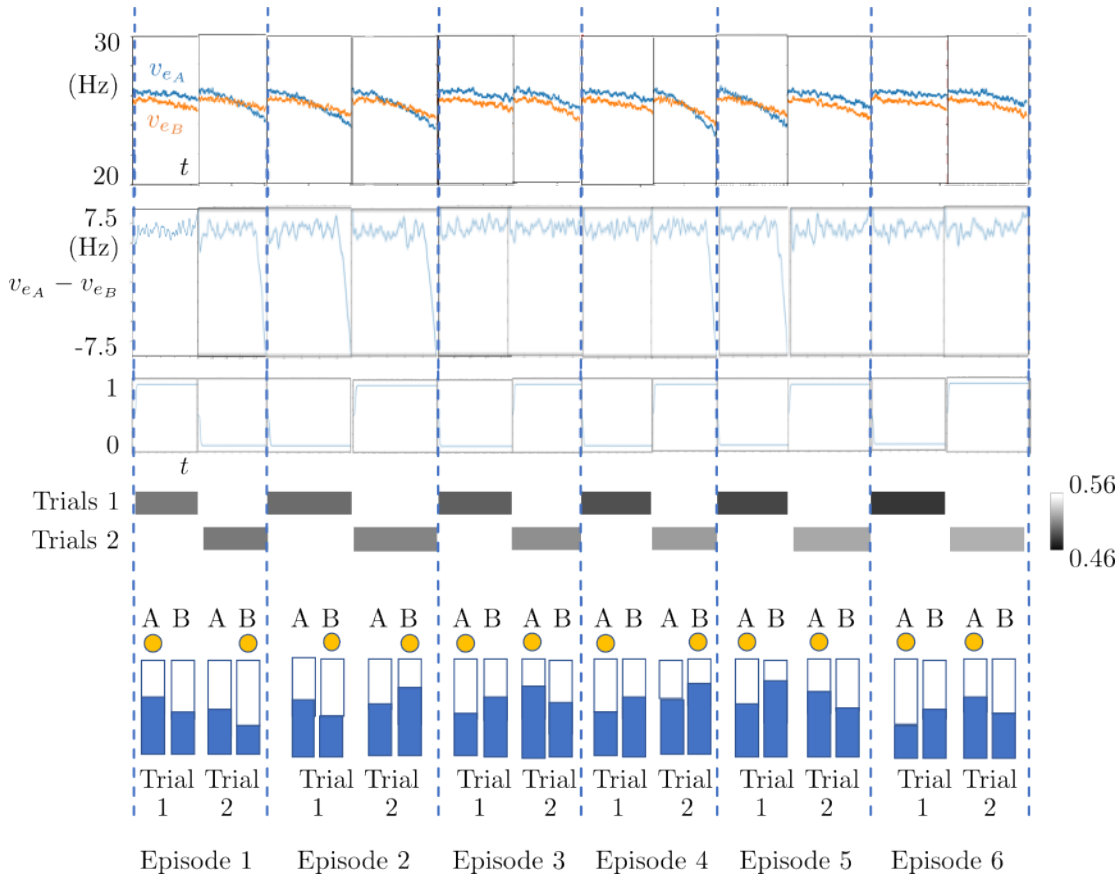


**Figure 5:** Simulation results of 6 episodes. First row: Time courses of the excitatory population firing rates. Second row: Time courses of the difference of the excitatory population firing rates. Third row: Time course of the regulatory function $\psi$. Fourth row: Time course of the reward function $\phi$ for each trial of the corresponding episode shown at the bottom row. Bottom row: External stimuli. Chosen stimuli are highlighted by the yellow dot at the top. Trial and episode numbers are given at the bottom.

where the solution $\omega_\alpha(t)$ is a Gaussian process. It can be written for a small time step $0 < h \ll 1$ as

$$\omega_\alpha(t+h) = \mathrm{e}^{-\theta_\alpha h}\,\omega_\alpha(t) + \xi(t), \tag{11}$$

where $\xi(t)$ is a centered Gaussian white noise with variance $\sigma_\xi^2 = \frac{\sigma^2}{2\theta_\alpha}(1 - \mathrm{e}^{-2\theta_\alpha h})$, and which is sampled independently for each $t$ and population $\alpha$. We observe that $\omega_\alpha(t+h)$ is equivalent to a zero mean Gaussian with variance

$$\sigma_{\omega_\alpha}^2(t+h) = \mathrm{e}^{-2\theta_\alpha h}\,\sigma_{\omega_\alpha}(t)^2 + \sigma_\xi^2. \tag{12}$$

Once we choose $\sigma_{\omega_\alpha}^2(0) = \sigma_0^2 = \frac{\sigma^2}{2\theta_\alpha}$, we observe that $\sigma_{\omega_\alpha}^2(t+h)$ is time independent and equal to $\frac{\sigma^2}{2\theta_\alpha}$. This allows us to generate $\omega_\alpha(t)$ from the Gaussian distribution $\mathcal{N}(0, \frac{\sigma^2}{2\theta_\alpha})$ independently at each instant $t > 0$, avoiding explicit forward time simulations of the OU process given in (9). This is the idea behind introducing the noise terms $\omega_\alpha$ in the mean-field system (1) and the noise term $\zeta_i$ appearing in the regulatory mechanism (4) as Gaussian white noise sampled directly from a Gaussian distribution at each time and for each population, but not as an explicit OU process evolving in time.

## 4.1 Extracellular media distortion effects

In the previous AdEx mean-field model [34], the background noise was introduced as an additive noise to the base drive potential $v_{\mathrm{AI}}$. This noise was evolving as an OU process and it modeled dynamically the distortion effects appearing in the firing rates due to the extracellular medium and synaptic perturbations.

In our model, we introduce the white Gaussian noise $\omega_\alpha$ scaled by the noise intensity $\sigma$ explicitly in the firing rate equations as shown in (6). In this way, we avoid that the noise undergoes the nonlinear effects of transfer functions since now it is additive in the model equations.

### 4.2 Exploratory behavior

We quantify the coherence between the choices of the participant and the preset strategy in terms of performance values. In each episode, the maximum performance refers to the case in which the participant makes the choice in coherence with the preset strategy in every trial of the episode. The maximum performance value is 1. The minimum performance refers to the case in which the participant makes the incoherent choice in every trial of the episode. The minimum performance value is 0. All other cases with partial coherence are scaled between 1 and 0; see (13).

One of the typical behaviors in the decision making task is that the participant makes decisions which do not comply with the preset strategy even after the participant learned the strategy. This behavioral pattern is observed as irregular and rather sparse deviations from the maximum performance value. That is, we observe few performance results which are less than 1 within the blocks of maximum performance results; see Figure 11 in Appendix. This behavior occurs due to two reasons: (i) perceptual difficulty in the human experiment, i.e., the difference between stimuli is small, thus the participant makes the wrong decision as a result of perceptual difficulty; (ii) the participant is willing to explore whatsoever related to the experiment setup, and this might happen in both human and macaque experiments.

The first reason is due to the stimuli, and it does not require an additional mechanism in the model. The second one is due to the participant's exploratory will, and we model this via the Gaussian white noise $\zeta_i$ introduced in the regulatory mechanism (4). This noise is sampled independently at each time instant, and noise level $\sigma$ is scaled by $(c_0 t^2)$, where $c_0 > 0$. Here $t$ denotes the time instant and it is reinitialized at the beginning of each trial. This scaling term introduces to the bicolumnar competition a strong initial perturbation, which decays in time. This decay is needed since otherwise the noise could dominate the whole bicolumnar competition instead of perturbing only the initial bias. The parameter $c_0$ determines how strong the initial perturbation is. The regulatory mechanism therefore, is not only for an online reward-driven learning of the preset strategy but also for providing a natural behavior which is open to making wrong decisions. This is similar to what we observe in both human and macaque experiments.

## 5 Quantification of behavior

### 5.1 Global measures

The performance deviations arising from the aforementioned exploratory behavior and perceptual difficulties are momentary, they are locally concentrated around certain episodes appearing after the participant learned the preset strategy. Therefore, quantification of the behavioral performance requires global measures which are robust to such local deviations and which can quantifiy these deviations as well. For this, we use three objects which are highlighted in Figure 11; see Appendix. They are defined as follows:

**Definition** (*Performance deviation*): A performance deviation is a point with a performance lower than the maximum performance value ($= 1$) in the set of performance samples, except for the first sample.

**Definition** (*Deviation cluster*): A deviation cluster is a set of at least 3 successive performance deviations with respect to the episode numbers.

**Definition** (*Performance cluster*): Performance cluster is the set of performance samples in which there is no deviation cluster and whose last performance sample is also the last sample of the whole experiment.

**Definition** (*In/out-cluster deviation*): A performance deviation is called *in-cluster* if it occurs in a performance cluster, and *out-cluster* otherwise.

A cluster starts at the episode where the participant begins to make decisions coherently with the preset strategy. Within the cluster, the participant makes coherent decisions almost in each episode until the end of the cluster. Almost; because the participant might make decisions in the aforementioned exploratory manner,

producing in-cluster deviations. Note that this does not break the performance cluster as long as in-cluster deviations do not constitute a deviation cluster.

## 5.2 Characterization of the model behavioral performance

Learning speed determines how quickly the model learns the preset strategy in an experiment. The higher it is, the faster the model identifies the strategy. Flexibility parameter determines how much the model deviates from the preset strategy after it learns the strategy. It models the exploratory behavior and the deviations caused by perceptual difficulties. Finally, gain rescales the learning speed. Here we provide the effects of these three parameters on the model behavior performance.

Initial stimuli in the model are $M_0^E \pm \frac{d}{2}$, where $M_0^E$ is generated from a uniform distribution independently for each episode and $d$ is a constant. Here $d$ determines the difficulty of the trial. If $d$ is small, then it is more difficult to make a distinction between two stimuli. The value of $d$ remains constant throughout one episode, however, this value need not necessarily be the same for different episodes.

We show in Figure 6a the performance plots of three cases in which the learning speed $k$ appearing in (6) varies from 0.1 to 0.3. We denote by $V_i^E$ the value of the chosen stimulus at the end of the $i$th trial of the $E$th episode. We express as $V_{\min}^E$ and $V_{\max}^E$ the maximum and minimum values, respectively, which $\sum_{i=1}^{N} V_i^E$ can attain among all possible scenarios of the $E$th episode. We measure the performance of the model in the $E$th episode by using

$$\text{Performance}(E) := \frac{\sum_{i=1}^{N} V_i^E - V_{\min}^E}{V_{\max}^E - V_{\min}^E}. \tag{13}$$

We observe in Figure 6a that as we increase $k$, the system captures the strategy earlier, and makes constantly right decisions thereafter, with very few deviations.

The gain $G$ is another factor affecting the learning speed. The learning speed decreases as the gain decreases since the gain $G$ is equal to the multiplying factor $(M_{i+1}^E - M_i^E)$ of the learning speed $k$ in (6). As shown in Figure 6b, the model learns the preset strategy faster as we increase the gain.

Finally, the flexibility parameter $c_0$ determines the number of deviations in the model performance results. The higher it is, the closer to the deterministic case the model is. This is due to the fact that the noise given in (4), and which produces performance deviations, vanishes very quickly after that the regulatory variable $\psi$ starts to evolve at the beginning of each trial; see Figure 6c. Moreover, $c_0$ determines the beginning of the performance cluster together with the learning speed. As $c_0$ increases, the performance cluster is more likely to begin from smaller episode numbers.

## 5.3 Characterization of the model reaction time

In Figure 7, we provide the histograms of the reaction time measurements obtained from the Horizon 1 simulations. In these simulations, we varied the learning speed $k$ and the flexibility parameter $c_0$, separately in Trial 1 and Trial 2. We observe that there is no distinguishing change in the histograms of the simulation results, suggesting that these two parameters do not have noticeable effects on the reaction times. This is due to the fact that the reaction times are calculated by checking the threshold crossing of the difference between the firing rates $v_{e_A}$ and $v_{e_B}$. Here $k$ and $c_0$ do not enter in the equations that describe the firing rates, so they do not influence the reaction times. As the decision threshold increases, the competition between $v_{e_A}$ and $v_{e_B}$ lasts longer, thus the reaction times increase as shown in the bottom row of Figure 7. In Appendix, similar results for Horizon 0 simulations can be found in Figure 15.

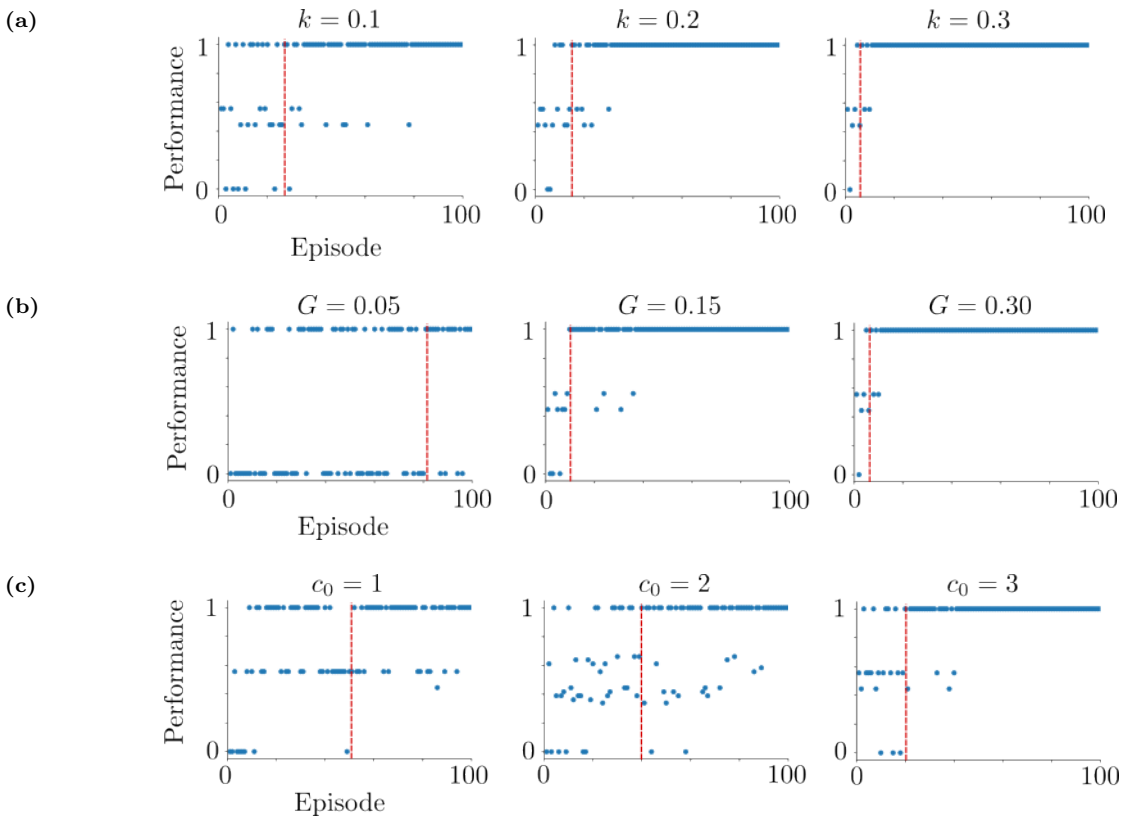


**Figure 6:** Simulation results of Horizon 1 performances in the human task. Initial episodes of the performance clusters are highlighted by the red vertical lines. The difficulty $d$ is 0.2. See Appendix for the rest of the parameters. **(a)** The learning speed $k$ is 0.1, 0.15 and 0.2 from left to right. The flexibility parameter $c_0$ is 2 and the gain $G$ is 0.3. The initial episodes of the performance clusters are episodes 24, 15 and 5 from left to right. **(b)** $G = 0.05$, $G = 0.15$ and $G = 0.3$ from left to right. Here $k = 0.3$ and $c_0 = 2$. The initial episodes of the performance clusters are episodes 81, 10 and 5 from left to right. **(c)** $c_0 = 1$, $c_0 = 2$ and $c_0 = 3$ from left to right. Here $k = 0.05$ and the gain $G = 0.3$. The initial episodes of the performance clusters are episodes 52, 40 and 20; the number of in-cluster deviations are 12, 5 and 4 from left to right.

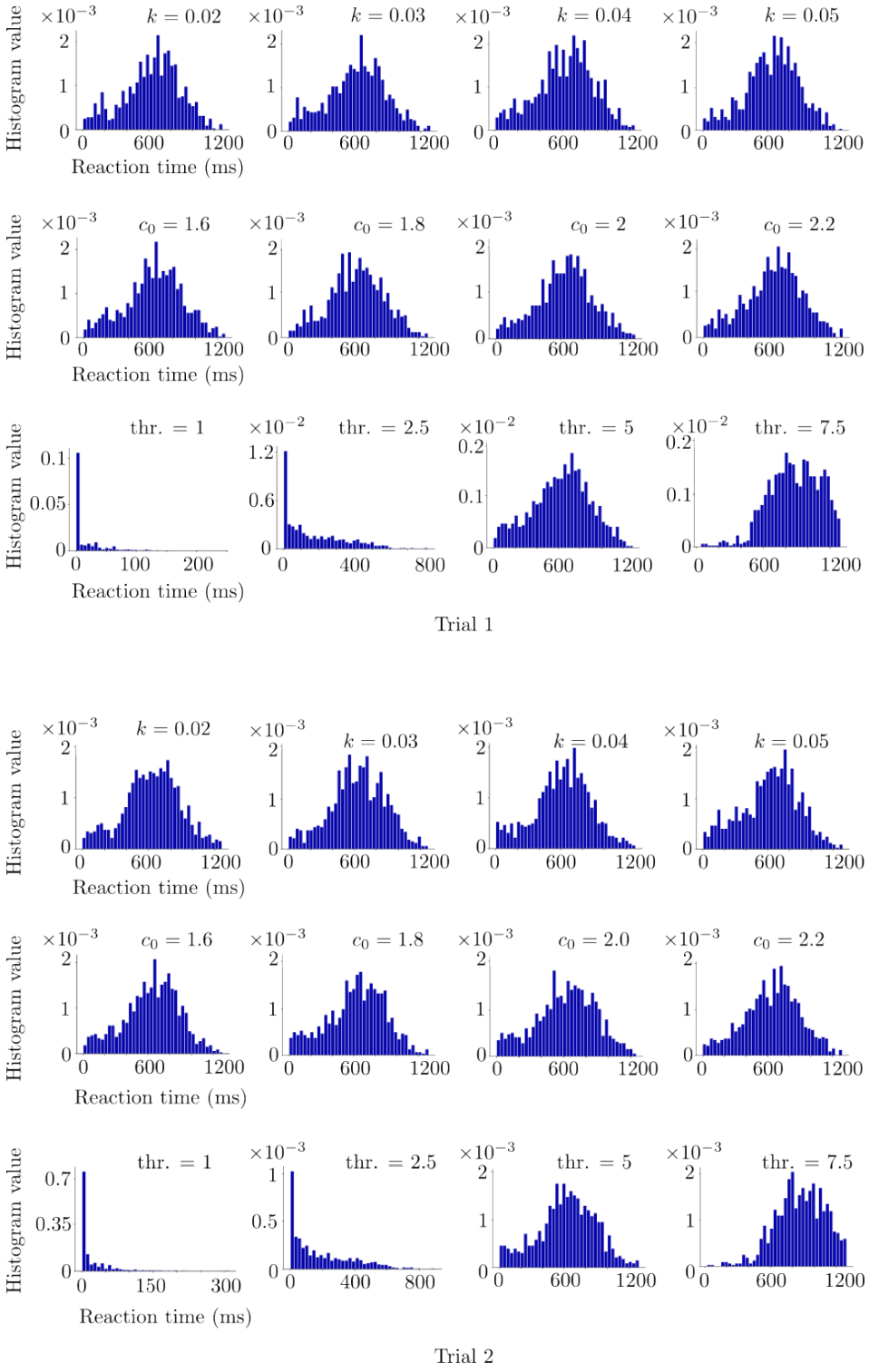


**Figure 7:** Model simulation results regarding Horizon 1 reaction time histograms of Trials 1 and 2. Varying the learning speed $k$ and the flexibility parameter $c_0$ does not affect noticeably the distribution of the reaction times in terms of mean and standard deviation. Increasing the decision threshold increases the mean and the standard deviation of the distribution. The histograms are obtained from 10 realizations of the same simulation setup for each parameter. The parameters are the same as those of Figure 8.

# 6 Results

We provide a comparison of the simulation results to the experiment results. We consider only Horizon 1 here. The comparison for Horizon 0 can be found in Appendix; see Figures 14 and 16. In Figure 8, we provide the performance results of the model and their comparison to the performance results of the human participant. Then, in Figure 9, we provide the histograms of the reaction times corresponding to the simulations and we compare them to the reaction time histograms of the human participant. Then we continue by providing the simulation performance results compared to the performance results obtained from the macaque experiment in Figure 10. The simulation results of the reaction time histograms and their comparison to the macaque histograms are given in Appendix; see Figure 17.

## 6.1 Comparison to the Horizon 1 human performance

We provide the Horizon 1 simulation and human experiment results as quantified based on the mean and standard deviation of the initial episode number of performance cluster, as well as the mean and standard deviation of the in-cluster performance deviations. This quantification fully describes the learned decision-making phase of the performance results. We obtain those simulation statistics from 10 realizations of each parameter set presented in Figure 8. Each realization is composed of 100 episodes. Red horizontal lines indicate the initial cluster episode number and the number of in-cluster deviations obtained from the human experiment. We provide the Horizon 0 results in the same fashion in Figure 14.

In Figure 8, we observe that as $k$ and $c_0$ increase, the initial episode number of performance clusters decreases as shown on the left column. This is the case in Horizon 0 as seen in Figure 14. However, differently from Horizon 0, the decreasing curves in Horizon 1 are concave, meaning that the decrease rate of the mean initial episode number of performance clusters increases. We observe that this cannot be compensated by the number of in-cluster deviations, which increases as $k$ and $c_0$ increase. This is different from the Horizon 0 simulations. We observe that the model can produce close results to the experimental values once $k$ and $c_0$ are chosen properly, for example $k = 0.05$ and $c_0 = 2$.

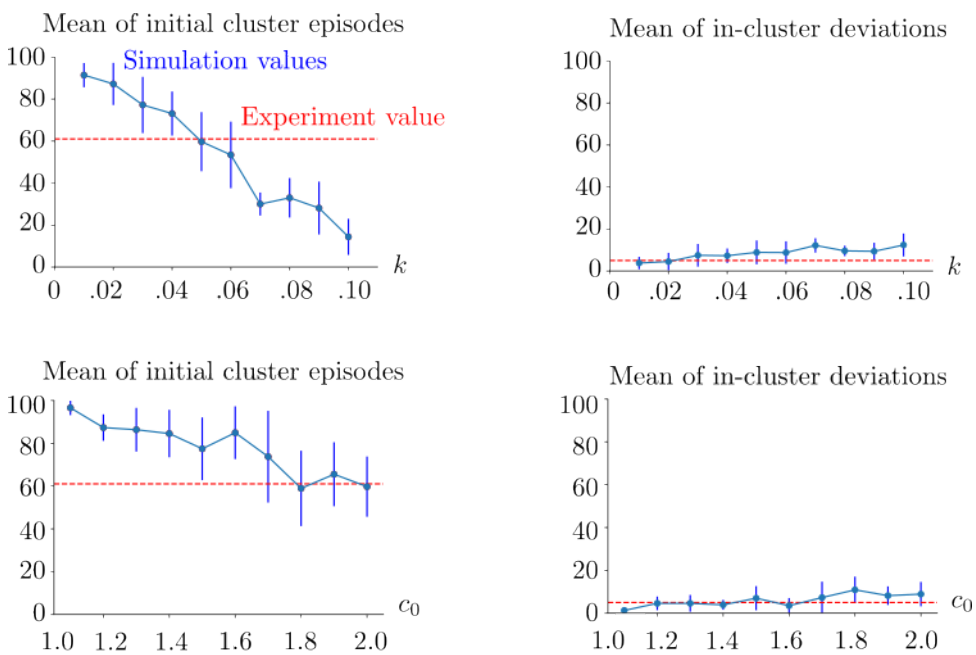


**Figure 8:** Horizontal 1 human case. Simulation statistics with respect to the varied learning speed $k$ in the top row, and with respect to the varied flexibility parameter $c_0$ in the bottom row. In the top row, $c_0 = 2$. In the bottom row, $k = 0.05$. The vertical blue lines show the standard deviations of the simulation results. The red horizontal lines indicate the initial cluster episode number and the number of in-cluster deviations obtained from the human experiment. The mean and standard deviation are obtained from 10 realizations of the same simulation setup for each $(k, c_0)$ pair. See Appendix for the rest of the parameters.

## 6.2 Comparison to the Horizon 1 human reaction times

Reaction time in one trial is measured as the time duration between the instant where the stimuli are shown simultaneously on the monitor and the moment when the participant starts to move the pointer. In the simulations, we measure the reaction time as the time duration between the instant where the two stimuli provided to the model and the time instant at which the decision is made, i.e., when the difference between the excitatory population firing rates $v_{e_A}$ and $v_{e_B}$ exceeds the decision threshold, which is fixed to 5 Hz.

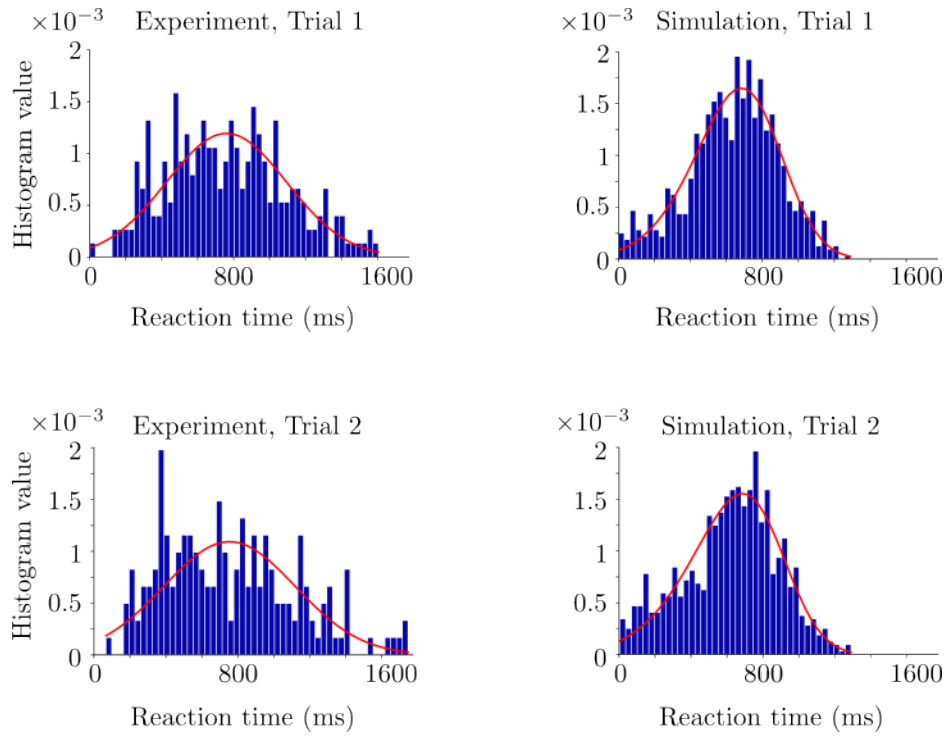


**Figure 9:** Horizon 1 human case. Histograms obtained from the human experiment and simulations for comparison. The red curves denote the fitted distributions via Python. Trial 1 and 2 histograms are in the top and bottom rows, respectively. The simulations are performed with $k = 0.05$, $c_0 = 2$ and decision threshold $= 0.5$.

In Figure 9, we show the simulation and experiment histograms of the reaction times corresponding to Horizon 1, together with the fitted distributions for the best performance-providing $k$ and $c_0$ tested values. In Figure 9, we fit distributions of different types to both simulation and experiment histograms. We use the *Fitter* function from the Python *Fitter* package. We consider the histograms of the reaction times obtained from the first and second trials separately. In Trial 1, skewrow distribution and hypergeometric distribution are fitted to the experiment and simulation histograms, respectively. Corresponding squared sum error values are 1.3e-5 and 9e-6 for the experiment and simulation histograms, respectively. Pearson correlation coefficient of these fitted distributions is 0.92. In Trial 2 histograms, skewrow distribution and hypergeometric distribution are fitted to the experiment and simulation histograms, respectively. Corresponding squared sum error values are 1.7e-5 and 1.1e-5 for the experiment and simulation histograms, respectively. Pearson correlation coefficient of these fitted distributions is 0.63. Trial 2 has a weaker overlap between the experiment and simulation histograms compared to Trial 1.

## 6.3 Comparison to the macaque performance

We observe that the trends in the performance results of macaque are similar to the ones found in the Horizon 1 human task simulations. We see that the simulation results overlap with the macaque experiments for properly chosen parameters, for example $k = 0.12$, $c_0 = 0.106$. For a comparison based on the reaction time histograms, we refer to Figure 9 given in Appendix.

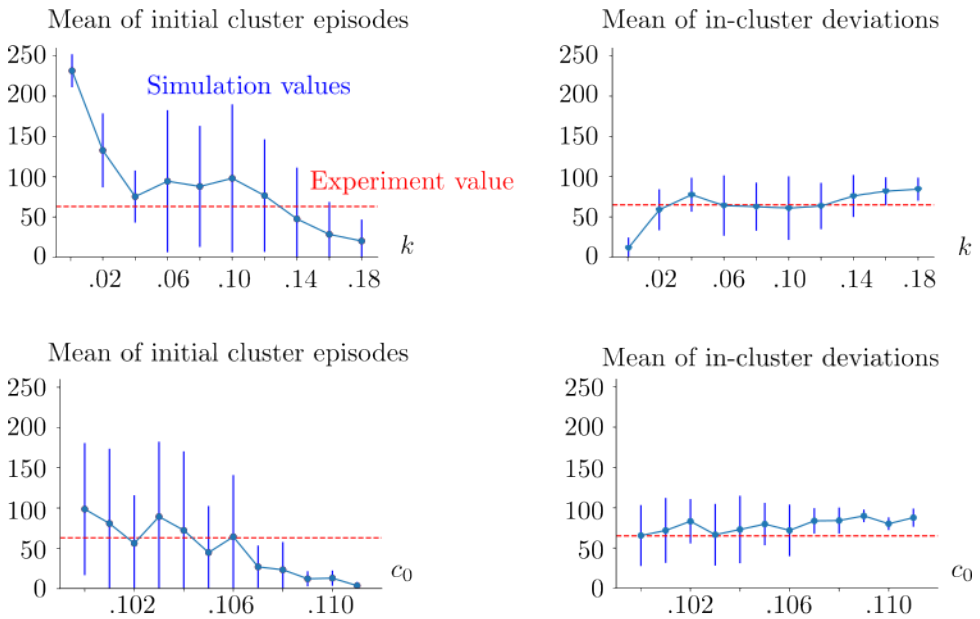


**Figure 10:** Macaque case. Comparison of the Horizontal 1 cluster statistics to the macaque experiment with respect to the varied learning speed $k$ in the top row, and with respect to the varied flexibility parameter $c_0$ in the bottom row. In the top row, $c_0 = 0.106$ and $k$ is varied. In the bottom row, $k = 0.12$ and $c_0$ is varied. The vertical blue lines show the standard deviation of the corresponding statistical sample. The red dashed lines show the value obtained from the macaque experiment. The mean and standard deviations are obtained from 10 realizations the same simulation setup for each $k, c_0$ pair. The decision threshold is 5 Hz. The rest of the parameters are given in Appendix.

# 7 Discussion

We presented a biologically plausible AdEx mean-field model based on interacting cortical columns. We extended the classical AdEx mean-field model [34] to two columns with bicolumnar competition. We applied this model to a reward driven consequential decision making task at behavioral level. We compared the simulation results of the model to the human and macaque experiment results in terms of behavioral performance and reaction time. Novelties of the model are at both connectivity and structure levels.

Regarding connectivity, intercolumnar connections between two columns are only excitatory. Two cross connections originate from the excitatory population of one column, and one of them is afferent to the excitatory population and the other one is afferent to the inhibitory population of the other column. This connectivity overlaps with the long range cortico-cortical connections found in the prefrontal cortex. Consequently, we can model columns as pools of excitatory and inhibitory cells, where inhibitory control over the excitatory population is triggered by the increasing activity of the excitatory population of the other column. This is not the case in the previously proposed models [2, 24]. In [24], there is only one inhibitory interneuron population which is connected to all excitatory populations via the long range connections. In [2], there is no separate inhibitory interneuron population, the excitatory populations inhibit mutually each other.

This bicolumnar architecture is a mathematically rigorous extension from the AdEx mean-field model [32–34], which was proposed in [34] for AdEx network of an isolated single column with adaptation. This extension is necessary to take into account long range cortico-columnar connections. In the model, we provoke the bicolumnar competition via those connections. The competition results in choosing one of the two alternatives by the model.

On the structure side, we model each column as the mean-field limit of a pair of excitatory and inhibitory neuronal populations. The parameters are directly linked to the biophysical mechanisms of RS and FS cells. This allows our model to be applied to not only behavioral data but also to relevant neurophysiological data, in particular to multi-site array recordings of the macaque prefrontal cortex. In this way, the model can bridge phenomenological decision-making models [2, 24, 36] to neural dynamics observed at mesoscopic level.

Moreover, we embedded a regulatory module in the AdEx system to introduce the plasticity which learns the preset strategy in the task. This machinery is essential for the model to find the optimal decision pattern. The regulatory module sustains the decision pattern once the strategy is learned. This is due to the fact that once the introduced bias provides high rewards successively over episodes, the regulatory module reinforces and retains the bias. This relates the regulatory module to working memory at phenomenological level.

Finally, the AdEx mean-field framework was never used for a cognitive task and at a behavioral level before. We show here that it can reproduce the output of whole-brain scale activity represented on a behavioral scale, in the context of the considered decision-making tasks. This supports the idea of that the AdEx mean-field framework can be embedded in a whole-brain simulator such as The Virtual Brain (TVB) [44] and can reproduce whole brain dynamics induced by cognitive tasks, in particular by decision-making tasks.

The model was tested at behavioral level by comparing its predictions to the experiment data obtained from human and macaque participants. The comparison was made based on three measures: mean value of initial episode number of performance cluster, mean value of in-cluster performance deviations and reaction time distributions. It was found that the model reproduces several characteristics of the experiment results. To begin with, the model predicts correctly the performance measures in the cases of both human and macaque for properly chosen parameter sets. It reproduces closely the reaction time distributions of the human data. The correlation values between the simulation and experiment reaction time histograms are lower in Trial 2 of Horizon 1 in both human and macaque compared to Trial 1 decision times. A possible reason for this is that the decision threshold is fixed. It is not dynamically adapted over the episodes as the strategy is learned.

Optimal decision-making refers to making choices which provide the maximum possible reward [45]. Albeit absence of precise definition, sub-optimal decision-making refers to decision-making pattern deviating from the preset strategy in terms of performance measures. Our model is designed for optimal decision-making, and it has the potential to be extended to sub-optimal decision-making. The model already reproduces partially sub-optimal behavior in the performance clusters thanks to its exploratory behavior as shown, for example, in Figures 6a–6c. This property can be improved by introducing in the regulatory module a term weighting flexibility parameter based on loss or augmentation of motivation of the participant to conduct the task. This can be combined with dynamically adapted decision thresholds to the performance so as to project the changes in the motivation of the participant. This modification could provide a better understanding of the neural dynamics which modulate behavioral strategies of the participant in accordance with the increased or decreased motivation of the participant. This has been an active research area in experimental studies [46–48], which will benefit enormously from computational support.

Finally, a further extension of the model can be towards large scale brain dynamics of decision-making. This requires simplification of the model, in which the second order terms (firing rate cross correlations) could be omitted in (1). This simplification allows to embed the AdEx framework in TVB [44]. In this way, the inputs to the embedded model can be provided from visual areas and the output of the model can be provided to motor areas in a biologically realistic way. Nevertheless, it is not straightforward how to and for which regions such embedding should be done.

# Appendix

## Simulation parameters

The parameters which are used in Figures 6-10 and Figures 14-17 are as follows:

$$\begin{gathered} T = 5\,\mathrm{ms}^{-1}, \quad \sigma = 0.01, \quad \tau_w = 5000\,\mathrm{ms}^{-1}\ \text{(for RS)}, \quad 1^{-9}\,\mathrm{ms}^{-1}\ \text{(for FS)}, \\ a = 4\ \text{(for RS)}, \quad 0\text{(for FS)}, \quad b = 40\ \text{(RS)},\ 0\ \text{(FS)}, \quad E_L = -65\,\mathrm{mV}, \\ N_{e_A} = N_{e_B} = 8000, \quad N_{i_A} = N_{i_B} = 2000, \end{gathered} \tag{14}$$

and it is given for (3) as

$$v_{\mathrm{AI}} = 5\,\mathrm{Hz}, \quad w_c^e = w_c^i = 2.5 \times 10^{-4}, \quad p_c = 0.8. \tag{15}$$

Finally, the parameters appearing in (4) and (5) are

$$\tau_\psi = T = 5, \qquad \sigma = \sigma_r = 0.01, \qquad w_r = 1. \tag{16}$$

Each trial lasts 15 seconds in both Horizontal 0 and 1 simulations. We set the decision threshold to 5 Hz in all simulations.

## Global measures

In Figure 11, we provide a visual highlight of the definitions of the global measures given in Section 5.1.

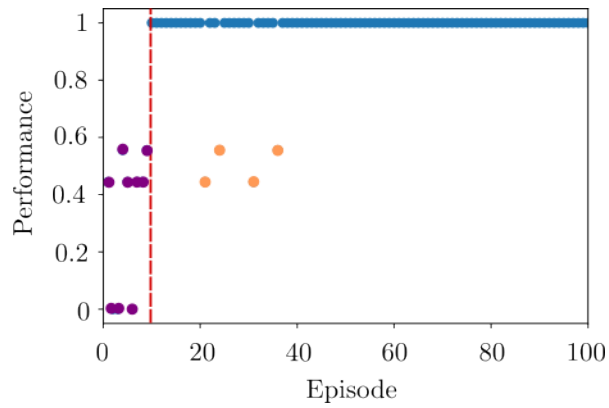


**Figure 11:** Highlighted global objects in Horizon 1 human experiment results. The violet samples correspond to a deviation block, and they are out-cluster deviations. The orange samples are in-cluster deviations, they do not form a deviation block. Both the violet and the orange samples are performance deviations. The blue samples are maximum performance samples and they form a performance cluster. The vertical red line marks the beginning of the performance cluster.

## 7.1 Effect of varying the bias

In this section, the results focus on the effects of varying the initial condition $\psi(0)$, therefore the effects of varying the bias. We fix the threshold to measure the reaction time $\bar{t}$ is $|v_{e_A}(\bar{t}) - v_{e_B}(\bar{t})| = 7.5$. We use Euler-Maruyama scheme with time step $\Delta t = 0.5$ and $t_f = 4$ for the results presented in Figure 12. Stimuli are applied at $t_0 = 2$. The parameters in this framework are as follows:

$$\begin{gathered} T = 0.003, \quad \sigma = 0.01, \quad \tau_w = 500 \text{ (for RS)}, \ 10^{-9} \text{ (for FS)}, \\ a = 1 \text{ (RS)}, \ 0\text{(FS)}, \quad b = 100 \text{ (RS)}, \ 0 \text{ (FS)}, \quad E_L = -65, \\ N_{e_A} = N_{e_B} = 8000, \quad N_{i_A} = N_{i_B} = 2000. \end{gathered} \tag{17}$$

Moreover, in (3) we fix

$$v_{\text{AI}} = 5\,\text{Hz}, \quad w_c^e = w_c^i = 0.01, \quad p_c = 0.025. \tag{18}$$

Finally, the parameters appearing in (4) and (5) are

$$\tau_\psi = 10\,T = 0.03, \qquad \sigma = 0.01, \qquad c_0 = 1000, \qquad w_r = 1. \tag{19}$$

In Figure 12a, we show an example of the applied stimuli and the results of the case with $\psi(0) = 0$. In Figure 12b, we provide the results of the cases with $\psi(0)$ varied from 0 to 1, where the same stimuli shown in Figure 12a were applied. In the cases where $\psi$ converges to 1, i.e., where Stimulus $A$ is chosen, the reaction time is lower ($\approx 0.1$ s) than the cases where $\psi$ converges to 0, i.e., where Stimulus $B$ is chosen ($\approx 0.18$ s). This is due to that we chose $v_A(0) > v_B(0)$ in the trials given in Figure 12a, and therefore $v_{e_A}(\bar{t}) - v_{e_B}(\bar{t}) = 7.5$ is reached in a shorter time duration compared to $v_{e_A}(\bar{t}) - v_{e_B}(\bar{t}) = -7.5$.

In Figure 12b, there is no clear difference in terms of reaction time between the cases with different initial conditions $\psi(0)$ as long as the stimuli remain the same and $\psi$ converges to the same value. It is due to the fact that, the convergence of the regulatory mechanism is rapid, therefore the competition is promoted towards the same pool and as we do not change the stimuli, the evolution of $v_{e_A}, v_{e_B}$ becomes different realizations of almost the same random processes. Consequently, the reaction times of those realizations fluctuate around the same value.

Finally, we refer to Figure 13 for the effects of varying the stimuli difference on the reaction time.

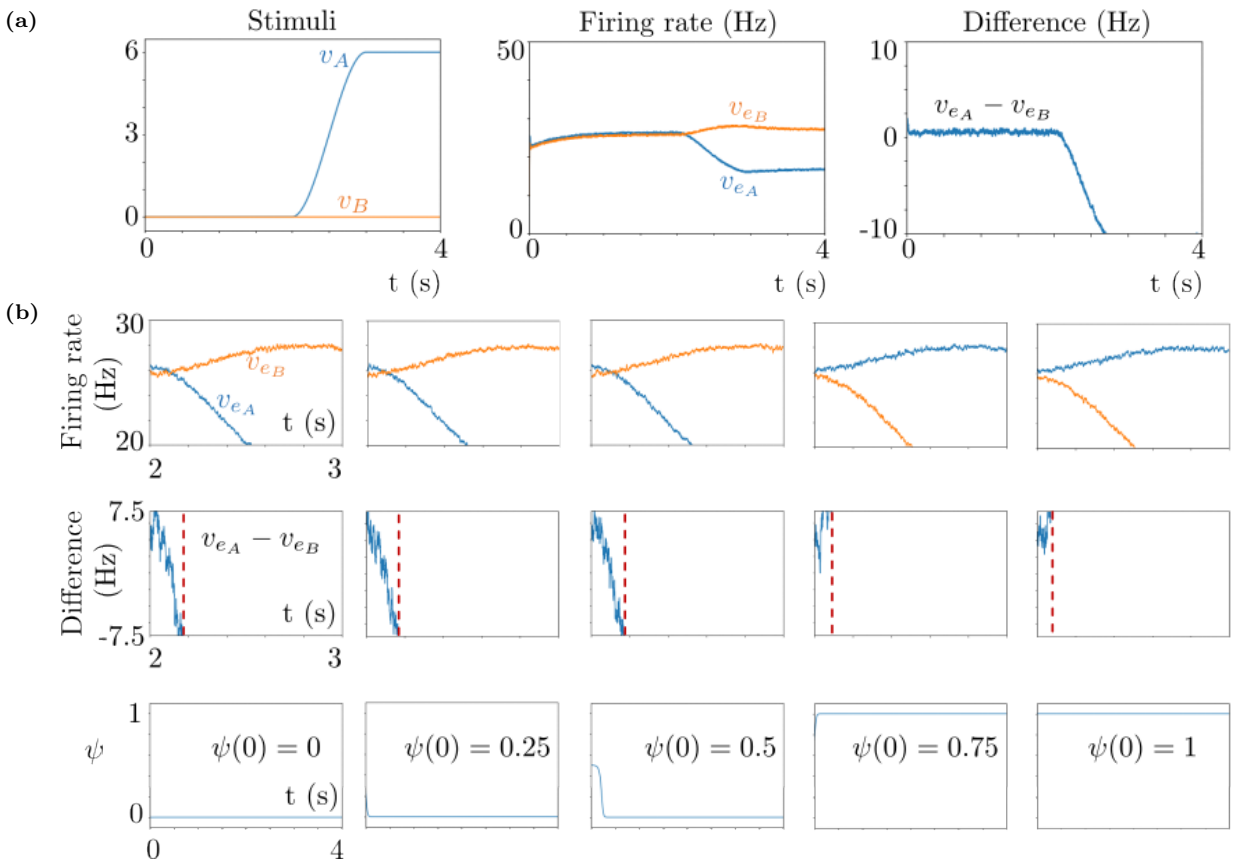


**Figure 12:** Simulation results of isolated trials, with different regulatory module initial conditions $\psi(0)$. **(a)** The result for $\psi(0) = 0$. Left: Applied stimuli $v_A$ and $v_B$. Middle: Time course of the excitatory population membrane voltages of both Pools $A$ and $B$. Right: Difference of the evolving membrane voltages in time. **(b)** The results for varied $\psi(0)$ values. Top: Time courses of $v_{e_A}$ and $v_{e_B}$. Middle: Difference of the firing rates where the red vertical line denotes the decision instant. Bottom: Time evolution of the regulatory pool $\psi(t)$. The initial values $\psi(0)$ are 0, 0.25, 0.5, 0.75, 1 from left to right.

## Effect of varying the difference between stimuli

In Figure 13, we observe the effects of varying the difference between the stimuli on the model reaction time. The reaction time is measured in the same way as in Figure 12b. We set the amplitude of Stimuli $A$ to 6 and vary the amplitude of Stimuli $B$ between 0 and 6. We keep constant ($= 0.75$) the initial value $\psi(0)$ of the regulatory function $\psi$ such that the system privileges always Stimulus $A$. We observe that the reaction time increases as the difference between the stimuli decreases. It is expected since it becomes more difficult to make a distinction between the stimuli as the difference between the filled quantities of the bars decreases.

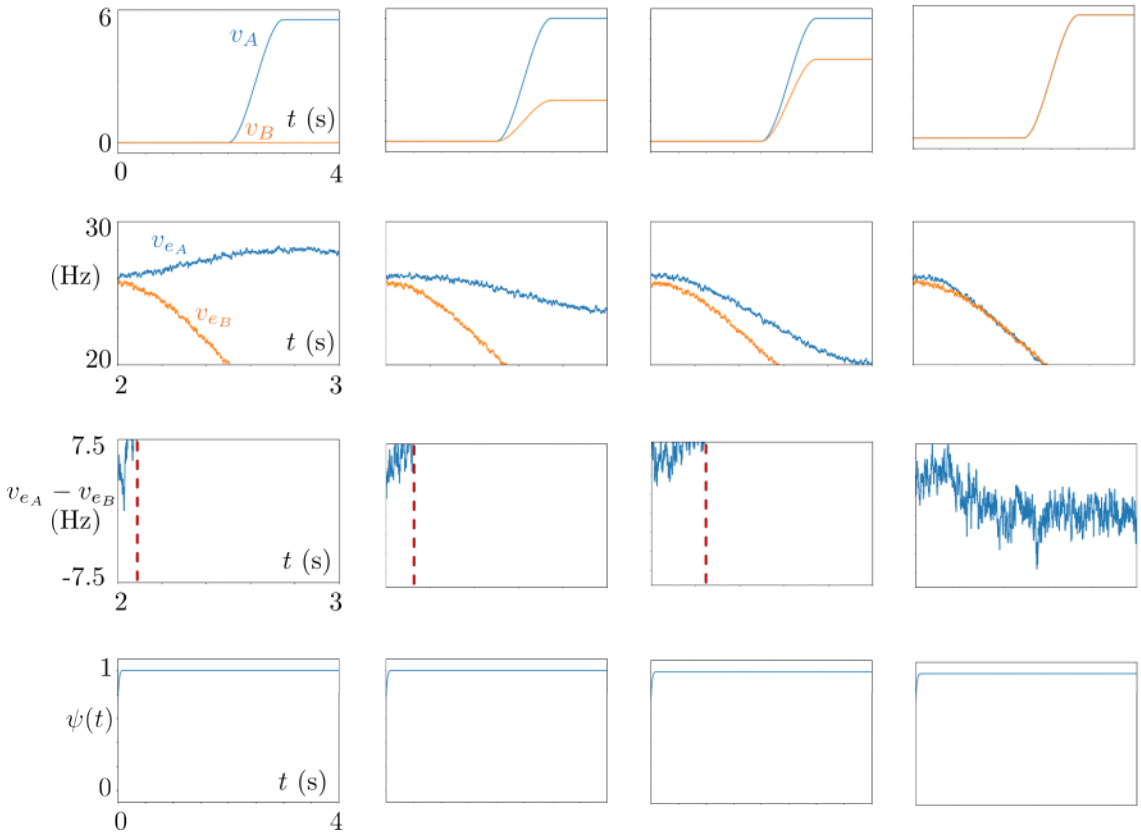


**Figure 13:** Simulation results with changing $v_B$. Top row: Time courses of the external stimuli $v_A$ and $v_B$. Second row: Time courses of $v_{e_A}$ and $v_{e_B}$. Third row: Time course of the difference between the membrane potentials where the red vertical line denotes the decision instant. Bottom row: Time course of the regulatory function $\psi$. The initial value $\psi(0)$ is 0.75 in all plots.

## Comparison to the Horizon 0 human performance

In Figure 14, we provide the results of the Horizon 0 simulations similarly to the Horizon 1 results presented in Figure 8. The red horizontal lines in Figure 14 show the experiment values. We observe in the left column that the learning speed decreases both with increasing $k$ and with increasing $c_0$, resulting in smaller values of the initial episode number of the performance clusters. The parameter $k$ directly controls the learning speed, and $c_0$ contributes to it by avoiding the deviation blocks which might break the performance clusters. In both cases, the decrease in the mean value profiles is convex. We see that in both cases, the model can reproduce close statistics once $k$ and $c_0$ are chosen properly, for example $k = 0.3$ and $c_0 = 1.1$.

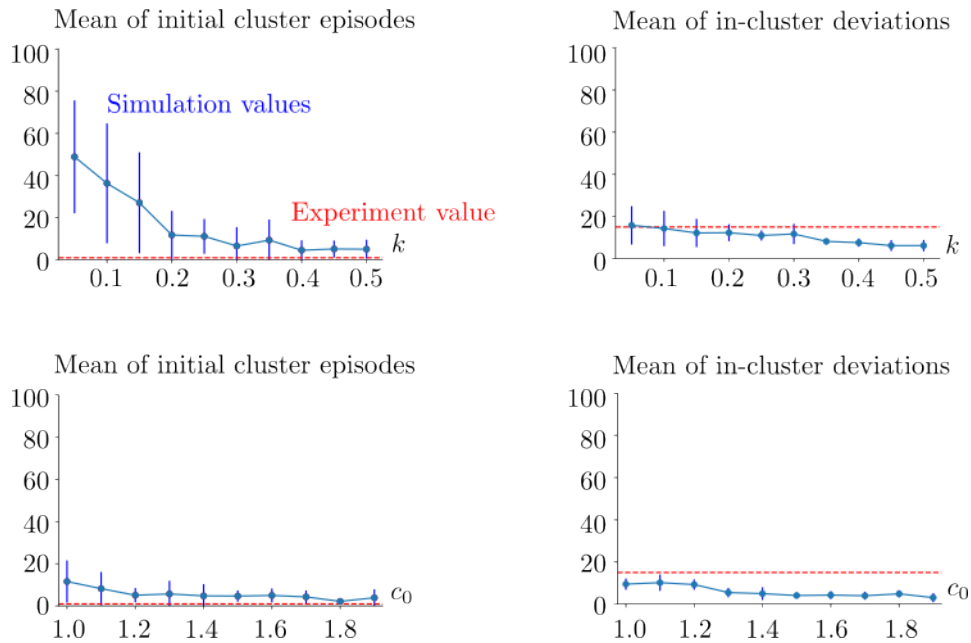


**Figure 14:** Horizon 0 human case. Simulation statistics with respect to the varied learning speed $k$ in the top row, and with respect to the varied flexibility parameter $c_0$ in the bottom row. In the top row, $c_0 = 1.1$. In the bottom row, $k = 0.3$. The blue vertical lines show the standard deviations of the corresponding statistical sample. The red dashed lines are the experiment values. The mean and standard deviations are obtained via repeating the simulation 10 times for each $(k, c_0)$ pair. See (14)-(16) for the rest of the parameters.

## Comparison to the Horizon 0 human reaction times

In Figure 15, we provide the histograms of the reaction times obtained from Horizontal 0 simulations with the varied learning speed $k$ and flexibility parameter $c_0$. We observe that there is no distinguishing change in the histograms as we vary $k$ and $c_0$, suggesting that these two parameters do not have noticeable effects on the reaction times.

In Figure 16, we show the simulation and experiment histograms of the reaction times corresponding to Horizon 0, with the fitted distributions for the best performance $k$ and $c_0$ values as in the Horizon 1 case. The best fit is achieved with skewrow distribution and hypergeometric distribution for the experiment and simulation histograms, respectively. Squared sum errors are 13e-6 and 12e-6 for the experiment and simulation histograms, respectively. Pearson correlation coefficient between the fitted distributions is 0.91.

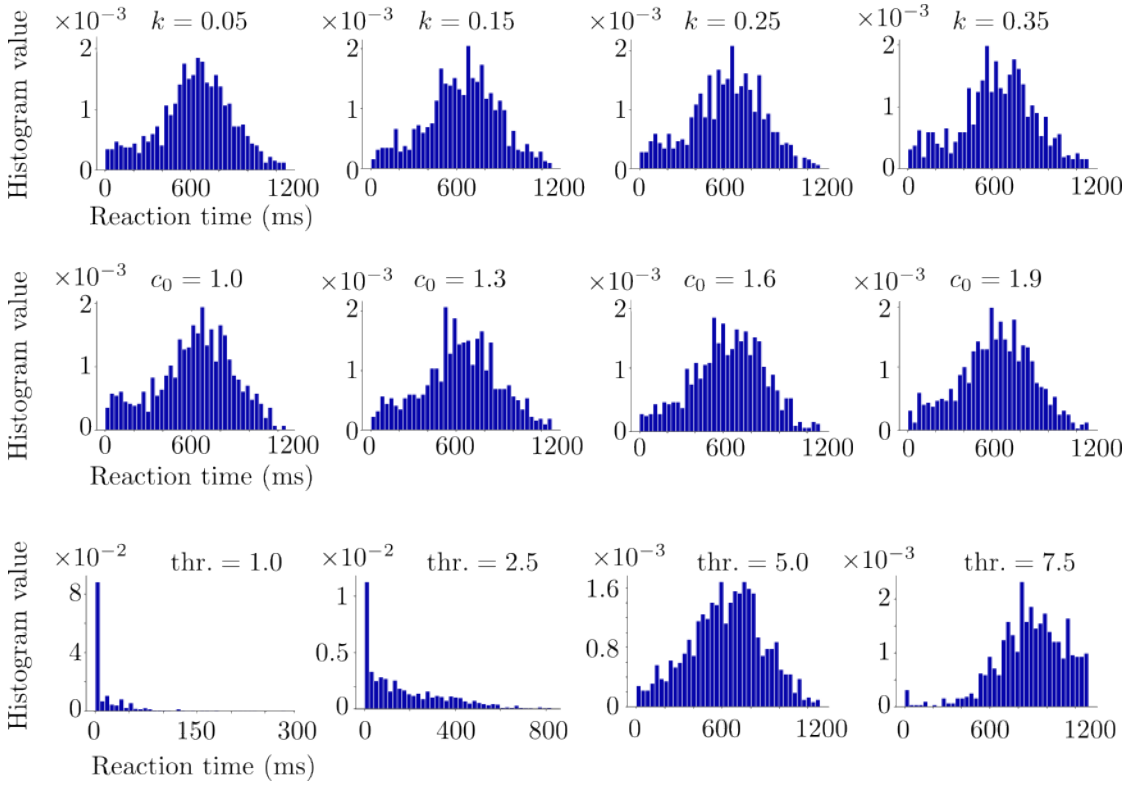


**Figure 15:** Horizon 0 reaction time histograms. Top: The histograms of the varied $k$, with $c_0 = 1.1$, decision threshold = 5Hz. Middle: The histograms of the varied $c_0$, with $k = 0.3$, decision threshold = 5 Hz. Bottom: The histograms of the varied decision threshold, where $k = 0.3$, $c_0 = 1.1$. The histograms are obtained for each parameter from 10 realizations of the same simulation setup. The parameters are the same as those of Figure 14.

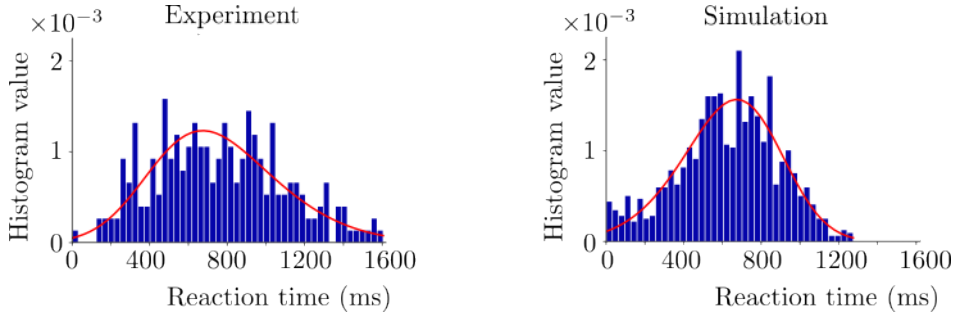


**Figure 16:** Horizon 0 human case. Histograms obtained from the human experiment and simulations for comparison. The red curves denote the fitted distributions via Python. The simulations are performed with $k = 0.3$, $c_0 = 1.1$ and decision threshold = 0.5.

### Comparison to the macaque reaction times

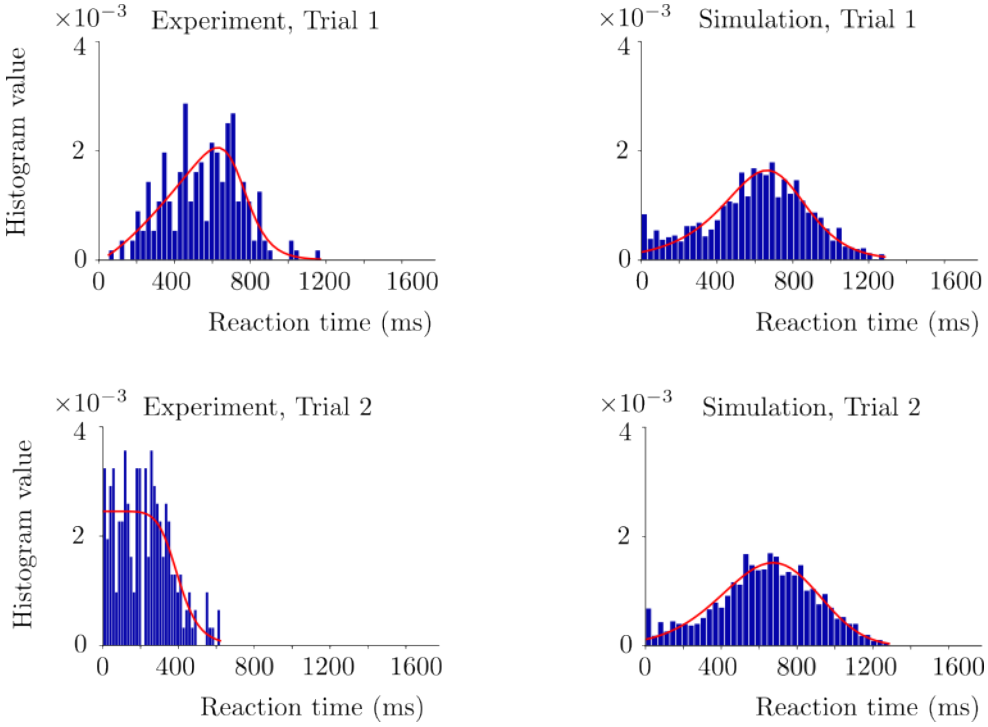


**Figure 17:** Horizon 1 macaque case. Histograms obtained from the experiment and simulations for comparison. The red curves denote the fitted distributions via Python. Trial 1 and 2 histograms are in the top and bottom rows, respectively. The simulations are performed with $k = 0.12$, $c_0 = 0.106$ and decision threshold $= 5$. The rest of the parameters is found in (14)-(16).

In Figure 17, we show the simulation and experiment histograms of the reaction times corresponding to the macaque experiment, together with the fitted distributions for the best performance $k$ and $c_0$ values. We fit distributions to both simulation and experiment histograms by using the *Fitter* function from the Python *Fitter* package as in the case of the human experiment. In Trial 1, genlogistic distribution and powernorm distribution are fitted to the experiment and simulation histograms, respectively. Corresponding squared sum error values are 1e-5 and 9e-6 for the experiment and simulation histograms, respectively. Pearson correlation coefficient of these fitted distributions is 0.97. In Trial 2 histograms, burr distribution and kappa3 distribution are fitted to the experiment and simulation histograms, respectively. Corresponding squared sum error values are 9.9e-5 and 5.73e-4 for the experiment and simulation histograms, respectively. Pearson correlation coefficient of these fitted distributions is 0.31. Trial 2 has a weaker overlap between the experiment and simulation histograms compared to Trial 1 as in the case of the human task.

## 8 Acknowledgments

We hereby acknowledge that this research was supported by the Human Brain Project (European Union grant H2020-945539).

## References

[1] A. K. Churchland, R. Kiani, and M. N. Shadlen, "Decision-making with multiple alternatives," *Nature Neuroscience*, vol. 11, no. 6, pp. 693–702, 2008.

[2] E. Marcos, P. Pani, E. Brunamonti, G. Deco, S. Ferraina, and P. Verschure, "Neural variability in premotor cortex is modulated by trial history and predicts behavioral performance," *Neuron*, vol. 78, no. 2, pp. 249–255, 2013.

[3] A. J. Parker and W. T. Newsome, "Sense and the single neuron: probing the physiology of perception," *Annual Review of Neuroscience*, vol. 21, no. 1, pp. 227–277, 1998.

[4] R. Romo and E. Salinas, "Touch and go: decision-making mechanisms in somatosensation," *Annual Review of Neuroscience*, vol. 24, no. 1, pp. 107–137, 2001.

[5] J. D. Schall, "Neural basis of deciding, choosing and acting," *Nature Reviews Neuroscience*, vol. 2, no. 1, pp. 33–42, 2001.

[6] J. Drugowitsch, R. Moreno-Bote, A. K. Churchland, M. N. Shadlen, and A. Pouget, "The cost of accumulating evidence in perceptual decision making," *Journal of Neuroscience*, vol. 32, no. 11, pp. 3612–3628, 2012.

[7] G. Mochol, R. Kiani, and R. Moreno-Bote, "Prefrontal cortex represents heuristics that shape choice bias and its integration into future behavior," *Current Biology*, vol. 31, no. 6, pp. 1234–1244, 2021.

[8] A. Bechara, D. Tranel, and H. Damasio, "Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions," *Brain*, vol. 123, no. 11, pp. 2189–2202, 2000.

[9] A. N. Hampton, P. Bossaerts, and J. P. O'doherty, "The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans," *Journal of Neuroscience*, vol. 26, no. 32, pp. 8360–8367, 2006.

[10] M. F. Rushworth, M. P. Noonan, E. D. Boorman, M. E. Walton, and T. E. Behrens, "Frontal cortex and reward-guided learning and decision-making," *Neuron*, vol. 70, no. 6, pp. 1054–1069, 2011.

[11] M. N. Shadlen and R. Kiani, "Decision making as a window on cognition," *Neuron*, vol. 80, no. 3, pp. 791–806, 2013.

[12] P. Domenech and E. Koechlin, "Executive control and decision-making in the prefrontal cortex," *Current Opinion in Behavioral Sciences*, vol. 1, pp. 101–106, 2015.

[13] M. P. Paulus, C. Rogalsky, A. Simmons, J. S. Feinstein, and M. B. Stein, "Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism," *Neuroimage*, vol. 19, no. 4, pp. 1439–1448, 2003.

[14] A. G. Sanfey, G. Loewenstein, S. M. McClure, and J. D. Cohen, "Neuroeconomics: cross-currents in research on decision-making," *Trends in Cognitive Sciences*, vol. 10, no. 3, pp. 108–116, 2006.

[15] Z. Guo, J. Chen, S. Liu, Y. Li, B. Sun, and Z. Gao, "Brain areas activated by uncertain reward-based decision-making in healthy volunteers," *Neural Regeneration Research*, vol. 8, no. 35, p. 3344, 2013.

[16] Y. Broche-Pérez, L. H. Jiménez, and E. Omar-Martínez, "Neural substrates of decision-making," *Neurología (English Edition)*, vol. 31, no. 5, pp. 319–325, 2016.

[17] I. E. Monosov and O. Hikosaka, "Regionally distinct processing of rewards and punishments by the primate ventromedial prefrontal cortex," *Journal of Neuroscience*, vol. 32, no. 30, pp. 10318–10330, 2012.

[18] S. Suzuki, V. M. Lawlor, J. A. Cooper, A. R. Arulpragasam, and M. T. Treadway, "Distinct regions of the striatum underlying effort, movement initiation and effort discounting," *Nature human behaviour*, vol. 5, no. 3, pp. 378–388, 2021.

[19] A. Sirigu and J.-R. Duhamel, "Reward and decision processes in the brains of humans and nonhuman primates," *Dialogues in Clinical Neuroscience*, 2022.

[20] R. Moreno-Bote, J. Rinzel, and N. Rubin, "Noise-induced alternations in an attractor network model of perceptual bistability," *Journal of Neurophysiology*, vol. 98, no. 3, pp. 1125–1139, 2007.

[21] L. Albantakis and G. Deco, "Changes of mind in an attractor network of decision-making," *PLoS Comput Biol*, vol. 7, no. 6, p. e1002086, 2011.

[22] B. Y. Hayden and R. Moreno-Bote, "A neuronal theory of sequential economic choice," *Brain and Neuroscience Advances*, vol. 2, p. 2398212818766675, 2018.

[23] N. Brunel and X.-J. Wang, "Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition," *Journal of Computational Neuroscience*, vol. 11, no. 1, pp. 63–85, 2001.

[24] X.-J. Wang, "Probabilistic decision making by slow reverberation in cortical circuits," *Neuron*, vol. 36, no. 5, pp. 955–968, 2002.

[25] C. D. Gilbert and T. N. Wiesel, "Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex," *Nature*, vol. 280, no. 5718, pp. 120–125, 1979.

[26] K. Martin and D. Whitteridge, "Form, function and intracortical projections of spiny neurones in the striate visual cortex of the cat.," *The Journal of Physiology*, vol. 353, no. 1, pp. 463–504, 1984.

[27] Z. F. Kisvárday, K. A. Martin, T. Freund, Z. Magloczky, D. Whitteridge, and P. Somogyi, "Synaptic targets of hrp-filled layer iii pyramidal cells in the cat striate cortex," *Experimental Brain Research*, vol. 64, no. 3, pp. 541–552, 1986.

[28] B. A. McGuire, C. D. Gilbert, P. K. Rivlin, and T. N. Wiesel, "Targets of horizontal connections in macaque primary visual cortex," *Journal of Comparative Neurology*, vol. 305, no. 3, pp. 370–392, 1991.

[29] Z.-C. Xiao, K. K. Lin, and L.-S. Young, "A data-informed mean-field approach to mapping of cortical parameter landscapes," *PLoS Computational Biology*, vol. 17, no. 12, p. e1009718, 2021.

[30] H. R. Wilson and J. D. Cowan, "Excitatory and inhibitory interactions in localized populations of model neurons," *Biophysical Journal*, vol. 12, no. 1, pp. 1–24, 1972.

[31] S.-i. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," *Biological Cybernetics*, vol. 27, no. 2, pp. 77–87, 1977.

[32] S. El Boustani and A. Destexhe, "A master equation formalism for macroscopic modeling of asynchronous irregular activity states," *Neural Computation*, vol. 21, no. 1, pp. 46–100, 2009.

[33] Y. Zerlaut, S. Chemla, F. Chavane, and A. Destexhe, "Modeling mesoscopic cortical dynamics using a mean-field model of conductance-based networks of adaptive exponential integrate-and-fire neurons," *Journal of Computational Neuroscience*, vol. 44, no. 1, pp. 45–61, 2018.

[34] M. Di Volo, A. Romagnoni, C. Capone, and A. Destexhe, "Biologically realistic mean-field models of conductance-based networks of spiking neurons with adaptation," *Neural Computation*, vol. 31, no. 4, pp. 653–680, 2019.

[35] E. Baspinar, G. Cecchini, R. Moreno-Bote, I. Cos, and A. Destexhe, "Jupyter notebook of a biophysically plausible decision-making model based on interacting cortical columns," *Zenodo*, 2023.

[36] G. Cecchini, M. DePass, E. Baspinar, M. Andujar, S. Ramawat, P. Pani, S. Ferraina, A. Destexhe, R. Moreno-Bote, and I. Cos, "A theoretical formalization of consequence-based decision-making," *bioRxiv*, 2023.

[37] E. Seidemann, E. Zohary, and W. T. Newsome, "Temporal gating of neural signals during performance of a visual discrimination task," *Nature*, vol. 394, no. 6688, pp. 72–75, 1998.

[38] R. Kiani, T. D. Hanks, and M. N. Shadlen, "Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment," *Journal of Neuroscience*, vol. 28, no. 12, pp. 3017–3029, 2008.

[39] C. D. Kopec, J. C. Erlich, B. W. Brunton, K. Deisseroth, and C. D. Brody, "Cortical and subcortical contributions to short-term memory for orienting movements," *Neuron*, vol. 88, no. 2, pp. 367–377, 2015.

[40] F. Van Ede, S. R. Chekroud, M. G. Stokes, and A. C. Nobre, "Decoding the influence of anticipatory states on visual perception in the presence of temporal distractors," *Nature Communications*, vol. 9, no. 1, pp. 1–12, 2018.

[41] Y. Zuo and M. E. Diamond, "Rats generate vibrissal sensory evidence until boundary crossing triggers a decision," *Current Biology*, vol. 29, no. 9, pp. 1415–1424, 2019.

[42] A. Finkelstein, L. Fontolan, M. N. Economo, N. Li, S. Romani, and K. Svoboda, "Attractor dynamics gate cortical information flow during decision-making," *Nature Neuroscience*, pp. 1–8, 2021.

[43] L. Arnold, "Stochastic differential equations," *New York*, 1974.

[44] J. S. Goldman, L. Kusch, B. H. Yalcinkaya, D. Depannemaecker, T.-A. E. Nghiem, V. Jirsa, and A. Destexhe, "Brain-scale emergence of slow-wave synchrony and highly responsive asynchronous states based on biologically realistic population models simulated in the virtual brain," *bioRxiv*, 2020.

[45] K. Doya, S. Ishii, A. Pouget, and R. P. Rao, *Bayesian brain: Probabilistic approaches to neural coding*. MIT press, 2007.

[46] L. A. Leotti and T. D. Wager, "Motivational influences on response inhibition measures.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 36, no. 2, p. 430, 2010.

[47] M. Giamundo, F. Giarrocco, E. Brunamonti, F. Fabbrini, P. Pani, and S. Ferraina, "Neuronal activity in the premotor cortex of monkeys reflects both cue salience and motivation for action generation and inhibition," *Journal of Neuroscience*, vol. 41, no. 36, pp. 7591–7606, 2021.

[48] S. Padmala and L. Pessoa, "Interactions between cognition and motivation during response inhibition," *Neuropsychologia*, vol. 48, no. 2, pp. 558–565, 2010.