Neural Networks 154 (2022) 283–302

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Brain-inspired meta-reinforcement learning cognitive control in conflictual inhibition decision-making task for artificial agents

Federica Robertazzi *, Matteo Vissani, Guido Schillaci, Egidio Falotico

The BioRobotics Institute, Sant'Anna School of Advanced Studies, Pisa, 56127, Italy
Department of Excellence in Robotics & AI, Sant'Anna School of Advanced Studies, Pisa, 56127, Italy

ARTICLE INFO

Article history:
Received 22 August 2021
Received in revised form 9 June 2022
Accepted 16 June 2022
Available online 22 June 2022

Keywords:
Meta-learning
Brain-inspired modeling
Inhibition cognitive control
Basal ganglia
Prefrontal cortex

ABSTRACT

Conflictual cues and unexpected changes in human real-case scenarios may be detrimental to the execution of tasks by artificial agents, thus affecting their performance. Meta-learning applied to reinforcement learning may enhance the design of control algorithms, where an outer learning system progressively adjusts the operation of an inner learning system, leading to practical benefits for the learning schema. Here, we developed a brain-inspired meta-learning framework for inhibition cognitive control that i) exploits the meta-learning principles in the neuromodulation theory proposed by Doya, ii) relies on a well-established neural architecture that contains distributed learning systems in the human brain, and iii) proposes optimization rules of meta-learning hyperparameters that mimic the dynamics of the major neurotransmitters in the brain. We tested an artificial agent in inhibiting the action command in two well-known tasks described in the literature: NoGo and Stop-Signal Paradigms. After a short learning phase, the artificial agent learned to react to the hold signal, and hence to successfully inhibit the motor command in both tasks, via the continuous adjustment of the learning hyperparameters. We found a significant increase in global accuracy, right inhibition, and a reduction in the latency time required to cancel the action process, i.e., the Stop-signal reaction time. We also performed a sensitivity analysis to evaluate the behavioral effects of the meta-parameters, focusing on the serotoninergic modulation of the dopamine release. We demonstrated that brain-inspired principles can be integrated into artificial agents to achieve more flexible behavior when conflictual inhibitory signals are present in the environment.

1. Introduction

Cognitive control is the ability to adapt behavior to current demands by promoting relevant information regardless of interference from the internal urge (Dreher & Berman, 2002). The disengagement from a motor response – action inhibition – is an important component in the transition between cognitive demands. This inhibition can either be a simple rest period before continuing the same task, or switch to a more compelling command. In real-world applications, critical circumstances may require artificial agents to counter the effect of unpredictable signal alerts as they carry out their duties. However, although modern artificial agents are extremely reliable and accurate when learning a single task after long exposure to stationary learning trials, they are not usually characterized by reactive behavior that adapts robustly to the unpredictability of the world (Pfeifer, Lungarella, & Iida, 2007; Spelke & Kinzler, 2007).

Meta-learning (ML), also known as "learning-to-learn", could lead to the development of control algorithms in which the agent optimizes for the maximal reward with increasing flexibility and efficiency (Baxter, 1998, 1998; Botvinick et al., 2019; Duan et al., 2016; Schmidhuber, Zhao, & Wiering, 1996; Wang et al., 2017). The elementary block of an ML system consists of one learning system – the outer loop – which progressively adjusts the operation of a second learning system – the inner loop. Machine learning algorithms can generalize the learning capability of artificial agents through: (i) knowledge transfer across a joint distribution of related tasks, environments, and objects; (ii) building on previous experience which increases learning speed; (iii) the generation of useful shifts in inductive bias by adjusting the current strategy.

In a recent review (Wang, 2020), Wang and colleagues proposed a classification of ML approaches based on the following categories: learning of mental schemas and hierarchical structures (Kesteren, Ruiter, Fernández, & Henson, 2012; Tse et al., 2007), learning of Bayesian priors and latent states (Lake, Salakhutdinov, & Tenenbaum, 2015; Solway et al., 2014), and learning of

* Corresponding author at: The BioRobotics Institute, Sant'Anna School of Advanced Studies, Pisa, 56127, Italy.
E-mail addresses: federica.robertazzi@santannapisa.it (F. Robertazzi), MVISSANI@mgh.harvard.edu (M. Vissani), guido.schillaci@santannapisa.it (G. Schillaci), egidio.falotico@santannapisa.it (E. Falotico).

https://doi.org/10.1016/j.neunet.2022.06.020

meta-parameters. ML in terms of being the learning of mental schemas is a learning control based on the consolidation of existing knowledge, i.e., structured hierarchical representations, and its effect on the learning of new information (Badre, 2008; Badre, Kayser, & D'Esposito, 2010; Collins & Frank, 2013). ML in terms of being the learning of Bayesian priors and latent states, proposes a framework where hierarchical priors are constructed using similar previous experiences which are used for the inference of probabilistic learning programs of new examples (Lake et al., 2015). ML as the learning of meta-parameters may be the most straightforward application as it is aimed at adjusting the parameters, i.e., meta-parameters or hyperparameters, of the learning process itself. This meta-learning approach has stimulated interest in the machine learning field as it reduced the hand-tuning of parameters and generalization errors, which thereby improves performance (Wang, 2020; Xu et al., 2020; Xu, Hasselt, & Silver, 2018).

Neurophysiological knowledge from extensive studies on cognitive control and learning in non-human and human primates can inspire the design of ML principles for robotic applications. Many studies have highlighted the existence of distributed learning modules in the brain, e.g., certain areas of the prefrontal cortex (PFC) and basal ganglia (BG) (Caligiore, Arbib, Miall, & Baldassarre, 2019). They implement global and local meta-parameter computations by the signaling and mediation of the four major neurotransmitters, e.g., dopamine, serotonin, noradrenaline, and acetylcholine (Doya, 2000, 2002; Schweighofer & Doya, 2003).

PFC and BG have the potential to implement reward-based learning mechanisms (Padoa-Schioppa & Assad, 2006; Rushworth & Behrens, 2008; Wang et al., 2018). In fact, the temporal profile of PFC neuron activity and the synaptic plasticity in the striatum dynamically compute and encode the expected action/state values and the history of rewards and choices (Daw, Niv, & Dayan, 2005; Daw & Tobler, 2014; Seo, Lee, & Averbeck, 2012; Tsutsui, Grabenhorst, Kobayashi, & Schultz, 2016; Wang et al., 2018).

The interdependence and relationship between these two distributed learning systems (i.e., PFC and BG) are still not completely understood. Dopaminergic projections (D_1 and D_2) from the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc) innervate the anterior cingulate cortex (ACC) (mesocorticolimbic pathway) and the striatum (nigrostriatal pathway), subserving the encoding of the difference between the estimated reward at any given state (i.e., the action value) and the actual reward received, i.e., the temporal difference error (Kim et al., 2020; Poulin et al., 2018; Starkweather & Uchida, 2021; Sutton & Barto, 2018).

D_1 and D_2 dopamine receptors mediate different functional mechanisms, and act prevalently in the opposite way to each other. The D_1/ D_2 balance thus regulates the interaction between the cortex and the basal ganglia circuitry, therefore playing a pivotal role in motor control, action selection, action switching, and action inhibition (Lapidus, Stern, Berlin, & Goodman, 2014). The ACC tracks and memorizes the action values, and dynamically meta-controls the trade-off between exploration and exploitation in reward-based tasks (Akam et al., 2021; Khamassi, Enel, Dominey, & Procyk, 2013). Serotonergic neurons are located deep in the midbrain, and send projections to extensive regions of the brain.

Serotonin ([γ]) is a major neuromodulator implicated in several physiological and behavioral functions such as impulsivity, behavioral inhibition, and self-control (Berger, Gray, & Roth, 2009; Fischer & Ullsperger, 2017). Depletion of serotonin may be associated with a lack of inhibitory control as such a reduced willingness to wait for a large reward, i.e., higher temporal discount factor (Ranade, Pi, & Kepecs, 2014). Although the interplay between serotonin and dopamine has not been fully understood (De Deurwaerdère, Chagraoui, & Di Giovanni, 2021; Fischer & Ullsperger, 2017), the blockade of [γ] receptors in humans increases the sensitivity of trial-to-trial punishment (Chamberlain, 2006). This potentially suggests the computational role of serotonin in inhibiting actions via the opposing process with dopamine (Daw, Kakade, & Dayan, 2002). Noradrenergic neurons located in the locus coeruleus increase coherently their firing rate with the accuracy of action selection, prompting the basis for dealing with the exploration/exploitation dilemma (Usher, Cohen, Servan-Schreiber, Rajkowski, & Aston-Jones, 1999). Higher levels of reuptake blocking of noradrenaline receptors result in a drastic reduction in the stochasticity of action selection and stereotyped behaviors.

Khamassi and coll. (Khamassi et al., 2013) implemented a cognitive control architecture for agent–environment interactions, such as action selection. Khamassi's model reproduces the interaction between the prefrontal cortex and basal ganglia, as well as the reinforcement learning update rules to encode the sensitivity of trial-to-trial punishments/rewards. This approach also included the dynamic regulation of exploration/exploitation rates based on the outcome history (i.e., vigilance Dehaene, Kerszberg, & Changeux, 1998) where uncertain events shift the system towards exploration. This model was tested in situations where environmental uncertainties were introduced by humans such as changing targets or cheating.

In this work, we developed a cognitive inhibition control framework for artificial agents by implementing ML brain-inspired rules that draw inspiration from the action of the neuromodulators in the human brain (Doya, 2002). Using the robotic model in Khamassi' s paper, we developed a cognitive computational framework to countermand inhibition signals that interfere with cortical motor commands that are planned or ongoing.

Briefly, we extended and adapted their framework by implementing: (i) neural dynamics and action values for the inhibition reaction to hold signals during the tasks, (ii) a different meta-learning control policy that dynamically modulates the level of the exploration/exploitation rate β in the action selection process, based on the current phasic dopaminergic release of D_1, (iii) a gain modulation of striatal output by means of a tonic D_2-based shift of the excitation properties of striatal neurons, and (iv) a serotoninergic modulation of the overall dopamine level (that is, D_1 and D_2).

We tested our model in two forms of action inhibition: action restraint and action cancellation (Eagle, Bari, & Robbins, 2008; Mosher, Mamelak, Malekmohammadi, Pouratian, & Rutishauser, 2021a; Pasquereau & Turner, 2017a; Schall, Palmeri, & Logan, 2017). The former refers to the inhibition of the motor response before the response has started, which was investigated through the NoGo Paradigm. The latter describes the inhibition of a motor response during its execution, and was studied using the Stop-Signal Paradigm where the delay between the onset of the response and the appearance of the hold signal plays a fundamental role. The artificial agent successfully adjusted the hyperparameter of the learning process through reinforcement learning and brain-inspired meta-learning principles, thus demonstrating the feasibility of this approach for robotic applications.

We focus on the dynamic regulation of the exploration/exploitation rate and the effect of the serotonin–dopamine interaction on the inhibition of cognitive control.

2. Materials and methods

We implemented a brain-inspired meta-learning framework to inhibit cognitive control for artificial agents which incorporates meta-learner representations in distributed regions of the brain, ranging from cortical (e.g., PFC) to subcortical (e.g., BG) areas. We tested the behavior of the artificial agent in two well-known conflictual tasks (e.g., NoGo Paradigm and Stop-Signal Paradigm), in which the agent was challenged to countermand an inhibitory hold signal at different time delays after the onset of the stimulus.

Table 1
Neuron dynamics.

Neuron dynamics									
Brain regions	N	Δt	τ	X_0	C_0	S_0	I_{tonic}	W_{source}	I_{thr}
PPC	3	0.1	0.5	0	0.6	5			0.75
ACC	3	0.1	0.6	0	0.6	5		+1 (PPC) +1 (PMC)	
VTA	1	0.1						+1 (ACC)	
LPFC	3	0.1	0.5	0	0.6	5		+1 (PPC) +1 (ACC)	
PMC	3	0.1	0.25	0	0.6	5		+1 (Thalamus)	
SNc	1	0.1							
Striatum	3	0.1	0.5	0	0.6	5		+1 (PPC) +1 (LPFC) +1 (PMC)	
SNr	3	0.1	0.5	0	0.6	5	1	−1 (Striatum)	
Thalamus	3	0.1	0.5	0	0.6	5	1	−1 (SNr)	

N (number of neurons in each layer), Δt (timestep), τ (time constant), X_0 (initial value neuron's firing activity), C_0 (sigmoid central value), S_0 (sigmoid slope),I_{tonic} (tonic input), W_{source} (weights inputted to the layers), and I_{thr} (threshold input).

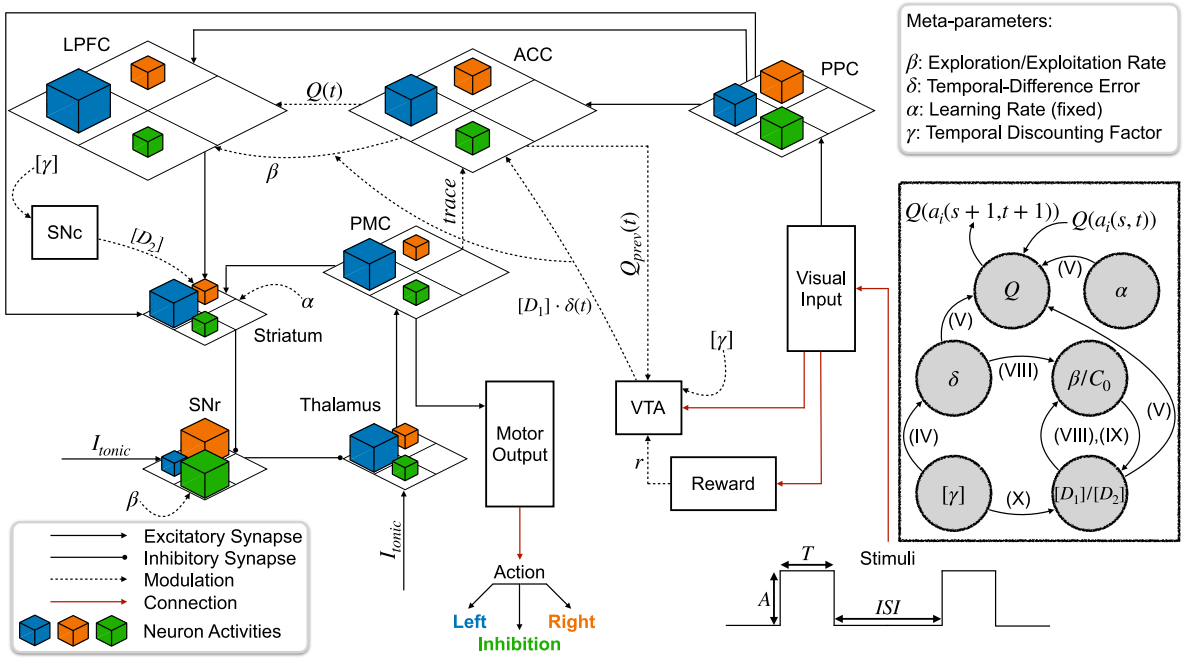

Fig. 1. The model architecture inspired by (Khamassi et al., 2011) and the meta-learning mechanism based on the principles of Doya's neuromodulation theory (Doya, 2002; Schweighofer & Doya, 2003). Each layer represents a brain region whose size is proportional to the biological dimension, preserving the ratio across brain areas. The volume of the cubes displays the intensity of the neuron's activity, topographically associated with the two directions (Left (blue), Right (orange)) and action inhibition (Inhibition (Green)). Excitatory (black arrow) and inhibitory (black circle arrow) neural synapses, reinforcement learning, meta-learning mechanisms (e.g., action values, dopamine, serotonin, noradrenaline, etc.) (black dashed arrows), and input/output connections (red lines) are displayed. Stimuli are fed in the model by simulating a square wave (A (amplitude) = 1 [a.u.], T (duration) = 100 [samples], ISI (interstimulus interval) = 200 [samples]) for neurons codifying Left or Right movements. On the right, the update rule for the action value Q and relationships among the parameters of the model are illustrated. Equations are ordered in the main text using Latin numbers. Abbreviations: posterior parietal cortex (PPC), anterior cingulate cortex (ACC), ventral tegmental area (VTA), lateral prefrontal cortex (LPFC), substantia nigra compacta (SNc), substantia nigra reticulata (SNr), and premotor cortex (PMC).

2.1. Model architecture

The model architecture (Khamassi, Lallée, Enel, Procyk, & Dominey, 2011) is made up of modules representing different brain regions: the prefrontal cortex (e.g., posterior parietal cortex (PPC), anterior cingulate cortex (ACC), ventral tegmental area (VTA), lateral prefrontal cortex (LPFC), premotor cortex (PMC)) and basal ganglia circuitry (e.g., striatum, substantia nigra reticulata (SNr), thalamus, substantia nigra compacta (SNc)) (Fig. 1, Table 1).

Anatomical connections in the model follow the neurophysiological evidence in studies on nonhuman and human primates (Alexander & Crutcher, 1990; Alexander, DeLong, & Strick, 1986; Amiez, 2006; Daw & Doya, 2006; Daw et al., 2005; Khamassi et al., 2013, 2011; Leisman, Braun-Benjamin, & Melillo, 2014; Mosher, Mamelak, Malekmohammadi, Pouratian, & Rutishauser, 2021b; Pasquereau & Turner, 2017b). Each layer is composed of three neurons that topographically codify different space regions (Middleton & Strick, 1994). In particular, we used two neurons to encode two different opposite directions, e.g., left and right,

and one neuron to encode action inhibition, i.e., refraining from responding (action restraint in NoGo Trials) and stopping an already prepared response (action cancellation in Stop-Signal Trials) (Chen et al., 2020; Heekeren, Marrett, & Ungerleider, 2008; Kaplan, Schuck, & Doeller, 2017; Mosher et al., 2021b; Pasquereau & Turner, 2017b; Schall, 2001; Shadlen & Newsome, 2001; Verbruggen et al., 2019). The neuron dynamics of the firing rate is described by the following first-order differential equation (Eq. (I)):

$$\tau \frac{dV}{dt} = -V(t) + I \tag{I}$$

where $V(t)$ is the neuron's membrane voltage, τ is the time constant of the neuron, and I is the input signal. The output of the differential equation of each neuron is mapped from 0 to 1 with a static and monotonic non-linear transfer function, i.e., a sigmoid function, as follows (Eq. (II)):

$$V_{out}(t) = \frac{1}{\left(1 + e^{-S_0(V(t)-C_0)}\right)} \tag{II}$$

where S_0 and C_0 are the slope and the central point of the sigmoid, respectively. The visual perception and the external reward modules regulate how the agent interacts with the environment in terms of the integration of visual stimuli and the release of the reward. The output module delivers the motor commands corresponding to the selected action. Stimuli are fed sequentially in the model (the overlap condition creates a motor cortex conflict in the basal ganglia) by simulating a square wave (amplitude=1 [a.u.], duration=100 [samples], interstimulus interval = 200 [samples], see Fig. 1) to the two neurons encoding the movement with a matrix of two categories (e.g., left and right) x 1000 stimuli. If the motor output activity corresponds to the visual input, the external reward is positive (+1), otherwise, a punishment signal is delivered (0). The connection between neurons is only interlayers (i.e., no intra-layer recurrent connections) mediated by one-to-one excitatory or inhibitory synapses (Nagel & Wilson, 2016). An external tonic input I_{tonic} set to 1 is fed as input to the SNr and thalamus to ensure thalamic triggering of the PMC when it is facilitated by the disynaptic disinhibition effect. The PPC receives the stimuli and projects them into the ACC, the LPFC and the striatum. The PPC does not receive any recurrent activity from the LPFC because we hypothesize that the PPC participates in the inhibition of the motor response in terms of encoding spatial visual stimuli as opposed to an executive block of fronto-basal-ganglia–thalamocortical network for inhibitory control (Wessel & Aron, 2017).

The ACC stores and updates the action value associated with each possible action that can be performed by the simulated agent. This is based on a temporal difference learning algorithm (Beninger, 1983; Daw & Doya, 2006; Sutton & Barto, 2018; Wickens, 1990; Wise & Rompre, 1989). Neurophysiological studies suggest that reinforcement learning mechanisms are performed in the ACC layer. These studies have shown that (i) the dopaminergic system projects strongly to ACC with respect to LPFC (Fluxe et al., 1974), (ii) the ACC reacts to the reward prediction error (Amiez, Joseph, & Procyk, 2005; Holroyd & Coles, 2002; Matsumoto & Hikosaka, 2007), and (iii) the ACC is involved in action value encoding (Kennerley, Walton, Behrens, Buckley, & Rushworth, 2006; Lee, Groman, London, & Jentsch, 2007; Rushworth & Behrens, 2008; Rushworth, Behrens, Rudebeck, & Walton, 2007). ACC neurons receive a filtered version of the PPC activity multiplied by the action values $Q(a_i(s,t))$, as follows (Eq. (III)):

$$I_{ACC}(a_i(s,t)) = Q(a_i(s,t)) \cdot V_{PPC}(a_i(s,t)) \tag{III}$$

where $a_i(s,t) \in \{Left, Right, Inhibition\}$ is the set of possible actions.

The action values are initialized at the beginning of the experiment, updated automatically after the action execution (salient events, such as after reward or punishment delivery) and reset after the completion of a block of trials (except for the action value encoding the action inhibition).

ACC neurons send action values to the VTA layer which computes the temporal difference error signal $\delta(t)$. VTA is a mesencephalic brain region which contains D_1 dopaminergic efferents projecting into the ACC (Carr & Sesack, 2000; Krichmar, 2008; Sesack & Pickel, 1992). The VTA computation of the $\delta(t)$ occurs only once the action is performed (salient event), as follows (Eq. (IV)):

$$\delta(t) = r - Q(a_i(s,t)) + [\gamma] \cdot max_a Q(a_i(s+1, t+1)) \tag{IV}$$

where $a_i(s,t) \in \{Left, Right, Inhibition\}$ is the possible action that can be performed, r is the reward (i.e., 1 if the right action is executed, otherwise 0) and $[\gamma]$ is the temporal discounting factor regulated by a meta-learning mechanism described in the next section. Then, the product between $\delta(t)$ and the amount of the dopamine D_1 efferents from VTA to ACC $[D_1]$, is sent to the ACC which updates the corresponding action value $Q(a_i(s,t))$, as follows (Eq. (V)):

$$Q(a_i(s,t)) \leftarrow Q(a_i(s,t)) \cdot \left(1 - F_{factor}\right) + \alpha \cdot \delta(t) \cdot [D_1] \cdot trace(a_i(s,t)) \tag{V}$$

where α is the learning rate, F_{factor} is the forgetting factor, and $\delta(t)$ has a synaptic plasticity effect on ACC neurons only when they receive an efferent copy of the performed action $trace(a_i(s,t))$; $a_i(s,t) \in \{Left, Right, Inhibition\}$ from the PMC layer (i.e., the reinforcement in ACC happens only when an action is effectively performed).

The absence of reward is thus not a punishment signal per se. Reinforcement of action-relevant contingencies is only enabled by a concomitant dopamine signaling and PMC drive in the striatum thereby blocking any sustaining appetitive behavior and favoring the learning of task-relevant action-outcome contingencies by responding to any salient event (Horvitz, 2000, 2000; Redgrave, Gurney, & Reynolds, 2008; Redgrave et al., 2008). The action values $Q(a_i(s,t))$ are sent to the LPFC layer which computes the probability of every possible action to take using the Boltzmann Softmax Function. The LFPC estimates the action probability of each action a_i, as follows (Eq. (VI)):

$$p_i = P(a_i(s,t)) = \frac{e^{(\beta \cdot Q(a_i,s))}}{\sum_{i=1}^{3} e^{(\beta \cdot Q(a_i,s))}} \tag{VI}$$

where β is the exploration/exploitation rate. If β is high, the agent implements an exploitation behavior because it increases the difference between the highest action probability among the others. If β is low, the model implements an exploratory behavior because the probability density of the actions tends to a uniform distribution. The LPFC activity is multiplied by the probability p_i in such a way the striatum receives an action probability-weighted activity from the prefrontal regions.

The basal ganglia layers transfer the information from the LFPC layer to the PMC, implementing a winner-take-all mechanism to prevent the execution of two actions at the same time (Alexander & Wickens, 1993; Berns & Sejnowski, 1996; Binas, Rutishauser, Indiveri, & Pfeiffer, 2014; Humphries, 2014). If the activity related to a non-selected action remains non-null, the competing mechanism is performed by a double inhibitory synapse, i.e., the first one between the striatum and the SNr and the second between the SNr and thalamus. The PMC layer sends as output the motor command related to the selected action as well as an efferent copy to the ACC layer for the action value computation (refer Eq. (V)) (Alexander & Crutcher, 1990; Doya, 1999).

Table 2
Parameters of the meta-learning and reinforcement learning.

Reinforcement learning and Meta-learning	
α	0.9
β_0	0.25
β_{reset}	0.2
β_{hold}	9
$[\gamma]$	0.5
$\left[k_{a[D_1]}, k_{b[D_1]}, k_{c[D_1]}\right]$	[10, 9.5, 0.5]
$\left[k_{a[D_2]}, k_{b[D_2]}, k_{c[D_2]}\right]$	[1.2, 0.67, 0.5]
$\left[k_{1[D_1](VTA)}, k_{2[D_1](VTA)}\right]$	[1.44, 1]
$\left[k_{3[D_2](SN_C)}, k_{4[D_2](SN_C)}\right]$	[1.44, 1]
$\left[a_{Q(hold)}, b_{Q(hold)}\right]$	[0.012, 0.2]
$\left[a_{\tau(hold)}, b_{\tau(hold)}\right]$	[0.008, 0.25]
F_{factor}	$F_{factor(Right)} = F_{factor(Left)} = 0.1$ $F_{factor(Inhibition)} = 0.01$
M_{thr}	-0.25
η	0.1

The table contains the parameters of the reinforcement learning and meta-learning mechanisms: α (learning rate), β_0 (initial value exploration/exploitation rate), β_{reset} (reset value exploration/exploitation rate), β_{hold} (hold value exploration/exploitation rate), $[\gamma]$ (temporal discounting factor as serotonin), $\left[k_{a[D_1]}, k_{b[D_1]}, k_{c[D_1]}\right]$ (parameters in $[D_1]$ and β relationship (Eq. (VIII))), $\left[k_{a[D_2]}, k_{b[D_2]}, k_{c[D_2]}\right]$ (parameters in $[D_2]$ and sigmoid central value C_0 relationship (Eq. (IX))), $\left[k_{1[D_1](VTA)}, k_{2[D_1](VTA)}\right]$ (parameters in serotonin $[\gamma]$ and dopamine $[D_1]$ relationship in VTA) and $\left[k_{3[D_2](SN_C)}, k_{4[D_2](SN_C)}\right]$ (parameters in serotonin $[\gamma]$ and dopamine $[D_2]$ relationship in SNc) (Eq. (X)), $\left[a_{Q(hold)}, b_{Q(hold)}\right]$ (slope and intercept of the linear relationships between delay d and action values reset Q_{reset}), $\left[a_{\tau(hold)}, b_{\tau(hold)}\right]$ (slope and intercept of the linear relationships between delay d and time constant τ), F_{factor} (forgetting factor), M_{thr} (hold signal meta-value threshold) and η (update rate of the meta-values for both directions and hold signal).

2.2. Meta-learning mechanisms

The model is extended with a meta-learning mechanism based on the neuromodulation theory proposed by Doya (Daw & Doya, 2006; Doya, 1999, 2000, 2002; Schweighofer & Doya, 2003), which provides a neurophysiological interpretation of the properties of the meta-parameters (Table 2). Based on this theory, the computational role of meta-parameters in meta-learning is played by the neurotransmitter dynamics (e.g., dopamine, serotonin, acetylcholine and noradrenaline) within the brain areas that implement reinforcement learning, i.e., basal ganglia circuitry and prefrontal cortex (Daw & Doya, 2006; Doya, 2000, 2002; Houk, Davis, & Beiser, 2019; Montague, Dayan, & Sejnowski, 1996).

Dopamine is important in the brain for the prediction of rewards, and in meta-learning is related to the computation of the temporal difference error $\delta(t)$ (Apicella, Ljungberg, Scarnati, & Schultz, 1991; Schultz, 1998; Schultz, Apicella, & Ljungberg, 1993; Schultz, Dayan, & Montague, 1997; Wickens, 1990). Dopaminergic neurons respond to any salient event or novel input in the environment (Avery & Krichmar, 2017; Berridge, 2004; Bromberg-Martin, Matsumoto, & Hikosaka, 2010), and their depletion potentially impacts task performance (Berridge & Robinson, 1998, 1998; Cannon & Palmiter, 2003).

Serotonin controls the release of dopamine as well as the level of impulsivity (Eagle & Baunez, 2010), and in meta-learning is associated with the temporal discounting factor $[\gamma]$ that controls the time-scale of reward prediction (Schweighofer et al., 2008; Schweighofer, Tanaka, & Doya, 2007; Tanaka et al., 2004, 2007). The serotonergic and dopaminergic systems appear to be highly interactive and interlinked, even though their relationship is not clear. They are primarily activated in opposition (Boureau & Dayan, 2011; Daw et al., 2002), but under certain circumstances dopamine transporters transport significant serotonin molecules to the dopamine terminals (Cools, Nakamura, & Daw, 2011; Fischer & Ullsperger, 2017; Nakamura, Matsumoto, & Hikosaka, 2008; Winstanley, Theobald, Dalley, & Robbins, 2005; Zhou et al., 2005).

Noradrenaline regulates the exploration/exploitation balance of the behavior of an agent and is thus associated with the inverse temperature parameter β, which controls the action selection process (Aston-Jones & Cohen, 2005; Avery & Krichmar, 2017; Cohen, McClure, & Yu, 2007; Ishii, Yoshida, & Yoshimoto, 2002). Interestingly, the noradrenergic system may also be crucial for the functional reorganization of brain activity when environmental factors mutate to allow a behavioral adaptation (Bouret & Sara, 2005).

Acetylcholine aids memory storage and renewal balance, and may be associated in meta-learning with the learning rate α which controls the brain's attention through its cortical projections and the speed of the memory update (Avery & Krichmar, 2017; Doya, 2002; Hasselmo & Bower, 1993; Hasselmo & Schnell, 1994; Partridge, Apparsundaram, Gerhardt, Ronesi, & Lovinger, 2002; Rasmusson, 2000; Yu & Dayan, 2002).

We formalized a meta-learning hyperparameter optimization which includes noradrenergic modulation (i.e., the exploration/exploitation rate β), the role of dopamine (e.g., $\delta(t)$) and the influence of serotonin (i.e., the control of dopamine release and $[\gamma]$). Humphries and coll. (Humphries, 2012) discussed the dopamine effect on the exploration/exploitation rate. The action probability density function, computed with the Boltzmann Softmax Function (see Eq. (VI)), feeding the β and action value Q over the possible actions, can be either uniform or with a prominent peak. If it is uniform, the system is in an exploration regime whereby actions are all equiprobable and it is associated with low levels of dopamine. If there is a prominent peak, the system is in an exploitation regime whereby one action prevails over the others, and it is related to high dopaminergic levels.

The metric used for detecting the peak presence in a probability density distribution is entropy. A distribution with a low level of entropy is quasi-deterministic, i.e., an action is highly likely to occur, whereas high levels of entropy describe flat distributions

where it is less possible to predict a chosen action. Thus, the entropy H of the action probability density function and the dopamine level can be assumed to be inversely correlated, as follows (Eq. (VII)):

$$H(D) = H(P(\beta(D), Q)) = -\sum_{i=1}^{3} p_i \log_2 p_i$$
$$= -\sum_{i=1}^{3} P(\beta(D) \cdot Q_i) \log_2 (P(\beta(D) \cdot Q_i)) \quad \text{(VII)}$$

where $D = [D_1] \cdot \delta(t)$. This inverse relation is satisfied when the dopamine level and the β meta-parameter are linked with a sigmoid function, as follows (Eq. (VIII)):

$$\beta = \beta([D_1] \cdot \delta(t)) = \frac{k_{a[D_1]}}{1 + e^{\left(-k_{b[D_1]}\left([D_1]\cdot\delta(t) - k_{c[D_1]}\right)\right)}} \quad \text{(VIII)}$$

where $\left[k_{a[D_1]}, k_{b[D_1]}, k_{c[D_1]}\right]$ are the parameters that define the maximum, the slope, and the central value of the sigmoid, respectively. At this point we have only considered $[D_1]$. We then included the SNc dopamine D_2 efferent density $[D_2]$ in the model that tonically (i.e. no $\delta(t)$ modulation) influences the striatal neuron excitability codified by the central value C_0 of the sigmoid function (Tanaka et al., 2004), as follows (Eq. (IX)):

$$C_0 = C_0([D_2]) = \frac{k_{a[D_2]}}{1 + e^{\left(-k_{b[D_2]}\left([D_2] - k_{c[D_2]}\right)\right)}} \quad \text{(IX)}$$

where $\left[k_{a[D_2]}, k_{b[D_2]}, k_{c[D_2]}\right]$ define the maximum value, the slope, and the central value of the sigmoid, respectively. We assumed that $[D_2]$ efferent density has a symmetric and specular behavior with respect to the efferent density $[D_1]$, so the relationship between $[D_2]$ and C_0 assumes the same form as the relationship between $[D_1]$ and β.

Schweighofer et al. (2007) highlighted that the relationship between serotonin and dopamine may be more complex than expected. In this model we assumed a monotonic decrescent logarithmic relationship between the dopaminergic efferent ($[D_1]$ and $[D_2]$ involved in the VTA/ACC and SNc/striatum pathways respectively), and serotonin concentration $[\gamma]$. This accounts for a differential effect both in space inside the striatum and in the reward temporal scale (Schweighofer et al., 2007; Tanaka et al., 2007), as follows (Eq. (X)):

$$[\mathrm{D_1}] = \mathrm{k_{1[D_1](VTA)}} \log_{10}\left([\gamma] + \mathrm{k_{2[D_1](VTA)}}\right) \quad \text{(X)}$$

$$[D_2] = k_{3[D_2](SN_C)} \log_{10}\left([\gamma] + k_{4[D_2](SN_C)}\right)$$

where $\left[k_{1[D_1](VTA)}, k_{2[D_1](VTA)}\right]$ and $\left[k_{3[D_2](SN_C)}, k_{4[D_2](SN_C)}\right]$ are the parameters in the equations that are estimated as such $[D_1]$, $[D_2]$ and $[\gamma]$ are bounded between 0 and 1 and $[\mathrm{D_1}] = [D_2] = 0$ *if* $[\gamma] = 0$. The ACC layer also computes the meta-value $M(a_i, t)\,; a_i(s,t) \in \{Left, Right, Inhibition\}$ of the directions and the hold signal. The meta-values indicate how these actions are associated with variations in the average reward. The agent learns that the presentation of the hold signal is initially followed by a drop in the average reward and thus the hold signal is associated with a reduction in the meta-value. Every time a reward average is computed (at the end of the current trial), meta-values are updated based on variations in the reward average, as follows (Eq. (XI)):

$$M(a_i, t) \leftarrow M(a_{i-1}, t) + \eta \cdot m(t) \quad \text{(XI)}$$

$$m(t) = -\frac{r(t-2) + r(t-3)}{2} + \frac{r(t-1) + r(t-2)}{2}$$

where η is the update rate and $m(t)$ is the estimated reward average, i.e., the local variation of the total computed reward.

When the meta-value associated with any object is below a certain threshold, the presentation of this cue to the agent automatically resets the action values (random values) and an increase in β to high values, for example, 9. Therefore, the simulated agent will display exploitation behavior after this reset. The point in time when the hold signal meta-value crosses the threshold divides the simulation into the training and test phases.

We also implemented a linear relationship in the model between the delay d and the action values reset Q_{reset} and the time constant τ, respectively, as follows (Eq. (XII)):

$$Q_{reset} = a_{Q(hold)} \cdot d + b_{Q(hold)} + U(0, 0.3) \quad \text{(XII)}$$

$$\tau = a_{\tau(hold)} \cdot d + b_{\tau(hold)}$$

where $\left[a_{Q(hold)}, b_{Q(hold)}\right]$, $\left[a_{\tau(hold)}, b_{\tau(hold)}\right]$ were estimated is such a way that the delay, the action values and the time constant were in the [0,25], [0.2,0.5] and [0.25, 0.45] range, respectively. U is a random uniform distribution with a support in the [0,0.3] range.

2.3. Task

We implemented a brain-inspired ML control model of the prefrontal cortex and basal ganglia that performs a conflictual decision-making task (Fig. 2) (Berns & Sejnowski, 1996; Bogacz & Gurney, 2007; Collins & Koechlin, 2012; Dunovan & Verstynen, 2016; Khamassi et al., 2013; Rosenbloom, Schmahmann, & Price, 2012). We generated a conflictual instance between the prefrontal cortex and motor output in a simulated agent.

We now detail the temporal organization of the tasks. After four correct selections of the target with reward (i.e., to ensure the correct encode of the target), the agent encounters a trial which can be either Go (50% of trials) or Hold (50% of trials) (Fig. 2). In Go Trials, the simulated agent goes in the direction that maximizes the expectation of the associated reward while continuing to perform the previous choice ($r = 1$ if $a_i(s,t) = Left \wedge a_i(s,t-1) = Left \vee \mathrm{r} = 1$ if $a_i(s,t) = Right \wedge a_i(s,t-1) = Right$). In contrast, Hold Trials comprise a hold signal that alerts the agent to inhibit the movement ($r = 1$ if $a_i(s,t) = Inhibition$). This hold signal may be delivered either before the onset of the movement (i.e., NoGo Trials, $d = 0$) or after a certain delay from the appearance of the stimulus (i.e., Stop-Signal Trials, $d \in \{10, 20, 30\}$ samples).

The task was thus designed to manipulate different modes of movement inhibition and, hence, to unpredictably countermand the initially planned response: going (action restraint) or cancelling (action cancellation), depending on the task conditions (NoGo and Stop-Signal Trials, respectively) (Eagle, Baunez, et al., 2008; Schall et al., 2017).

The agent has to learn to change strategy when there is a hold signal because there is a change in strategy associated with the expected reward. After the learning phase, the agent embeds the hold signal presentation concept by means of a meta-learning system where we have an optimization of the hyperparameter evolution and the meta-values that inherit the significance level that the model gives to the hold signal and directions.

2.4. Computed parameters

We computed the following parameters to evaluate the agent's performance in both the training and the test phases during the task. First, we divided Hold Trials into correct and failure trials according to the reward outcome. Second, we computed the reaction time (RT) in all the trials. The reaction time is defined as the time at which the action-selected normalized (max–min

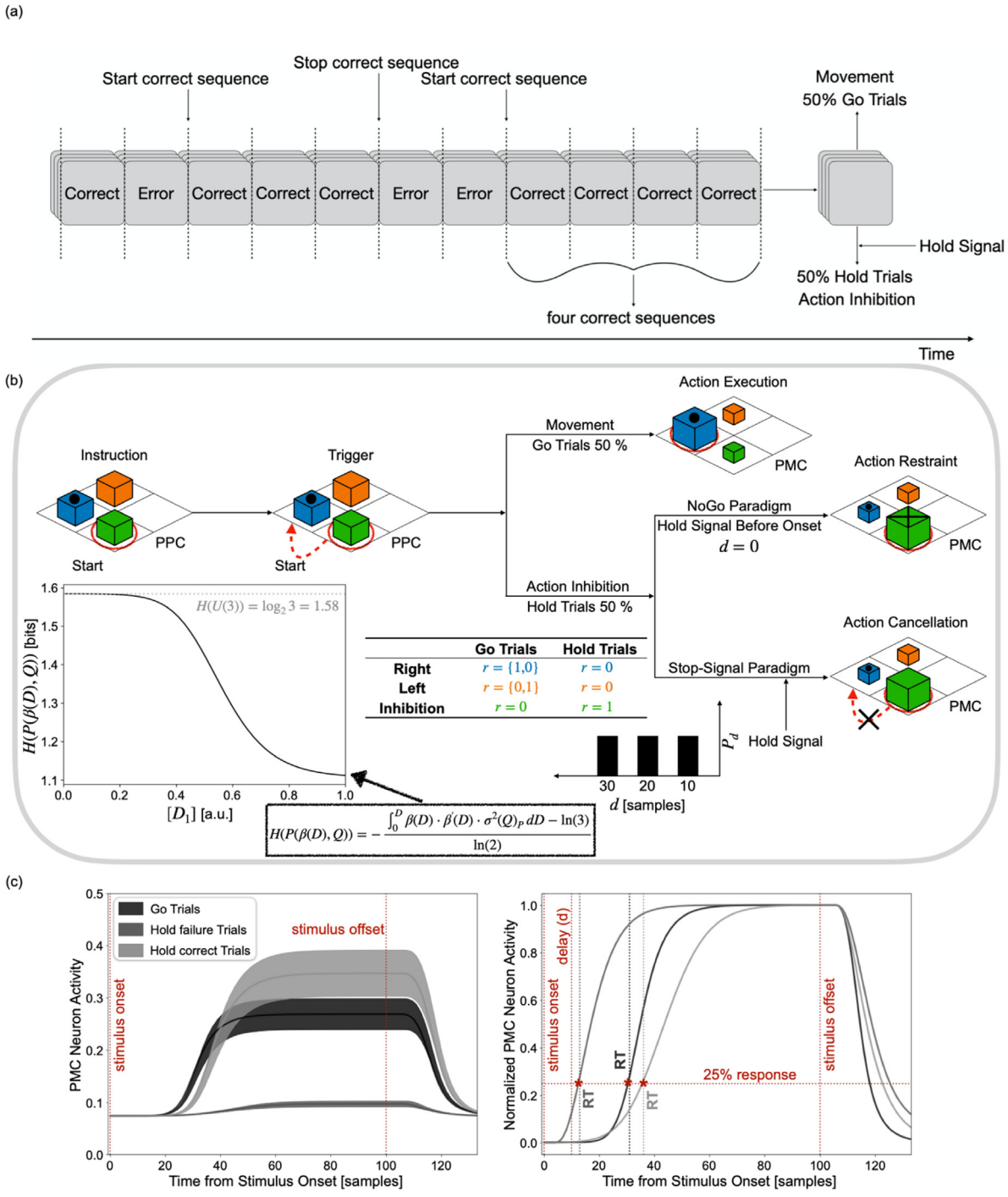

Fig. 2. (a) Time course of the task: each gray square represents the appearance of a stimulus. After four correct sequences, either a Go Trial or Hold Trial appears. (b) The temporal sequence of the trial: instructions, trigger, and agent's response. In Go Trials (50% of trials) the agent plans (red dashed arrow) and executes the desired action (big blue cube) from the starting point (red circle) to reach the target (black circle) after the trigger (i.e., stimulus onset). In Hold Trials (50% of trials), a hold signal may occur either before the initiation of the action (NoGo Paradigm, $d = 0$) or at a certain random delay (Stop-Signal Paradigm, $d \in \{10, 20, 30\}$ samples uniformly distributed) after the trigger appearance. In Hold Trials the agent performs different forms of action inhibition (black cross and big green cube): action restraint in NoGo Paradigm, i.e., the refraining from responding, and action cancellation in Stop-Signal Paradigm, i.e., stopping of an already-prepared response. The table indicates the rewards (r) associated with each action in the two conditions illustrated in the figure: Go and Hold Trials. For the sake of simplicity, we only represent the PPC layer during the instruction and trigger task segments and the PMC layer for the response. The volume of the cube indicates the neuron's firing rate. The relationship (Eq. (VII)) between entropy $H\,(P\,(\beta\,(D)\,,\,Q\,))$ and the amount of the dopamine D_1 efferents $[D_1]$ is a monotonic decrescent sigmoid-like function as $\beta\,(D)$ and D are linked with a sigmoid function (see Materials and Methods and Appendix for details). (c) (Left) Mean of PMC activity across Go Trials (black, $n = 77$), Hold failure Trials (dark gray, $n = 44$), and Hold correct Trials (light gray, $n = 46$). (Right) Max–min normalized PMC activity during exemplary Go (black), Hold failure at $d = 10$ (dark gray), and Hold correct at $d = 10$ (light gray) Trials are illustrated. The reaction time (RT) is displayed with vertical dotted lines using the same colors. The onset/offset of the stimulus (vertical dotted red lines), the delay (d) (vertical dotted red lines), and the threshold for the RT computation (horizontal dotted red lines) are displayed. Shaded areas stand for SEM.

normalization) PMC neuron activity reaches 25% of the peak of activity after the stimulus onset, as follows (Eq. (XIII)):

$$RT : t* \mid \frac{PMC(t*) - min(PMC(t))}{max(PMC(t)) - min(PMC(t))} \geq 0.25 \quad \text{(XIII)}$$

Second, we computed the Stop-signal reaction time (SSRT) in the Stop-Signal Trials (only test phase) using the quantile method (Band, van der Molen, & Logan, 2003; Pasquereau & Turner, 2017b; Williams, Ponesse, Schachar, Logan, & Tannock, 1999). Briefly, we sorted the RTs in the Go Trials into ascending order, and we selected the RT percentile corresponding to the same proportion of failure ($P_{failure}$) in inhibition during trials (i.e., we picked the median of RT if $P_{failure} = 0.50$). We defined the SSRT as this quantile subtracted by the stop-signal delay, providing an estimate of the average stopping latency, i.e., the time taken to cancel the action. We also computed the probability and the peak of the PMC activity of the chosen action. The global accuracy of the model (i.e., proportion of Go + Hold Trials) was evaluated along with the right inhibition (i.e., proportion of Hold Trials). Finally, we analyzed the performance of the agent at different concentrations of serotonin ($[\gamma] \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, and, consequently, at different levels of both dopamine receptors [D_1], [D_2].

2.5. Statistical data analysis

Data were statistically analyzed using SPSS Statistics v. 26.0 (SPSS Inc, Chicago, IL). All data are expressed as Mean ± Standard Error (SEM), unless otherwise specified. Relationships in simulation data between network parameters and inhibition performance metrics (e.g., right inhibition, $RT_{(Hold,failure)}$, $RT_{(Hold,correct)}$) with respect to delay were assessed by linear regression analysis (e.g., slope and intercept, ρ, R^2, p-value with a significance level of 0.05). We compared different conditions computing Cohen's effect size d_c (Lakens, 2013) using Cohen's interpretation of small ($d_c \sim 0.2$), medium ($d_c \sim 0.5$), and large ($d_c \sim 0.8$) effect size. We assessed the effect of the serotonin concentration $[\gamma] \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ computing the η^2 again using Cohen's interpretation of small ($\eta^2 \sim 0.01$), medium ($\eta^2 \sim 0.06$) and large ($\eta^2 \sim 0.14$) effect sizes.

2.6. Model implementation

The code was written in Python 3.8.1. Executing one simulation with 1000 stimuli takes ∼3 min on an Intel Core ™ i9-9880H CPU @2.3GHz × 8. The code is available on GitHub at the following link: https://github.com/FedericaRobertazzi/Robertazzi_et_al_code_BG_MetaLearning_DecisionMakingTasks_NeuralNetworks_2022

3. Results

We analyzed the behavior of the artificial agent with a moderate level of serotonin concentration (e.g., $[\gamma] = 0.5$) during the various trials (i.e., Go Trials, NoGo Trials and Stop-Signal Trials) running 40 simulations of 1000 stimuli. We then performed a sensitivity analysis sweeping $[\gamma]$ parameter from 0.1 to 0.9 with a discrete step of 0.2.

3.1. Meta-learning effects during the learning phase

The artificial agent performed 166.37 ± 0.56 trials (Go Trials: 85.57 ± 1.08; Hold Trials: 80.8 ± 1.02). The training period was defined as the time required to achieve a fixed hold signal meta-value threshold M_{thr} set at −0.25. The remaining time was labeled as the test period in which the artificial agent embedded (i.e., learning to react) the hold signal (see Fig. 3). Each time the action selection was followed by a variation in the estimated reward average $m(t)$, the meta-value $M(t)$ of the actions and hold signal changed accordingly (see Eq. (XI)). Because of the higher likelihood of a drop in $m(t)$ after the appearance of the hold signal during the training phase, the agent learned this association decreasing over stimuli the hold signal meta-value. The agent crossed the hold signal meta-value threshold M_{thr} at 51.27 ± 2.40 stimuli with a slope of −0.0048 ± 0.00 [stimulus^{-1}] reaching a minimum plateau of −2.31 ± 0.06. In addition, the agent learned to select the correct command – action inhibition – after the hold signal as suggested by an increase in the maximum action value of the inhibition command $Q_{Inh(Max)}$ after learning (training: 0.57 ± 0.003 vs test:1.55 ± 0.01; $d_c = 11.41$). Interestingly, the meta-value associated with directions (Left and Right actions) did not decline because the perception of the target was on average followed by the same amount of positive and negative rewards. Finally, the number of incorrect stimuli before starting a new trial decreased with learning (training: 2.47 ± 0.08 vs test:1.69 ± 0.01; $d_c = -1.392$).

3.2. NoGo Paradigm results

We analyzed the results in the NoGo Paradigm ($d = 0, n = 40$ simulations) to evaluate the unpredictable countermand of the not-yet-initiated response (i.e., action restraint) (Fig. 4, Table 3, Table 4). Both the accuracy (Mean ± SD; training: 53.32 ± 15.51% vs test: 85.22 ± 4.42%; $d_c = 2.060$) and the right inhibition (Mean ± SD; training: 0% vs test: 70.67 ± 9.20%; $d_c = 7.675$) increased during the test phase, suggesting a positive effect of the meta-learning control framework in the performance of the task. Because the right inhibition in the training phase is 0%, parameters and comparisons during NoGo correct Trials in this phase were not defined.

We found that RTs in Go Trials did not differ during the training and test phases (Mean ± SD; training: 28.03 ± 0.64 samples vs test: 27.95 ± 0.22 samples; $d_c = -0.118$). In contrast, RTs in NoGo failure Trials decreased considerably with learning (Mean ± SD; training: 28.21 ± 1.07 samples vs test: 21.69 ± 1.56 samples; $d_c = -3.489$). During the test phase, RTs in NoGo failure Trials were faster than in Go Trials (training: $d_c = 0.168$; test: $d_c = -4.081$). RTs in NoGo correct Trials were slower than in the Go Trials (test: $d_c = 9.154$) and in NoGo failure Trials (test: $d_c = 6.909$) in the test phase.

The PMC_{peak} in Go Trials was similar during the training and test phases (Mean ± SD; training: 0.25 ± 0.12 vs test: 0.29 ± 0.02; $d_c = 0.283$). Learning also had a differential effect on the PMC_{peak} with a decrease in NoGo failure Trials (Mean ± SD; training: 0.24 ± 0.11 vs test: 0.08 ± 0.01; $d_c = -1.488$). The PMC_{peak} in NoGo failure Trials was also lower than in Go Trials (training: $d_c = -0.038$; test: $d_c = -7.020$) in the test phase. The PMC_{peak} in NoGo correct Trials was higher than in Go Trials (test: $d_c = 1.771$) and in NoGo failure Trials (test: $d_c = 6.371$) during the test phase.

The P_{Max} of the chosen action increased during the test phase both in Go Trials (Mean ± SD; training: 0.13 ± 0.05 vs test: 0.23 ± 0.02; $d_c = 1.545$) and NoGo failure Trials (Mean ± SD; training: 0.15 ± 0.09 vs test: 0.23 ± 0.04; $d_c = 0.873$). No significant differences were detected between Go and NoGo Trials in either the training or test phases.

3.3. Stop-Signal Paradigm results

We analyzed the results in the Stop-Signal Paradigm ($d \in \{10, 20, 30\}$, $n = 40$ simulations), in which the agent was alerted

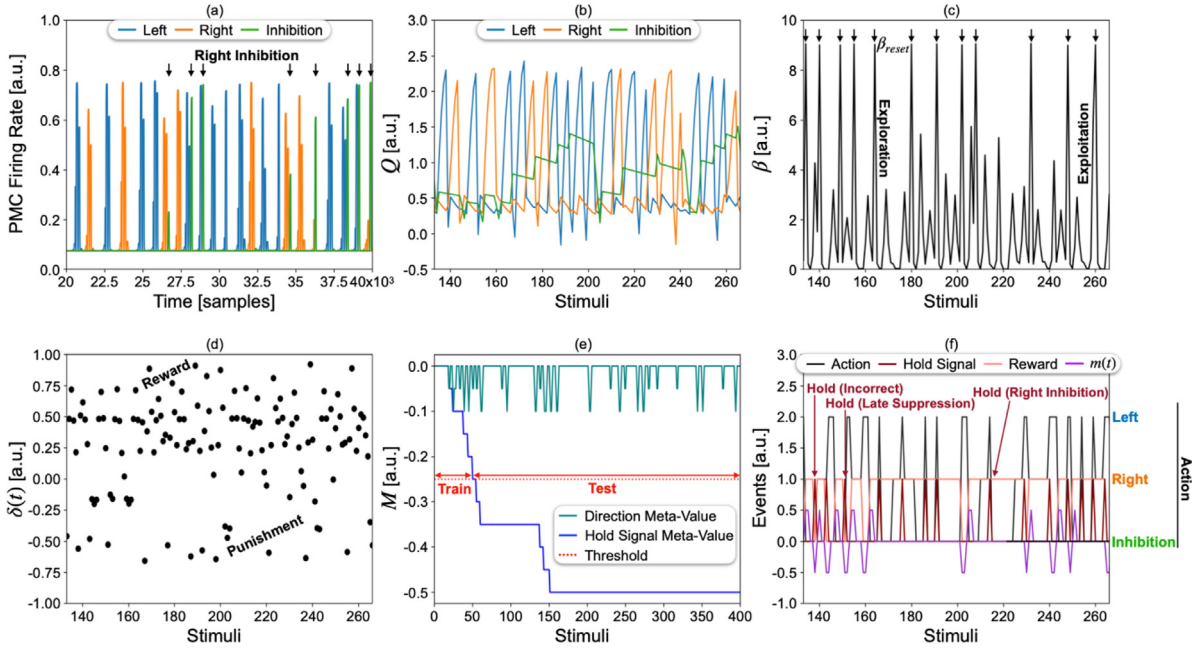

Fig. 3. Simulation with 400 stimuli. (a) Neuron activities in PMC layer (Left (blue), Right (orange) and Inhibition (green)) in a window ranging from 20,000 to 40,000 samples (timestep = 0.1 ms). Neuron activities are mapped from 0 to 1 by the non-linear sigmoid function. Black arrows highlight the insurgence of a right inhibition of the motor command (i.e., action inhibition). (b) Action values (Q) of the possible actions in the ACC layer are illustrated, which are updated according to Eq. (V) after the administration of the reward. Action values of the non-selected actions follow an exponential decay set by the forgetting factor F_{factor} over time. (c) The meta-parameter β decreases to allow explorative behavior, while it increases during the exploitation phase. β peaks represent the meta-learning mechanism used to react to the hold signal. (d) Each point illustrates the temporal difference error $\delta(t)$ (correct behavior ($\delta(t) > 0$) and punishment ($\delta(t) < 0$) after the execution of each action. (e) Meta-Values of the actions (Left, Right; aqua green) and meta-values of the hold signal (blue) are plotted. The hold signal meta-value threshold (dotted red line) at -0.25 divides the training and test phases. (f) The possible actions that can be performed by the agent are represented by the black line (Left (2), Right (1) and Inhibition (0)). The reward administered is represented by the pink track. The surge of the hold signal that confounds the previous received instruction is represented by the red triangle. The estimated reward average $m(t)$ (refer Eq. (XI)), defined as the local variation of the total computed reward is illustrated in violet. When the agent chooses a possible action (black), the reward is administered (pink). The reward is 1 if the action of the agent coincides with the desired target, and otherwise is 0. During the task, a hold signal (red) can occur in Hold Trials. The agent can react to the hold signal: (i) inhibits the action (right inhibition), (ii) inhibits (e.g., in a subsequential trial) the motor command late (late inhibition), and (iii) does not suppress the action (incorrect inhibition). (b–f) The parameters that are updated in terms of a stimulus presentation according to the reinforcement learning and meta-learning mechanisms are plotted with respect to the stimuli.

Table 3
Cohen's effect size d_c between training and test phases.

	Stop-Signal Paradigm	NoGo Paradigm
Variable	Training vs Test	Training vs Test
$RT_{(Go)}$	–	$d_c = -0.118$
$RT_{(Hold,failure)}$	–	$d_c = -3.489$
$RT_{(Hold,correct)}$	–	–
$PMC_{peak(Go)}$	$d_c = 0.406$	$d_c = 0.283$
$PMC_{peak(Hold,failure)}$	$d_c = -1.385$	$d_c = -1.488$
$PMC_{peak(Hold,correct)}$	–	–
$P_{Max(Go)}$	$d_c = 1.581$	$d_c = 1.545$
$P_{Max(Hold,failure)}$	$d_c = 1.111$	$d_c = 0.873$
$P_{Max(Hold,correct)}$	–	–
Right Inhibition	$d_c = 4.555$	$d_c = 7.675$
Accuracy	$d_c = 1.952$	$d_c = 2.060$

Cohen's effect size d_c of the parameters in Go and Hold Trials in both Stop-Signal Paradigm and NoGo Paradigm between training and test phases across 40 simulations. SSRT is not included because of the lack of definition during the training phase. $d_c > 0$ indicate higher values during the test phase.

to cancel the ongoing action – action cancellation – after a certain delay from the appearance of the stimulus (Fig. 5, Table 3, Table 4). The accuracy (Mean ± SD; training: 45.53 ± 15.54% vs test: 73.29 ± 4.76%; $d_c = 1.952$) as well as the right inhibition (Mean ± SD; training: 0% vs test: 47.83 ± 10.46%; $d_c = 4.555$) increased during the test phase. Because the right inhibition in the training phase is 0%, parameters and comparisons during Stop-Signal correct Trials in this phase are not defined.

The PMC_{peak} increased in Go Trials (Mean ± SD; training: 0.23 ± 0.14 vs test: 0.28 ± 0.03; $d_c = 0.406$) and decreased in Stop-Signal failure Trials (Mean ± SD; training: 0.24 ± 0.10 vs test: 0.09 ± 0.009; d_c= - 1.385) with learning. Additionally, the PMC_{peak} in Stop-Signal failure Trials was lower than in Go Trials (training: $d_c = 0.064$; test: d_c= - 6.809) in the test phase. The PMC_{peak} in Stop-Signal correct Trials was higher than in Stop-Signal failure Trials (test: $d_c = 4.020$) and in Go Trials (test: $d_c = 1.061$) during the test phase.

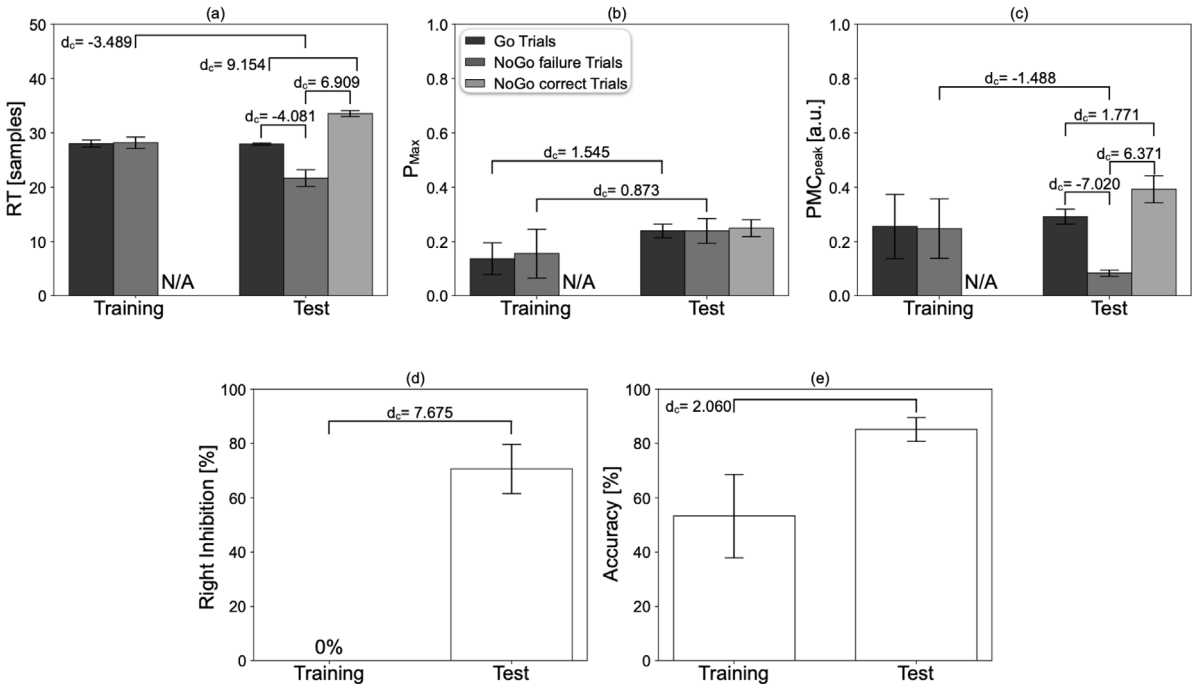

Fig. 4. The results of 40 simulations of 1000 stimuli during the training and test phases in the NoGo Paradigm ($d = 0$). RT (a), P_{Max}, (b) PMC_{peak} (c) are obtained by averaging across the various trial conditions: Go Trials (black), NoGo failure Trials (dark gray) and NoGo correct Trials (light gray). Right Inhibition (d) and Accuracy (e) are computed using the NoGo Trials and all trials, respectively. N/A stands for non-available. Results are expressed as Mean ± SD. Cohen's effect sizes are reported for large effects d_c (see Methods). d_c is positive/negative if the right/left group mean is higher.

Table 4
Cohen's effect size d_c between different trial conditions.

	Go vs Stop-Signal failure		Go vs Stop-Signal correct		Stop-Signal failure vs Stop-Signal correct	
Variable	Train	Test	Train	Test	Train	Test
RT	–	–	–	–	–	–
PMC_{peak}	$d_c = 0.064$	$d_c = -6.809$	–	$d_c = 1.061$	–	$d_c = 4.020$
P_{Max}	$d_c = 0.056$	$d_c = -0.133$	–	$d_c = -0.090$	–	$d_c = 0.064$
	Go vs NoGo failure		Go vs NoGo correct		NoGo failure vs NoGo correct	
RT	$d_c = 0.168$	$d_c = -4.081$	–	$d_c = 9.154$	–	$d_c = 6.909$
PMC_{peak}	$d_c = -0.038$	$d_c = -7.020$	–	$d_c = 1.771$	–	$d_c = 6.371$
P_{Max}	$d_c = 0.167$	$d_c = -2.344e{-}17$	–	$d_c = 0.186$	–	$d_c = 0.189$

Cohen's effect size d_c of the parameters in training and test phases between Go, Stop-Signal Trials (failure/correct) and NoGo Trials (failure/correct) across 40 simulations. SSRT is not included because of the lack of definition during the training phase and in Go Trials. $d_c > 0$ indicate higher values in the right group of the comparison; for example, Go < Stop-Signal correct if $d_c > 0$ in Go vs Stop-Signal correct.

The P_{Max} of the selected action increased in the test phase in Go Trials (Mean ± SD; training: 0.13 ± 0.06 vs test: 0.23 ± 0.02; $d_c = 1.581$) as well as Stop-Signal failure Trials (Mean ± SD; training: 0.14 ± 0.08 vs test: 0.22 ± 0.03; $d_c = 1.111$). No significant differences were detected between Go and Stop-Signal Trials in either the training or test phases.

The SSRT was defined only in the test phase (Mean ± SD; 12.21 ± 1.08 samples) because the right inhibition is 0%.

3.4. Serotonin modulation effects

In the Hold Trials (i.e., both NoGo and Stop-Signal Paradigm) in the test phase, we analyzed the relationship between the right inhibition and $RT_{(Hold)}$ respectively, (Fig. 6) at $[\gamma] = 0.5$. We found a negative correlation between the right inhibition and the delay ($\rho = 0.83$, $R^2 = 0.69 \pm 0.03$, p-value < 0.001). Both $RT_{(Hold,failure)}$ ($\rho = 0.74$, $R^2 = 0.54 \pm 0.05$, p-value < 0.001) and $RT_{(Hold,correct)}$ ($\rho = 0.87$, $R^2 = 0.75 \pm 0.03$, p-value < 0.001) correlated positively with the delay. These results suggest slower and less accurate reactions to the hold signal as the delay increases (Table 6).

Thirdly, we replicated these analyses by sweeping the serotonin concentration $[\gamma] \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ to investigate the performance of the artificial agents in low and high serotonin concentration regimes. As expected from Eqs. (VIII), (IX), high levels of serotonin were characterized by a concomitant increase in C_0 and δ (Table 5(a)). For all levels of serotonin, we found a positive and significant correlation between the right inhibition and the delay (Table 6). Interestingly, slopes and intercepts were modulated with the serotonin level (slope: $\eta^2 = 0.536$; intercept: $\eta^2 = 0.697$) (Table 7). High regimes of serotonin thus reduced the sensitivity of the performance during Hold Trials to the delay of appearance of the hold signal (both the slope and the explained variance R^2 of the linear fitting decreased).

Both $RT_{(Hold,failure)}$ and $RT_{(Hold,correct)}$ correlated positively with the delay in all the serotonin concentrations (Table 6). Again, serotonin modulated the coefficients of these correlations ($RT_{(Hold,failure)}$, slope: $\eta^2 = 0.249$ and intercept: $\eta^2 = 0.925$; $RT_{(Hold,correct)}$, slope: $\eta^2 = 0.272$ and intercept: $\eta^2 = 0.485$) (Table 7).

For the sake of completeness, we investigated the interaction between the serotonin concentration and the effect of learning (e.g., training vs test phase). The results showed a strong

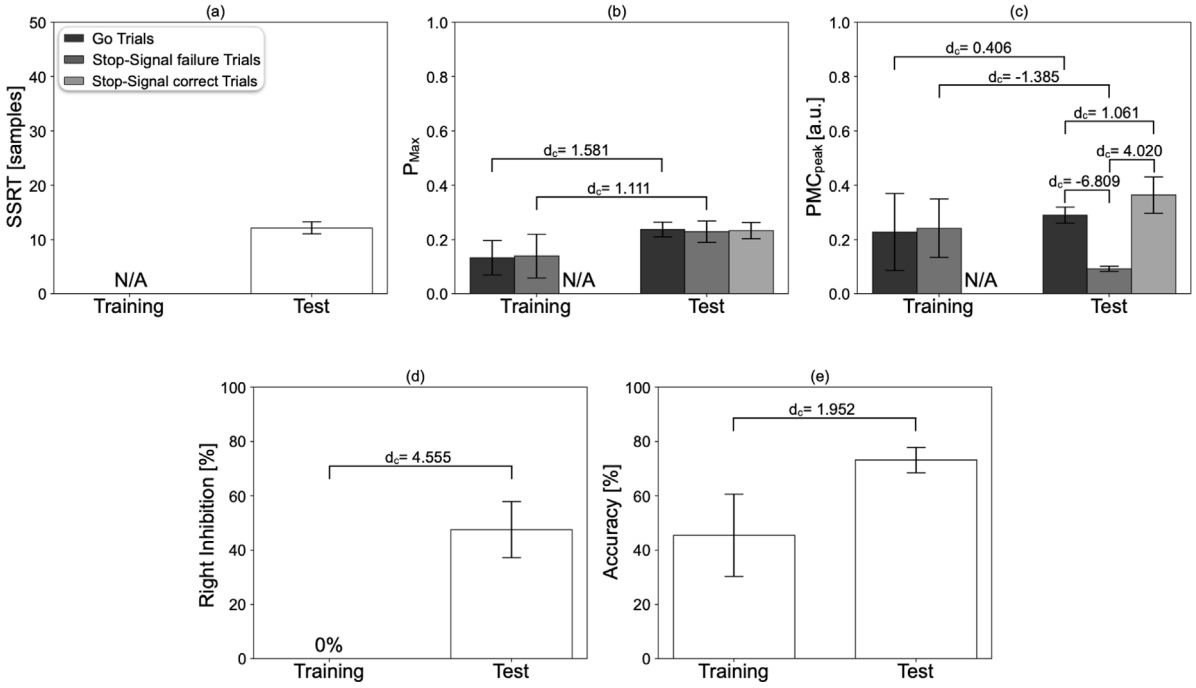

Fig. 5. The results of 40 simulations of 1000 stimuli during the training and test phases in the Stop-Signal Paradigm ($d \in \{10, 20, 30\}$). SSRT (a), P_{Max}, (b) PMC_{peak} (c) are obtained by averaging across the various trial conditions: Go Trials (black), Stop-Signal failure Trials (dark gray) and Stop-Signal correct Trials (light gray). Right Inhibition (d) and Accuracy (e) are computed using the Stop-Signal Trials and all trials, respectively (white). N/A stands for non-available. Results are expressed as Mean ± SD. Cohen's effect sizes are reported for large effects d_c (see Methods). d_c is positive/negative if the right/left group mean is higher.

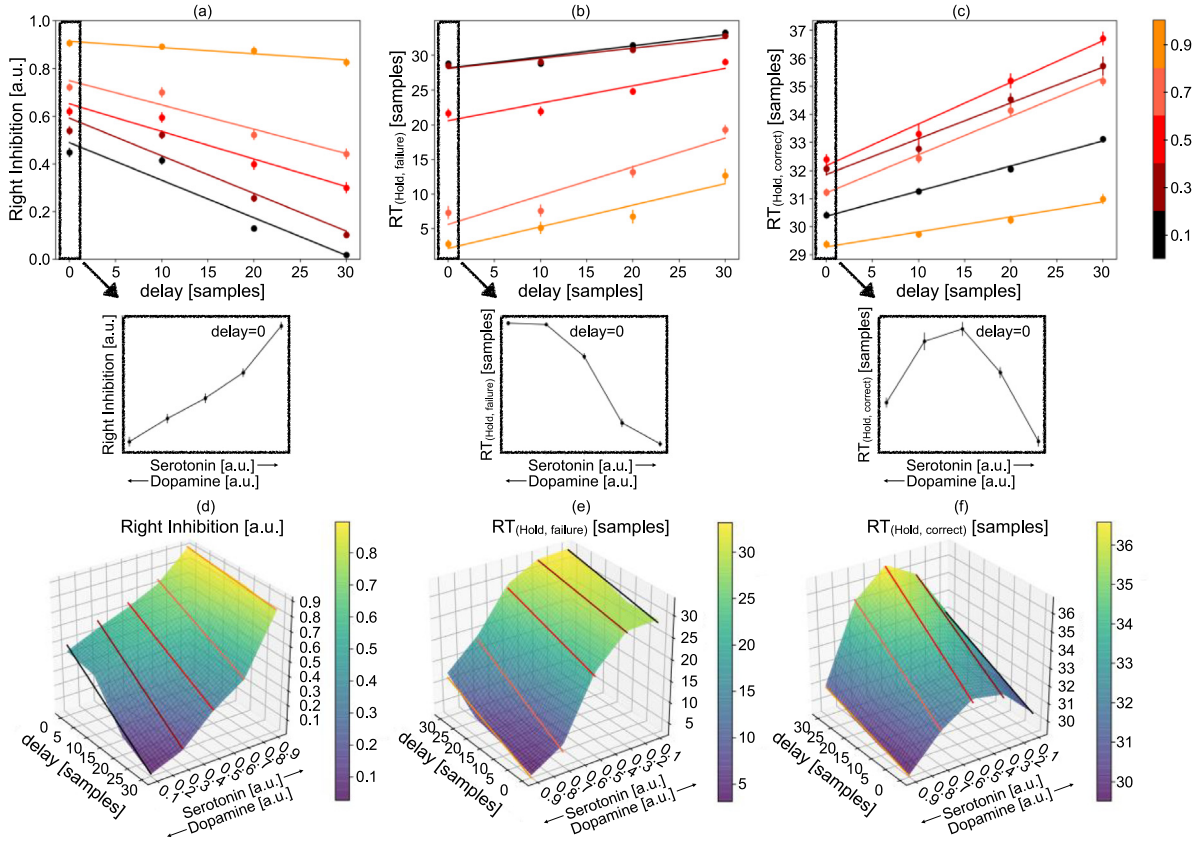

Fig. 6. The results are averaged across 40 simulations of 1000 stimuli during the test phase in Hold Trials. The relationship (simulation data (black) and linear fit (gradient palette between black and orange)) of the Right Inhibition (a), $RT_{(Hold,failure)}$ (b) and $RT_{(Hold,correct)}$ (c) with the delay d are illustrated for different serotonin concentrations $[\gamma] \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The box highlights the relationship between the parameters in only NoGo Trials ($d = 0$). Results are expressed as Mean ± SEM. (d–f) 3D representation of the relationship among task parameters, delay of the hold signal, and serotonin concentration.

Table 5
Serotonin effects on C_0, δ, Q_{Inh} and M_{Hold}.

(a)	C_0		δ	
Serotonin	Min	Max	Min	Max
0.1	0.52 ± 0.00	0.6 ± 0.00	−0.1 ± 0.00	0.13 ± 0.00
0.3	0.57 ± 0.00	0.6 ± 0.00	−0.35 ± 0.001	0.37 ± 0.005
0.5	0.6 ± 0.00	0.61 ± 0.00	−0.77 ± 0.03	1.01 ± 0.03
0.7	0.6 ± 0.00	0.65 ± 0.00	−2.45 ± 0.16	2.11 ± 0.07
0.9	0.6 ± 0.00	0.68 ± 0.00	−10.09 ± 2.19	8.35 ± 1.81

(b)	$Q_{Inh(Max)}$	
Serotonin	Training	Test
0.1	0.56 ± 0.02	0.68 ± 0.03
0.3	0.57 ± 0.02	1.03 ± 0.08
0.5	0.57 ± 0.02	1.55 ± 0.09
0.7	0.58 ± 0.01	2.08 ± 0.11
0.9	0.58 ± 0.05	3.65 ± 0.22

(c)	M_{Hold}		
Serotonin	# stimuli to M_{thr}	M_{Hold} plateau	Slope at M_{thr}[stimuli^{-1}]
0.1	45.37 ± 9.21	−3.27 ± 0.32	−0.0052 ± 0.001
0.3	50.05 ± 13.89	−2.62 ± 0.33	−0.0048 ± 0.001
0.5	51.27 ± 15.19	−2.31 ± 0.4	−0.0048 ± 0.001
0.7	57.65 ± 20.28	−1.62 ± 0.36	−0.0043 ± 0.001
0.9	57.57 ± 17.91	−0.6 ± 0.28	−0.0048 ± 0.004

Serotonin effects on (a) C_0 (central value of the sigmoid function), δ (temporal difference error), (b) $Q_{Inh(Max)}$ (maximum action value of the inhibition command in training and test phase) and (c) M_{Hold} (hold signal meta-value). Data are collected over 40 simulations (Mean ± SD).

Table 6
Linear fitting analysis.

	Serotonin	Slope	Intercept	ρ	R^2	p-value
Right Inhibition vs d	0.1	−0.01 ± 0.00	0.48 ± 0.01	0.91	0.82 ± 0.02	<0.001
	0.3	−0.01 ± 0.00	0.59 ± 0.01	0.88	0.78 ± 0.03	<0.001
	0.5	−0.01 ± 0.00	0.65 ± 0.01	0.83	0.69 ± 0.03	<0.001
	0.7	−0.01 ± 0.00	0.74 ± 0.01	0.80	0.64 ± 0.04	<0.001
	0.9	−0.002 ± 0.00	0.91 ± 0.01	0.65	0.42 ± 0.04	<0.001
$RT_{(Hold,failure)}$ vs d	0.1	0.16 ± 0.009	28.17 ± 0.23	0.86	0.74 ± 0.03	<0.001
	0.3	0.14 ± 0.01	28.11 ± 0.22	0.83	0.70 ± 0.04	<0.001
	0.5	0.25 ± 0.02	20.58 ± 0.57	0.74	0.54 ± 0.05	<0.001
	0.7	0.41 ± 0.03	5.63 ± 0.71	0.76	0.57 ± 0.04	<0.001
	0.9	0.31 ± 0.03	2.18 ± 0.53	0.68	0.47 ± 0.04	<0.001
$RT_{(Hold,correct)}$ vs d	0.1	0.08 ± 0.004	30.37 ± 0.12	0.89	0.80 ± 0.02	<0.001
	0.3	0.12 ± 0.01	31.86 ± 0.2	0.79	0.63 ± 0.04	<0.001
	0.5	0.14 ± 0.009	32.17 ± 0.23	0.87	0.75 ± 0.03	<0.001
	0.7	0.13 ± 0.008	31.20 ± 0.12	0.85	0.72 ± 0.03	<0.001
	0.9	0.05 ± 0.008	29.28 ± 0.12	0.75	0.56 ± 0.05	<0.001

The results obtained from a linear regression analysis of simulation data for each of the parameters presented in Fig. 6 with the delay d at different serotonin concentrations $[\gamma] \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The slope and intercept (Mean ± SEM) of the linear regression, ρ, R^2, and p-values are reported.

Table 7
Effect size η^2 of serotonin effect on linear regression analysis.

Coefficient	Right Inhibition vs d	$RT_{(Hold,failure)}$ vs d	$RT_{(Hold,correct)}$ vs d
Slope	$\eta^2 = 0.536$	$\eta^2 = 0.249$	$\eta^2 = 0.272$
Intercept	$\eta^2 = 0.697$	$\eta^2 = 0.925$	$\eta^2 = 0.485$

The effect size of serotonin on the coefficients (e.g., slope and intercept) extracted from the linear regression analysis (Table 6) ($[\gamma] \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$). Effect size is computed using η^2 (see Methods).

serotonin-learning interaction effect for most parameters, as suggested by the η^2 ($RT_{(Hold,failure)}$: $\eta^2 = 0.166$; $PMC_{peak(Hold,failure)}$: $\eta^2 = 0.301$; $P_{Max(Go)}$: $\eta^2 = 0.069$; right inhibition: $\eta^2 = 0.125$) (Fig. 7, Table 8). Indeed, both $RT_{(Go)}$ and $RT_{(Hold,failure)}$ decreased as the serotonin concentration increased during training ($RT_{(Go)}$: $\eta^2 = 0.885$; $RT_{(Hold,failure)}$: $\eta^2 = 0.900$) as well as in test phase ($RT_{(Go)}$: $\eta^2 = 0.995$; $RT_{(Hold,failure)}$: $\eta^2 = 0.947$). Interestingly, $RT_{(Hold,correct)}$ presented an inverse U-shaped profile with a peak at $[\gamma] = 0.5$ in the test phase ($RT_{(Hold,correct)}$: $\eta^2 = 0.860$). $PMC_{peak(Go)}$, $PMC_{peak(Hold,failure)}$, and $PMC_{peak(Hold,correct)}$ increased at higher serotonin concentrations during training ($PMC_{peak(Go)}$: $\eta^2 = 0.952$; $PMC_{peak(Hold,failure)}$: $\eta^2 = 0.974$) and in the test phase ($PMC_{peak(Go)}$: $\eta^2 = 0.997$; $PMC_{peak(Hold,failure)}$: $\eta^2 = 0.411$; $PMC_{peak(Hold,correct)}$: $\eta^2 = 0.957$). Finally, $P_{Max(Go)}$, $P_{Max(Hold,failure)}$ and $P_{Max(Hold,correct)}$ had a similar increasing pattern in both training ($P_{Max(Go)}$: $\eta^2 = 0.813$; $P_{Max(Hold,failure)}$: $\eta^2 = 0.799$) and test ($P_{Max(Go)}$: $\eta^2 = 0.950$; $P_{Max(Hold,failure)}$: $\eta^2 = 0.810$; $P_{Max(Hold,correct)}$: $\eta^2 = 0.939$) phases.

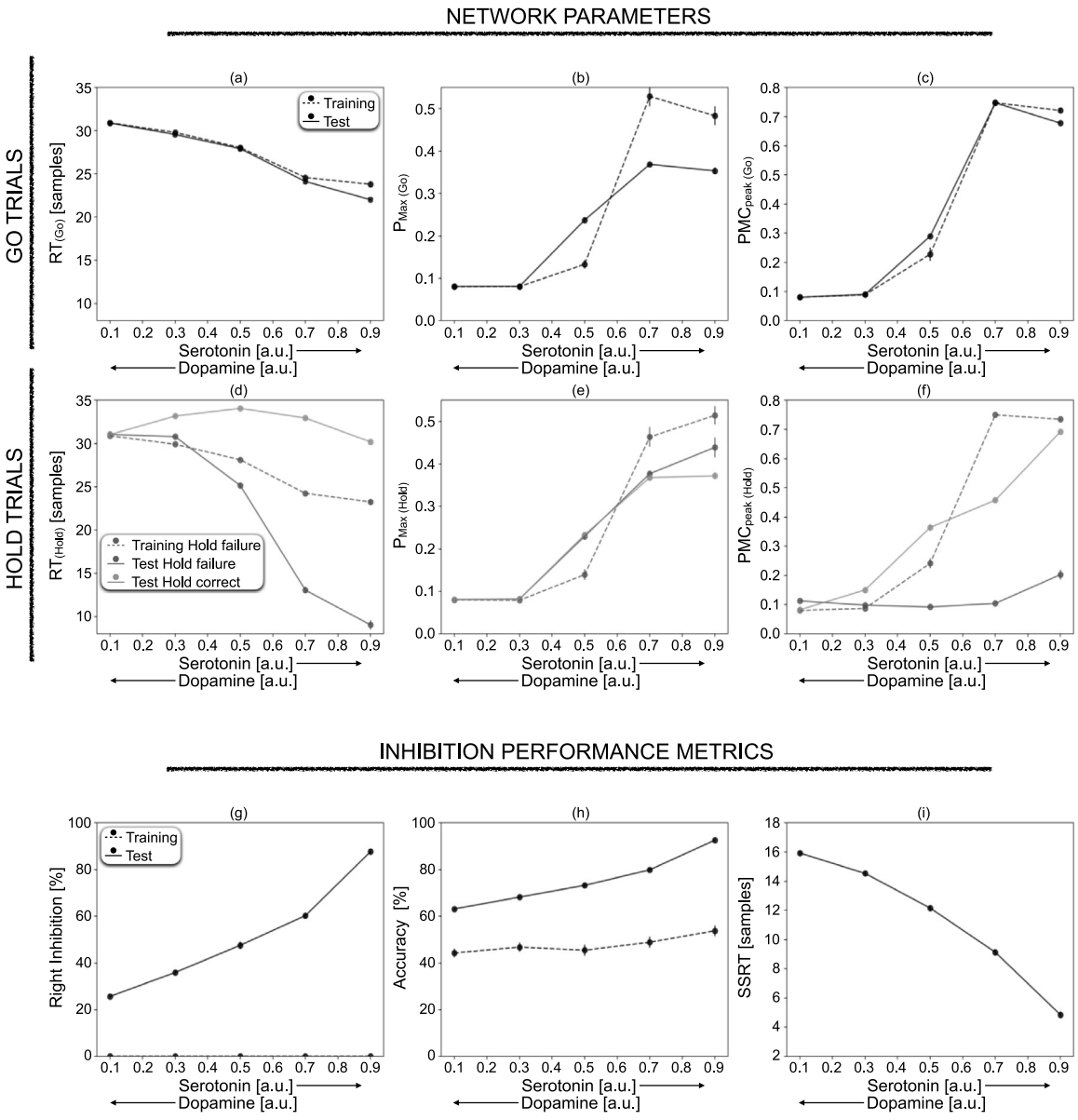

Fig. 7. Serotonin modulation ($[\gamma] \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) of the network parameters (RT, P_{max}, PMC_{peak}) and inhibition performance metrics (Right Inhibition, Accuracy, SSRT) averaged across 40 simulations of 1000 stimuli during the training (dashed line) and test (solid line) phases. Network parameters are displayed in Go Trials (a–c) and Hold Trials (d–f). In (d–f) Hold Trials are divided into failure (dark gray) and correct (light gray) Trials. Inhibition performance metrics (g–i) are illustrated. Results are expressed as Mean ± SEM.

Table 8
Effect size η^2 of the serotonin sensitivity analysis.

Variable	Serotonin	Learning	Serotonin x Learning
$RT_{(Go)}$	$\eta^2 = 0.929$	$\eta^2 = 0.007$	$\eta^2 = 0.010$
$RT_{(Hold,failure)}$	$\eta^2 = 0.653$	$\eta^2 = 0.132$	$\eta^2 = 0.166$
$RT_{(Hold,correct)}$	–	–	–
$PMC_{peak(Go)}$	$\eta^2 = 0.970$	$\eta^2 = 4.22e{-}5$	$\eta^2 = 0.003$
$PMC_{peak(Hold,failure)}$	$\eta^2 = 0.411$	$\eta^2 = 0.252$	$\eta^2 = 0.301$
$PMC_{peak(Hold,correct)}$	–	–	–
$P_{Max(Go)}$	$\eta^2 = 0.769$	$\eta^2 = 0.010$	$\eta^2 = 0.069$
$P_{Max(Hold,failure)}$	$\eta^2 = 0.775$	$\eta^2 = 0.001$	$\eta^2 = 0.028$
$P_{Max(Hold,correct)}$	–	–	–
Right Inhibition	$\eta^2 = 0.125$	$\eta^2 = 0.716$	$\eta^2 = 0.125$
Accuracy	$\eta^2 = 0.128$	$\eta^2 = 0.536$	$\eta^2 = 0.035$

Effect size of the interaction between serotonin and learning (training vs test phase) computed for each parameter in Fig. 7. Effect size of SSRT is not included because of the lack of definition during the training phase. Effect size is computed using the η^2 (see Methods).

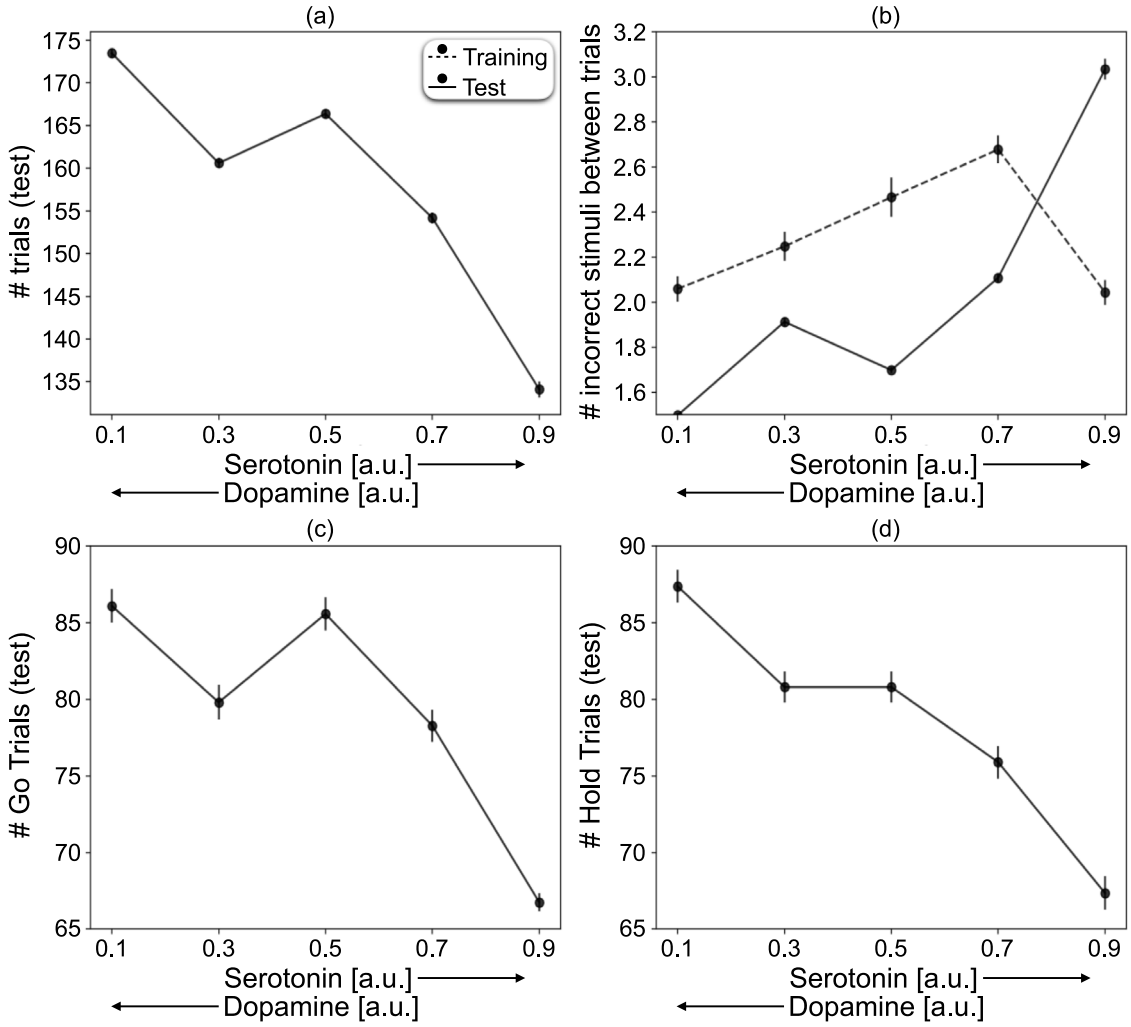

Fig. 8. Serotonin modulation ($[\gamma] \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) of the number of all trials (a), Go Trials (c) and Hold Trials (d) in the test phase performed by the artificial agent and the number of incorrect stimuli between two trials (b) in training (dashed black line) and test (black line) phases are illustrated. Results are averaged across 40 simulations of 1000 stimuli (Mean ± SEM).

Noteworthy, SSRT decreased with regimes of high serotonin concentration ($\eta^2 = 0.914$) during the test phase. Both the right inhibition ($\eta^2 = 0.881$) and the accuracy (training: $\eta^2 = 0.054$; test: $\eta^2 = 0.860$); increased sharply with high levels of serotonin (Fig. 7).

Finally, we investigated the serotonin modulation in relation to the number of trials (both Go and Hold Trials) performed by the artificial agent and the number of incorrect stimuli between the two trials. We found a decrease in the number of trials performed moving from low to high serotoninergic regimes during the test phase ($[\gamma] = 0.1$: 173.47 ± 0.44 (total), 86.1 ± 1.08 (Go Trials), 87.37 ± 1.06 (Hold Trials); $[\gamma] = 0.9$: 134.1 ± 0.94 (total), 66.75 ± 0.58 (Go Trials), 67.35 ± 1.09 (Hold Trials)). In line with this, the number of incorrect stimuli between two trials increased ($[\gamma] = 0.1$: 1.49 ± 0.01; $[\gamma] = 0.9$: 3.03 ± 0.04). Both metrics presented a local maximum (minimum) at intermediate serotonin values (e.g., $[\gamma] = 0.5$) (Fig. 8).

The maximum action value of the inhibition command $Q_{Inh(Max)}$ after learning increased with the increase in serotonin concentration (Mean ± SD; $[\gamma] = 0.1$: 0.68 ± 0.03; $[\gamma] = 0.9$: 3.65 ± 0.22). Finally, the number of stimuli before crossing the M_{thr} increased with the serotonin concentration (Mean ± SD; $[\gamma] = 0.1$: 45.37 ± 9.21; $[\gamma] = 0.9$: 57.57 ± 17.91). Consequently, both the slope of the hold signal meta-value M_{Hold} (Mean ± SD; $[\gamma] = 0.1$: −0.0052 ± 0.001; $[\gamma] = 0.9$: −0.0048 ± 0.004 [stimulus^{-1}]) and the minimum plateau increased with the serotonin concentration (Mean ± SD; $[\gamma] = 0.1$: −3.27 ± 0.32; $[\gamma] = 0.9$: −0.6 ± 0.28). In summary, the training section lasted longer at high serotonin regimes (Table 5(b–c)).

4. Discussion

In this work we implemented a meta-learning framework for optimizing hyperparameters based on reinforcement learning, thus mimicking the dynamics and the interplay of the neurotransmitters in the human brain as described by Doya (Doya, 2002). Our starting point was the control architecture developed by Khamassi and coll. (Khamassi et al., 2013, 2011), which follows meta-learning principles to dynamically adapt the exploration/exploitation rate β according to variations in the average outcome and the occurrence of certain events (e.g., cues).

We followed the same rationale, but we implemented a different control mechanism for β by leveraging the fact that the average uncertainty level of the action selection process (i.e., the entropy of the action probability distribution) is modulated by the tonic level of the D_1 dopamine efferent density (Humphries,

2012). We included a differential computational role for D_1 and D_2 dopamine receptors as well as an external serotoninergic modulation, accounting for a differential effect both in space in the striatum and in the reward temporal scale (Guiard, Mansari, Merali, & Blier, 2008; Tanaka et al., 2004). We showed that the brain-inspired meta-learning mechanism enables flexible and dynamic learning in artificial agents, solving the conflictual instances between motor planning and action execution (e.g., inhibition control) that occur in Stop-Signal and NoGo Paradigms. The tasks were designed to manipulate different modes of action inhibition, such as refraining from responding (action restraint in NoGo Trials) and cancelling an already-prepared response (action cancellation in Stop-Signal Trials) (Eagle, Bari, & Robbins, 2008).

The study led to five key results.

First, the agent learned (i) an internal inhibition command temporally associated with the hold signal that increases during the training phase, and (ii) to use it only when it is necessary to switch to high β values that drive the system in an exploitation regime. The agent tunes its action inhibition process, favoring the ongoing ramp-up of the action inhibition command and, consequently, improving its global accuracy and right inhibition both in NoGo and Stop-Signal Paradigms after learning (Figs. 4, 5). In line with this, in the Stop-Signal Paradigm the artificial agent reduced the time needed to cancel the command as suggested by the reduction of the SSRT. In Khamassi's model the behavioral shift that favors the flexibility of the action selection process is implemented through reset exploration in response to such uncertain cues and events. On the other hand, we implemented the withholding from responding by exploiting an internal inhibition command that is established over time (in Fig. 2 the green line increases over time). Thus, in this context, action inhibition did not arise in the artificial agent from an expansive strategy that requires the exploration of the action space.

Secondly, learning affected RT in the NoGo Paradigm. In the race model, cancelling an action is more complex than executing the motor command, and it becomes more difficult as the latency of the activation of motor commands decreases. In fact, RT in NoGo failure Trials were faster than in Go Trials and NoGo correct Trials (Fig. 4). (Boucher, Palmeri, Logan, & Schall, 2007; Mosher et al., 2021b; Schmidt, Leventhal, Mallet, Chen, & Berke, 2013; Wiecki & Frank, 2013). Interestingly, we had neither entrainment nor habituation in the Go Trials (Fig. 4).

Third, learning also affected aspects of artificial agents that are less related to the behavioral outcome and more to the dynamics of the underlying network. Indeed, both the confidence (i.e., $P_{\max}$) and the vigor (i.e., $\mathrm{PMC}_{\mathrm{peak}}$) of the selected action increased with learning in both the Go Trials and the Hold correct Trials. Vigor and confidence were uncoupled in the Hold failure Trials, where the former decreased drastically.

Fourth, serotonin modulated the overall level of the inhibition control process, reducing the impulsive behavior at high concentration regimes. The right inhibition increased overall (pooling all delays together) and the detrimental effect of the delay on the right inhibition was attenuated as serotonin increased (Pasquereau & Turner, 2017b). This interaction between serotonin and the delay in right inhibition (Table 6, Table 7) might be explained by the fact that, when the serotonin concentration is high, the control inhibition of the agent is sharper, so much so that it becomes virtually insensitive to the delay of the hold cue. Moreover, the RT in the Hold failure Trials became shorter as the serotonin increased, without significantly changing its relationship with the delay. Interestingly, the agent became sufficiently fast that it was inhibited at high levels of serotonin, as corroborated by the negative relationship between SSRT and serotonin. Increased control of inhibition induced by serotonin also reverberated during Go Trials, reducing their execution.

Fifth, learning with high levels of serotonin increased the model's performance (e.g., right inhibition, accuracy and speed) at the expense of vigor and confidence (crossing effect between training and test lines in Fig. 7) and the number of executed trials (Fig. 8).

In our framework, the serotonin modulation prevalently mimics an increasing concentration of the drug treatment (e.g., citalopram). Preclinical studies suggest that serotonin plays a role in action restraint thus reducing the impulsive response (Homberg et al., 2007; Robinson et al., 2008). Therefore, central serotonin depletion (e.g., by acute depletion of tryptophan) alters NoGo inhibition. However, the link between serotonin and action cancellation (Stop-Signal Paradigm) appears more puzzling and less consolidated. Animal and human control studies have shown mixed results regarding how serotonin impacts on Stop-signal tasks (Bari, Eagle, Mar, Robinson, & Robbins, 2009; Eagle, Baunez, et al., 2008). Interestingly, selective serotonin reuptake inhibitors improved impulsivity both in action restraint and action cancellation in individuals with serotonergic deficits, such as Parkinsonian patients (Ye et al., 2014). This suggests a potential relationship between the impact of serotonin manipulation and the baseline serotonin functioning because such an acute change in serotonin transmission may be more impactful on dysregulated serotonergic conditions.

Although reinforcement learning gives a well-established framework for autonomous agents to acquire adaptive behaviors based on reward feedback, setting the meta-parameters to match the demands of the task and the environment is crucial. In fact, most RL studies, do not include meta-learning principles, and they rely on long, computationally expensive and hand-tuned optimization procedures that are highly prone to the noise in the environment (Eriksson, Capi, & Doya, 2003; Sutton & Barto, 2018).

A major precursor of meta-learning as learning of meta-parameters in the context of reinforcement learning is Doya with the theory of neuromodulation, whose aim is to optimize meta-parameters following brain-inspired rules (Doya, 2002; Schweighofer & Doya, 2003). Seminal works stemmed from this rationale such as the model developed by Khamassi and colleagues (Khamassi et al., 2013, 2011) which implements a brain-inspired meta-learning hyperparameter optimization framework for reinforcement learning by linking the exploration/exploitation meta-parameter β with the activity of two types of neurons (Dehaene et al., 1998; Holroyd & Coles, 2002) (i.e., correct and error neurons) that react to positive and negative temporal difference errors $\delta(t)$, respectively. Authors tested the model both in silico and in a humanoid robot (i-Cub) in two problem solving tasks where environmental uncertainties are included either by changing the cue or cheating.

In the Cyber Rodent project (Doya & Uchibe, 2005), i.e., colony of robots that can recharge themselves from battery packs in the environment and to communicate with each other through their infrared communication port, brain-inspired evolutionary principles, i.e., genetic algorithms, have been adopted to investigate the underlying mechanisms of self-reproduction, self-preservation as well as foraging in artificial agents (Capi & Doya, 2005; Doya & Uchibe, 2005; Eriksson et al., 2003). This evolutionary approach was adopted to co-evolve the meta-parameters (e.g., exploration/exploitation rate, learning rate and temporal discounting factor) in synergy with shaping rewards, significantly accelerating the learning curve in foraging (Elfwing, Uchibe, Doya, & Christensen, 2008) and mating (Elfwing, Uchibe, & Doya, 2009). Lowe and Ziemke (2011) used genetic algorithms to investigate meta-learned exploration and planning in a multi-episode two-armed bandit navigation problem under different representations (e.g., absence/presence) of external rewards and punishments.

In future works, we plan to build on the results obtained in the present study.

First, we will replicate what we did in simulation in robotics applications with i-Cub by implementing a conflictual oculomotor saccadic movement task. At the beginning of each trial, a fixation point and two peripheral visual targets (e.g., triangle on the left and square on the right) appear on the display. Then, an onset cue, which can be either a triangle or a square, appears in the middle of the display and the robot is asked to perform a saccadic movement towards the direction that matches the onset cue and the correct peripherical target. In 50% of trials (Hold Trials) we will deliver a hold signal (e.g., black cross) either before the movement onset (action restraint), or at a range of delays during the movement (action cancellation). We expect the robot to learn to successfully inhibit the saccadic command when the hold signal appears.

Second, instead of using the same parameter values to implement the serotonergic modulation of the D_1 and D_2 dopamine receptors, we will evaluate the effects of the differential and selective relationship between serotonin and D_1 and D_2 dopamine receptors. Finally, we will implement a closed-loop regulation between serotonin and dopamine including the dopaminergic feedback on the serotonin release. Such brain-inspired approaches could pave the way to the design of robotic systems that dynamically adapt their behavior to unpredictable constraints in the environment, such as the countermanding of the motor command.

Although in the present paper we focused on brain-inspired meta-learning principles for learning action inhibition for artificial agents, our model could contribute to the neurophysiological description of the neural processes that underlie action inhibition. A series of virtual ablation experiments involving different regions of the model could lead to interesting predictions to test whether the disruption of the integrity of the: i) LPFC may lead to higher failure rates of inhibition due to the lack of involvement of the subcortical regions; ii) ACC may cause a loss of learning capacity in the system due to either a lack of storage and updating of action values or the loss of plasticity in the fronto-striatal connections; iii) PPC may cause a failure in the action selection process due to suboptimal visual stimulus encoding; iv) basal ganglia (BG) may lead to multiple unwanted motor commands due to the failure of the winner-takes-all mechanism.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

F.R. & E.F. have received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3). G.S. has received funding from the European Union's 2020 Framework Programme for Research and Innovation under the Marie Skłodowska-Curie grant agreement No. 838861 (Predictive Robots). M.V. has received internal fundings from Sant'Anna School of Advanced Studies.

Appendix

In the main text we claimed that a sigmoid relationship $\beta = \beta([D_1]\cdot\delta(t))$ is a reasonable choice to ensure that the entropy H is a decrescent monotonic function with respect to the dopamine efferent density $[D_1]$. Here, we establish under what circumstances a general function $\beta([D_1]\cdot\delta(t))$ satisfies this criterion.

We redefine the argument $D = [D_1]\cdot\delta(t) \in R$ and we use the definition of entropy H,

$$H(D) = H(P(\beta(D), Q)) = -\sum_{i=1}^{3} p_i \log_2 p_i = -\sum_{i=1}^{3} P(\beta(D)\cdot Q_i)\log_2(P(\beta(D)\cdot Q_i)) \quad (1)$$

where $Q_i = Q(a_i, s)$ is the ith action value and,

$$p_i = P(\beta(D)\cdot Q_i) = \frac{e^{(\beta(D)\cdot Q_i)}}{\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}}; \qquad 0 \le pi \le 1\ \forall i = 1, 2, 3 \quad (2)$$

is the Boltzmann Softmax Function that computes the ith action probability. For definition, the entropy is always positive $H(D) \ge 0 \forall D$. The requirement of $H(D)$ as such it is a decrescent monotonic equation can be equivalent formulated imposing a negative sign on the first derivate of $H(D)$,

$$\frac{\partial H(D)}{\partial D} < 0 \quad (3)$$

We insert (2) in (1) to obtain,

$$H(P(\beta(D), Q)) = -\sum_{i=1}^{3} \frac{e^{(\beta(D)\cdot Q_i)}}{\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}} \cdot \log_2\left(\frac{e^{(\beta(D)\cdot Q_i)}}{\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}}\right) \quad (4)$$

For the logarithmic rule $\log_2(x) = \frac{\ln(x)}{\ln(2)}$ and,

$$\ln\left(e^{(\beta(D)\cdot Q_i)}\right) - \ln\left(\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}\right) = \beta(D)\cdot Q_i - \ln\left(\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}\right) \quad (5)$$

And we can rewrite (4) as,

$$\begin{aligned} H(P(\beta(D), Q)) &= -\frac{1}{\ln(2)}\sum_{i=1}^{3} \frac{e^{(\beta(D)\cdot Q_i)}}{\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}} \cdot \beta(D)\cdot Q_i \\ &\quad + \frac{\ln\left(\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}\right)}{\ln(2)} = \\ &= -\frac{1}{\ln(2)}\beta(D)\cdot\langle Q\rangle_P + \frac{ln\left(\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}\right)}{\ln(2)} \\ &= H_1(P(\beta(D), Q)) + H_2(P(\beta(D), Q)) \end{aligned} \quad (6)$$

where $\langle Q\rangle_P = \sum_{i=1}^{3} \frac{e^{(\beta(D)\cdot Q_i)}}{\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}} \cdot Q_i = \sum_{i=1}^{3} p_i \cdot Q_i = \langle Q, P\rangle$ is the expected value of the action value Q with respect to the probability distribution P (equivalently the dot product between Q and P).

We compute the derivative of the entropy expressed as above,

$$\frac{\partial H(P(\beta(D), Q))}{\partial D} = \frac{\partial H_1(P(\beta(D), Q))}{\partial D} + \frac{\partial H_2(P(\beta(D), Q))}{\partial D} \quad (7)$$

where:

$$\frac{\partial H_2(P(\beta(D), Q))}{\partial D} = \frac{\partial}{\partial D}\left(\frac{\ln\left(\sum_{i=1}^{3} e^{(\beta(D)\cdot Q_i)}\right)}{\ln(2)}\right)$$

$$= \frac{1}{\ln(2)} \frac{\beta'(D) \cdot Q_i \cdot \sum_{i=1}^{3} e^{(\beta(D) \cdot Q_i)}}{\left(\sum_{i=1}^{3} e^{(\beta(D) \cdot Q_i)}\right)} = \frac{\beta'(D) \cdot \langle Q \rangle_P}{\ln(2)} \tag{8}$$

and,

$$\frac{\partial H_1(P(\beta(D), Q))}{\partial D} = -\frac{1}{\ln(2)} \left[\beta'(D) \cdot \langle Q \rangle_P + \beta(D) \cdot \frac{\partial \langle Q \rangle_P}{\partial D}\right] \tag{9}$$

The derivative of H can be expressed as,

$$\frac{\partial H(P)}{\partial D} = -\frac{1}{\ln(2)} \left[\beta'(D) \cdot \langle Q \rangle_P + \beta(D) \frac{\partial \langle Q \rangle_P}{\partial D} - \beta'(D) \cdot \langle Q \rangle_P\right] = -\frac{1}{\ln(2)} \beta(D) \cdot \frac{\partial \langle Q \rangle_P}{\partial D} \tag{10}$$

To compute the derivative of $\langle Q \rangle_P$ with respect to D we use the chain rule:

$$\frac{\partial \langle Q \rangle_P}{\partial D} = \frac{\partial \langle Q \rangle_P}{\partial \beta} \frac{\partial \beta}{\partial D} = \frac{\partial \langle Q \rangle_P}{\partial \beta} \cdot \beta'(D) \tag{11}$$

We plug in (11) the definition of $\langle Q \rangle_P$:

$$\frac{\partial \langle Q \rangle_P}{\partial \beta} = \frac{\partial \sum_{i=1}^{3} Q_i \cdot p_i}{\partial \beta} = \sum_{i=1}^{3} Q_i \cdot \frac{\partial p_i}{\partial \beta} = \sum_{i=1}^{3} Q_i \cdot (Q_i - \langle Q \rangle_P) \cdot p_i = \langle Q(Q - \langle Q \rangle_P) \rangle_P = \left\langle Q^2 \right\rangle_P - \langle \langle Q \rangle_P Q \rangle_P = \left\langle Q^2 \right\rangle_P - \langle Q \rangle_P^2 = \sigma^2(Q)_P \tag{12}$$

where:

$$\frac{\partial p_i}{\partial \beta} = \frac{Q_i \cdot e^{(\beta(D) \cdot Q_i)} \cdot \sum_{i=1}^{3} e^{(\beta(D) \cdot Q_i)} - \sum_{i=1}^{3} Q_i \cdot e^{(\beta(D) \cdot Q_i)} \cdot e^{(\beta(D) \cdot Q_i)}}{\left(\sum_{i=1}^{3} e^{(\beta(D) \cdot Q_i)}\right)^2} = \frac{Q_i \cdot e^{(\beta(D) \cdot Q_i)}}{\sum_{i=1}^{3} e^{(\beta(D) \cdot Q_i)}} - \frac{e^{(\beta(D) \cdot Q_i)} \cdot \sum_{i=1}^{3} Q_i \cdot e^{(\beta(D) \cdot Q_i)}}{\sum_{i=1}^{3} e^{(\beta(D) \cdot Q_i)} \cdot \sum_{i=1}^{3} e^{(\beta(D) \cdot Q_i)}} = p_i \cdot Q_i - p_i \cdot \langle Q \rangle_P = p_i \cdot (Q_i - \langle Q \rangle_P) \tag{13}$$

Finally, we can compute the derivative of $\langle Q \rangle_P$ as,

$$\frac{\partial \langle Q \rangle_P}{\partial D} = \frac{\partial \langle Q \rangle_P}{\partial \beta} \frac{\partial \beta}{\partial D} = \frac{\partial \langle Q \rangle_P}{\partial \beta} \cdot \beta'(D) = \sigma^2(Q)_P \cdot \beta'(D) \tag{14}$$

We substitute (14) into (10) to obtain the derivative of H as,

$$\frac{\partial H(P(\beta(D), Q))}{\partial D} = -\frac{1}{\ln(2)} \beta(D) \cdot \frac{\partial \langle Q \rangle_P}{\partial D} = -\frac{\beta(D) \cdot \beta'(D) \cdot \sigma^2(Q)_P}{\ln(2)} \tag{15}$$

Since $\ln(2) \geq 0 \bigwedge \sigma^2(Q)_P \geq 0 \forall Q$ for the definition of variance, the derivative of entropy is negative,

$$\frac{\partial H(D)}{\partial D} \leq 0 \Longleftrightarrow \beta(D) \cdot \beta'(D) \cdot \sigma^2(Q)_P \geq 0 \tag{16}$$

and since $\beta(D) \geq 0 \Longrightarrow \beta'(D) > 0$.

This means that $\beta(D)$ must be an increasing monotonic function. Moreover, integrating (16) we can compute H as,

$$H(P(\beta(D), Q)) = -\frac{\int_0^D \beta(D) \cdot \beta'(D) \cdot \sigma^2(Q)_P}{\ln(2)} dD + C \tag{17}$$

where C is the constant of integration,

$$C = H(0) + \frac{1}{\ln(2)} \int_0^0 \beta(0) \cdot \beta'(0) \cdot \sigma^2(Q)_{P(\beta(0))} dD = H(0) = H(U_3(Q)) = \log_2(3) \tag{18}$$

where U is the discrete uniform distribution with 3 values ($U_3(Q) = P(\beta(0), Q)$) and the entropy of a uniform distribution of n states is $\log_2(n)$.

Furthermore, the entropy $H(P(\beta(D), Q))$ should be positive,

$$H((\beta(D), Q)) \geq 0 \Longrightarrow \frac{\int_0^D \beta(D) \cdot \beta'(D) \cdot \sigma^2(Q)_P dD}{\ln(2)} \leq C \Longrightarrow$$

$$\Longrightarrow \int_0^D \beta(D) \cdot \beta'(D) \cdot \sigma^2(Q)_P dD \leq \ln(2) \cdot C = \ln(2) \cdot \log_2(3)$$

Since the variance is finite $\left(\exists M \sigma^2(Q)_P \leq M\right)$ and the integration interval is limited, the integral could not be limited if and only if the integrand $\beta(D) \cdot \beta'(D)$ is not limited. Thus, $\beta'(D)$ and $\beta(D)$ must be limited (i.e., differentiable and with saturation). Hence, $\beta(D)$ must be positive, limited, differentiable and increasing monotonic. The sigmoid function satisfies all these criteria.

Finally, we can write the entropy H as,

$$H(P(\beta(D), Q)) = -\frac{\int_0^D \beta(D) \cdot \beta'(D) \cdot \sigma^2(Q)_P dD}{\ln(2)} + \log_2(3) = -\frac{\int_0^D \beta(D) \cdot \beta'(D) \cdot \sigma^2(Q)_P dD - \ln(3)}{\ln(2)} \tag{19}$$

References

Akam, T., Rodrigues-Vaz, I., Marcelo, I., Zhang, X., Pereira, M., Oliveira, R. F., et al. (2021). The anterior cingulate cortex predicts future states to mediate model-based action selection. *Neuron*, *109*(1), 149–163. http://dx.doi.org/10.1016/j.neuron.2020.10.013, e7.

Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in Neurosciences*, *13*(7), 266–271. http://dx.doi.org/10.1016/0166-2236(90)90107-L.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381. http://dx.doi.org/10.1146/annurev.ne.09.030186.002041.

Alexander, M. E., & Wickens, J. R. (1993). Analysis of striatal dynamics: The existence of two modes of behaviour. *Journal of Theoretical Biology*, *163*(4), 413–438. http://dx.doi.org/10.1006/jtbi.1993.1128.

Amiez, C. (2006). Local morphology predicts functional organization of the Dorsal Premotor Region in the human brain. *Journal of Neuroscience*, *26*(10), 2724–2731. http://dx.doi.org/10.1523/JNEUROSCI.4739-05.2006.

Amiez, C., Joseph, J.-P., & Procyk, E. (2005). Anterior cingulate error-related activity is modulated by predicted reward. *European Journal of Neuroscience*, *21*(12), 3447–3452. http://dx.doi.org/10.1111/j.1460-9568.2005.04170.x.

Apicella, P., Ljungberg, T., Scarnati, E., & Schultz, W. (1991). Responses to reward in monkey dorsal and ventral striatum. *Experimental Brain Research*, *85*(3), http://dx.doi.org/10.1007/BF00231732.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of LOCUS CoeruleUS-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*(1), 403–450. http://dx.doi.org/10.1146/annurev.neuro.28.061604.135709.

Avery, M. C., & Krichmar, J. L. (2017). Neuromodulatory systems and their interactions: A review of models, theories, and experiments. *Frontiers in Neural Circuits*, *11*(108), http://dx.doi.org/10.3389/fncir.2017.00108.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200. http://dx.doi.org/10.1016/j.tics.2008.02.004.

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal cortex and the discovery of abstract action rules. *Neuron*, *66*(2), 315–326. http://dx.doi.org/10.1016/j.neuron.2010.03.025.

Band, G. P. H., van der Molen, M. W., & Logan, G. D. (2003). Horse-race model simulations of the stop-signal procedure. *Acta Psychologica*, *112*(2), 105–142. http://dx.doi.org/10.1016/S0001-6918(02)00079-3.

Bari, A., Eagle, D. M., Mar, A. C., Robinson, E. S. J., & Robbins, T. W. (2009). Dissociable effects of noradrenaline, dopamine, and serotonin uptake blockade on stop task performance in rats. *Psychopharmacology*, *205*(2), 273–283. http://dx.doi.org/10.1007/s00213-009-1537-0.

Baxter, J. (1998). Theoretical models of learning to learn. In S. Thrun, & L. Pratt (Eds.), *Learning to learn* (pp. 71–94). Springer US., http://dx.doi.org/10.1007/978-1-4615-5529-2_4.

Beninger, R. J. (1983). The role of dopamine in locomotor activity and learning. *Brain Research Reviews*, *6*(2), 173–196. http://dx.doi.org/10.1016/0165-0173(83)90038-3.

Berger, M., Gray, J. A., & Roth, B. L. (2009). The expanded biology of serotonin. *Annual Review of Medicine*, *60*(355), http://dx.doi.org/10.1146/annurev.med.60.042307.110802.

Berns, G. S., & Sejnowski, T. J. (1996). How the basal ganglia make decisions. In A. R. Damasio, H. Damasio, & Y. Christen (Eds.), *Neurobiology of decision-making* (pp. 101–113). Springer Berlin Heidelberg., http://dx.doi.org/10.1007/978-3-642-79928-0_6.

Berridge, K. C. (2004). Motivation concepts in behavioral neuroscience. *Physiology & Behavior*, *81*(2), 179–209. http://dx.doi.org/10.1016/j.physbeh.2004.02.004.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, *28*(3), 309–369. http://dx.doi.org/10.1016/S0165-0173(98)00019-8.

Binas, J., Rutishauser, U., Indiveri, G., & Pfeiffer, M. (2014). Learning and stabilization of winner-take-all dynamics through interacting excitatory and inhibitory plasticity. *Frontiers in Computational Neuroscience*, *8*, http://dx.doi.org/10.3389/fncom.2014.00068.

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Computation*, *19*(2), 442–477. http://dx.doi.org/10.1162/neco.2007.19.2.442.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. http://dx.doi.org/10.1016/j.tics.2019.02.006.

Boucher, L., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2007). Inhibitory control in mind and brain: An interactive race model of countermanding saccades. *Psychological Review*, *114*(2), 376–397. http://dx.doi.org/10.1037/0033-295X.114.2.376.

Boureau, Y.-L., & Dayan, P. (2011). Opponency revisited: Competition and cooperation between dopamine and serotonin. *Neuropsychopharmacology*, *36*(1), 74–97. http://dx.doi.org/10.1038/npp.2010.151.

Bouret, S., & Sara, S. J. (2005). Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, *28*(11), 574–582. http://dx.doi.org/10.1016/j.tins.2005.09.002.

Bromberg-Martin, E. S., Matsumoto, M., & Hikosaka, O. (2010). Dopamine in motivational control: Rewarding, aversive, and alerting. *Neuron*, *68*(5), 815–834. http://dx.doi.org/10.1016/j.neuron.2010.11.022.

Caligiore, D., Arbib, M. A., Miall, R. C., & Baldassarre, G. (2019). The super-learning hypothesis: Integrating learning processes across cortex, cerebellum and basal ganglia. *Neuroscience and Biobehavioral Reviews*, *100*, 19–34. http://dx.doi.org/10.1016/j.neubiorev.2019.02.008, Scopus.

Cannon, C. M., & Palmiter, R. D. (2003). Reward without Dopamine. *The Journal of Neuroscience*, *23*(34), 10827–10831. http://dx.doi.org/10.1523/JNEUROSCI.23-34-10827.2003.

Capi, G., & Doya, K. (2005). Evolution of neural architecture fitting environmental dynamics. *Adaptive Behavior*, *13*(1), 53–66. http://dx.doi.org/10.1177/105971230501300103.

Carr, D. B., & Sesack, S. R. (2000). Projections from the rat prefrontal cortex to the Ventral Tegmental Area: Target specificity in the synaptic associations with Mesoaccumbens and Mesocortical neurons. *The Journal of Neuroscience*, *20*(10), 3864–3873. http://dx.doi.org/10.1523/JNEUROSCI.20-10-03864.2000.

Chamberlain, S. R. (2006). Neurochemical modulation of response inhibition and probabilistic learning in humans. *Science*, *311*(5762), 861–863. http://dx.doi.org/10.1126/science.1121218.

Chen, W., Hemptinne, C. de., Miller, A. M., Leibbrand, M., Little, S. J., Lim, D. A., et al. (2020). Prefrontal-subthalamic hyperdirect pathway modulates movement inhibition in humans. *Neuron*, *106*(4), 579–588. http://dx.doi.org/10.1016/j.neuron.2020.02.012.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *362*(1481), 933–942. http://dx.doi.org/10.1098/rstb.2007.2098.

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering and generalizing task-set structure. *Psychological Review*, *120*(1), 190–229. http://dx.doi.org/10.1037/a0030852.

Collins, A., & Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology*, *10*(3), Article e1001293. http://dx.doi.org/10.1371/journal.pbio.1001293.

Cools, R., Nakamura, K., & Daw, N. D. (2011). Serotonin and dopamine: Unifying affective, activational, and decision functions. *Neuropsychopharmacology*, *36*(1), 98–113. http://dx.doi.org/10.1038/npp.2010.121.

Daw, N. D., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, *16*(2), 199–204. http://dx.doi.org/10.1016/j.conb.2006.03.006.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4–6), 603–616. http://dx.doi.org/10.1016/S0893-6080(02)00052-7.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. http://dx.doi.org/10.1038/nn1560.

Daw, N. D., & Tobler, P. N. (2014). Value learning through reinforcement. In *Neuroeconomics* (pp. 283–298). Elsevier, http://dx.doi.org/10.1016/B978-0-12-416008-8.00015-2.

De Deurwaerdère, P., Chagraoui, A., & Di Giovanni, G. (2021). Serotonin/dopamine interaction: Electrophysiological and neurochemical evidence. In *Progress in brain research vol. 261* (pp. 161–264). Elsevier., http://dx.doi.org/10.1016/bs.pbr.2021.02.001.

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proceedings of the National Academy of Sciences*, *95*(24), 14529–14534. http://dx.doi.org/10.1073/pnas.95.24.14529.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7–8), 961–974. http://dx.doi.org/10.1016/S0893-6080(99)00046-5.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*(6), 732–739. http://dx.doi.org/10.1016/s0959-4388(00)00153-7.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4), 495–506. http://dx.doi.org/10.1016/S0893-6080(02)00044-8.

Doya, K., & Uchibe, E. (2005). The cyber rodent project: Exploration of adaptive mechanisms for self-preservation and self-reproduction. *Adaptive Behavior*, *13*(2), 149–160. http://dx.doi.org/10.1177/105971230501300206.

Dreher, J.-C., & Berman, K. F. (2002). Fractionating the neural substrate of cognitive control processes. *Proceedings of the National Academy of Sciences*, *99*(22), 14595–14600. http://dx.doi.org/10.1073/pnas.222193299.

Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). Rl^2: Fast reinforcement learning via slow reinforcement learning. arXiv:1611.02779 [Cs, Stat]. http://arxiv.org/abs/1611.02779.

Dunovan, K., & Verstynen, T. (2016). Believer-skeptic meets actor-critic: Rethinking the role of basal ganglia pathways during decision-making and reinforcement learning. *Frontiers in Neuroscience*, *10*, http://dx.doi.org/10.3389/fnins.2016.00106.

Eagle, D. M., Bari, A., & Robbins, T. W. (2008). The neuropsychopharmacology of action inhibition: Cross-species translation of the stop-signal and go/no-go tasks. *Psychopharmacology*, *199*(3), 439–456. http://dx.doi.org/10.1007/s00213-008-1127-6.

Eagle, D. M., & Baunez, C. (2010). Is there an inhibitory-response-control system in the rat? Evidence from anatomical and pharmacological studies of behavioral inhibition. *Neuroscience & Biobehavioral Reviews*, *34*(1), 50–72. http://dx.doi.org/10.1016/j.neubiorev.2009.07.003.

Eagle, D. M., Baunez, C., Hutcheson, D. M., Lehmann, O., Shah, A. P., & Robbins, T. W. (2008). Stop-signal reaction-time task performance: Role of prefrontal cortex and subthalamic nucleus. *Cerebral Cortex*, *18*(1), 178–188. http://dx.doi.org/10.1093/cercor/bhm044.

Elfwing, S., Uchibe, E., & Doya, K. (2009). Emergence of different mating strategies in artificial embodied evolution. In C. S. Leung, M. Lee, & J. H. Chan (Eds.), *Neural information processing, vol. 5864* (pp. 638–647). Springer Berlin Heidelberg., http://dx.doi.org/10.1007/978-3-642-10684-2_71.

Elfwing, S., Uchibe, E., Doya, K., & Christensen, H. I. (2008). Co-evolution of shaping rewards and meta-parameters in reinforcement learning. *Adaptive Behavior*, *16*(6), 400–412. http://dx.doi.org/10.1177/1059712308092835.

Eriksson, A., Capi, G., & Doya, K. (2003). Evolution of meta-parameters in reinforcement learning algorithm. In *Proceedings 2003 IEEE/RSJ international conference on intelligent robots and systems (IROS 2003) (Cat. No. 03CH37453), vol. 1* (pp. 412–417). http://dx.doi.org/10.1109/IROS.2003.1250664.

Fischer, A. G., & Ullsperger, M. (2017). An update on the role of serotonin and its interplay with dopamine for reward. *Frontiers in Human Neuroscience*, *11*(484), http://dx.doi.org/10.3389/fnhum.2017.00484.

Fluxe, K., Hökfelt, T., Johansson, O., Jonsson, G., Lidbrink, P., & Ljungdahl, A. (1974). The origin of the dopamine nerve terminals in limbic and frontal cortex. Evidence for meso-cortico dopamine neurons. *Brain Research*, *82*(2), 349–355. http://dx.doi.org/10.1016/0006-8993(74)90618-0.

Guiard, B. P., Mansari, M. El., Merali, Z., & Blier, P. (2008). Functional interactions between dopamine, serotonin and norepinephrine neurons: An in-vivo electrophysiological study in rats with monoaminergic lesions. *International Journal of Neuropsychopharmacology*, *11*(5), 625–639. http://dx.doi.org/10.1017/S1461145707008383.

Hasselmo, M. E., & Bower, J. M. (1993). Acetylcholine and memory. *Trends in Neurosciences*, *16*(6), 218–222. http://dx.doi.org/10.1016/0166-2236(93)90159-J.

Hasselmo, M., & Schnell, E. (1994). Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region CA1: Computational modeling and brain slice physiology. *The Journal of Neuroscience*, *14*(6), 3898–3914. http://dx.doi.org/10.1523/JNEUROSCI.14-06-03898.1994.

Heekeren, H. R., Marrett, S., & Ungerleider, L. G. (2008). The neural systems that mediate human perceptual decision making. *Nature Reviews Neuroscience*, *9*(6), 467–479. http://dx.doi.org/10.1038/nrn2374.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. http://dx.doi.org/10.1037/0033-295X.109.4.679.

Homberg, J. R., Pattij, T., Janssen, M. C. W., Ronken, E., De Boer, S. F., Schoffelmeer, A. N. M., et al. (2007). Serotonin transporter deficiency in rats improves inhibitory control but not behavioural flexibility: Serotonin transporter knockout and impulse control. *European Journal of Neuroscience*, *26*(7), 2066–2073. http://dx.doi.org/10.1111/j.1460-9568.2007.05839.

Horvitz, J. C. (2000). Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience*, *96*(4), 651–656. http://dx.doi.org/10.1016/S0306-4522(00)00019-1.x.

Houk, J. C., Davis, J. L., & Beiser, D. G. (2019). Models of information processing in the Basal Ganglia. http://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9780262275774.

Humphries, M. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience*, *6*, http://dx.doi.org/10.3389/fnins.2012.00009.

Humphries, M. D. (2014). Basal ganglia: Mechanisms for action selection. In D. Jaeger, & R. Jung (Eds.), *Encyclopedia of computational neuroscience* (pp. 1–7). New York: Springer, http://dx.doi.org/10.1007/978-1-4614-7320-6_83-3.

Ishii, S., Yoshida, W., & Yoshimoto, J. (2002). Control of exploitation–exploration meta-parameter in reinforcement learning. *Neural Networks*, *15*(4–6), 665–687. http://dx.doi.org/10.1016/S0893-6080(02)00056-4.

Kaplan, R., Schuck, N. W., & Doeller, C. F. (2017). The role of mental maps in decision-making. *Trends in Neurosciences*, *40*(5), 256–259. http://dx.doi.org/10.1016/j.tins.2017.03.002.

Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J., & Rushworth, M. F. S. (2006). Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience*, *9*(7), 940–947. http://dx.doi.org/10.1038/nn1724.

Kesteren, M. T. R. van, Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. http://dx.doi.org/10.1016/j.tins.2012.02.001.

Khamassi, M., Enel, P., Dominey, P. F., & Procyk, E. (2013). Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. In *Progress in brain research vol. 202* (pp. 441–464). Elsevier, http://dx.doi.org/10.1016/B978-0-444-62604-2.00022-8.

Khamassi, M., Lallée, S., Enel, P., Procyk, E., & Dominey, P. F. (2011). Robot cognitive control with a neurophysiologically inspired reinforcement learning model. *Frontiers in Neurorobotics*, *5*, http://dx.doi.org/10.3389/fnbot.2011.00001.

Kim, H. R., Malik, A. N., Mikhael, J. G., Bech, P., Tsutsui-Kimura, I., Sun, F., et al. (2020). A unified framework for dopamine signals across timescales. *Cell*, http://dx.doi.org/10.1016/j.cell.2020.11.013.

Krichmar, J. L. (2008). The neuromodulatory system: A framework for survival and adaptive behavior in a challenging world. *Adaptive Behavior*, *16*(6), 385–399. http://dx.doi.org/10.1177/1059712308095775.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. http://dx.doi.org/10.1126/science.aab3050.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, http://dx.doi.org/10.3389/fpsyg.2013.00863.

Lapidus, K. A. B., Stern, E. R., Berlin, H. A., & Goodman, W. K. (2014). Neuromodulation for obsessive–compulsive disorder. *Neurotherapeutics*, *11*(3), 485–495. http://dx.doi.org/10.1007/s13311-014-0287-9.

Lee, B., Groman, S., London, E. D., & Jentsch, J. D. (2007). Dopamine D2/D3 receptors play a specific role in the reversal of a learned visual discrimination in monkeys. *Neuropsychopharmacology*, *32*(10), 2125–2134. http://dx.doi.org/10.1038/sj.npp.1301337.

Leisman, G., Braun-Benjamin, O., & Melillo, R. (2014). Cognitive-motor interactions of the basal ganglia in development. *Frontiers in Systems Neuroscience*, *8*, http://dx.doi.org/10.3389/fnsys.2014.00016.

Lowe, R., & Ziemke, T. (2011). The feeling of action tendencies: On the emotional regulation of goal-directed behavior. *Frontiers in Psychology*, *2*, http://dx.doi.org/10.3389/fpsyg.2011.00346.

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature*, *447*(7148), 1111–1115. http://dx.doi.org/10.1038/nature05860.

Middleton, F., & Strick, P. (1994). Anatomical evidence for cerebellar and basal ganglia involvement in higher cognitive function. *Science*, *266*(5184), 458–461. http://dx.doi.org/10.1126/science.7939688.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *16*(5), 1936–1947.

Mosher, C. P., Mamelak, A. N., Malekmohammadi, M., Pouratian, N., & Rutishauser, U. (2021a). Distinct roles of dorsal and ventral subthalamic neurons in action selection and cancellation. *Neuron*, http://dx.doi.org/10.1016/j.neuron.2020.12.025.

Mosher, C. P., Mamelak, A. N., Malekmohammadi, M., Pouratian, N., & Rutishauser, U. (2021b). Distinct roles of dorsal and ventral subthalamic neurons in action selection and cancellation. *Neuron*, *109*(5), 869–881. http://dx.doi.org/10.1016/j.neuron.2020.12.025, e6.

Nagel, K. I., & Wilson, R. I. (2016). Mechanisms underlying population response dynamics in inhibitory interneurons of the drosophila antennal lobe. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *36*(15), 4325–4338. http://dx.doi.org/10.1523/JNEUROSCI.3887-15.2016.

Nakamura, K., Matsumoto, M., & Hikosaka, O. (2008). Reward-dependent modulation of neuronal activity in the primate dorsal raphe nucleus. *Journal of Neuroscience*, *28*(20), 5331–5343. http://dx.doi.org/10.1523/JNEUROSCI.0021-08.2008.

Padoa-Schioppa, C., & Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, *441*(7090), 223–226. http://dx.doi.org/10.1038/nature04676.

Partridge, J. G., Apparsundaram, S., Gerhardt, G. A., Ronesi, J., & Lovinger, D. M. (2002). Nicotinic acetylcholine receptors interact with dopamine in induction of striatal long-term depression. *The Journal of Neuroscience*, *22*(7), 2541–2549. http://dx.doi.org/10.1523/JNEUROSCI.22-07-02541.2002.

Pasquereau, B., & Turner, R. S. (2017a). A selective role for ventromedial subthalamic nucleus in inhibitory control. *ELife*, *6*, Article e31627. http://dx.doi.org/10.7554/eLife.31627.

Pasquereau, B., & Turner, R. S. (2017b). A selective role for ventromedial subthalamic nucleus in inhibitory control. *ELife*, *6*, Article e31627. http://dx.doi.org/10.7554/eLife.31627.

Pfeifer, R., Lungarella, M., & Iida, F. (2007). Self-organization, embodiment, and biologically inspired robotics. *Science*, *318*(5853), 1088–1093. http://dx.doi.org/10.1126/science.1145803.

Poulin, J.-F., Caronia, G., Hofer, C., Cui, Q., Helm, B., Ramakrishnan, C., et al. (2018). Mapping projections of molecularly defined dopamine neuron subtypes using intersectional genetic approaches. *Nature Neuroscience*, *21*(9), 1260–1271. http://dx.doi.org/10.1038/s41593-018-0203-4.

Ranade, S., Pi, H.-J., & Kepecs, A. (2014). Neuroscience: Waiting for serotonin. *Current Biology*, *24*(17), R803–R805. http://dx.doi.org/10.1016/j.cub.2014.07.024.

Rasmusson, D. D. (2000). The role of acetylcholine in cortical synaptic plasticity. *Behavioural Brain Research*, *115*(2), 205–218. http://dx.doi.org/10.1016/S0166-4328(00)00259-X.

Redgrave, P., Gurney, K., & Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Research Reviews*, *58*(2), 322–339. http://dx.doi.org/10.1016/j.brainresrev.2007.10.007.

Robinson, E. S. J., Dalley, J. W., Theobald, D. E. H., Glennon, J. C., Pezze, M. A., Murphy, E. R., et al. (2008). Opposing roles for 5-HT2a and 5-HT2C receptors in the nucleus accumbens on inhibitory response control in the 5-choice serial reaction time task. *Neuropsychopharmacology*, *33*(10), 2398–2406. http://dx.doi.org/10.1038/sj.npp.1301636.

Rosenbloom, M. H., Schmahmann, J. D., & Price, B. H. (2012). The functional neuroanatomy of decision-making. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *24*(3), 266–277. http://dx.doi.org/10.1176/appi.neuropsych.11060139.

Rushworth, M. F. S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*(4), 389–397. http://dx.doi.org/10.1038/nn2066.

Rushworth, M. F. S., Behrens, T. E. J., Rudebeck, P. H., & Walton, M. E. (2007). Contrasting roles for cingulate and orbitofrontal cortex in decisions and social behaviour. *Trends in Cognitive Sciences*, *11*(4), 168–176. http://dx.doi.org/10.1016/j.tics.2007.01.004.

Schall, J. D. (2001). Neural basis of deciding, choosing and acting, nature reviews. *Neuroscience*, *2*(1), 33–42. http://dx.doi.org/10.1038/35049054.

Schall, J. D., Palmeri, T. J., & Logan, G. D. (2017). Models of inhibitory control. *Philosophical Transactions of the Royal Society, Series B (Biological Sciences)*, *372*(1718), Article 20160193. http://dx.doi.org/10.1098/rstb.2016.0193.

Schmidhuber, J., Zhao, J., & Wiering, M. (1996). Simple principles of metalearning. SEE.

Schmidt, R., Leventhal, D. K., Mallet, N., Chen, F., & Berke, J. D. (2013). Canceling actions involves a race between basal ganglia pathways. *Nature Neuroscience*, *16*(8), 1118–1124. http://dx.doi.org/10.1038/nn.3456.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*(1), 1–27. http://dx.doi.org/10.1152/jn.1998.80.1.1.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of Neuroscience*, *13*(3), 900–913. http://dx.doi.org/10.1523/JNEUROSCI.13-03-00900.1993.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. http://dx.doi.org/10.1126/science.275.5306.1593.

Schweighofer, N., Bertin, M., Shishida, K., Okamoto, Y., Tanaka, S. C., Yamawaki, S., et al. (2008). Low-serotonin levels increase delayed reward discounting in humans. *Journal of Neuroscience*, *28*(17), 4528–4532. http://dx.doi.org/10.1523/JNEUROSCI.4982-07.2008.

Schweighofer, N., & Doya, K. (2003). Meta-learning in reinforcement learning. *Neural Networks: The Official Journal of the International Neural Network Society*, *16*(1), 5–9. http://dx.doi.org/10.1016/s0893-6080(02)00228-9.

Schweighofer, N., Tanaka, S. C., & Doya, K. (2007). Serotonin and the evaluation of future rewards: Theory, experiments, and possible neural mechanisms. *Annals of the New York Academy of Sciences*, *1104*(1), 289–300. http://dx.doi.org/10.1196/annals.1390.011.

Seo, M., Lee, E., & Averbeck, B. B. (2012). Action selection and action value in frontal-striatal circuits. *Neuron*, *74*(5), 947–960. http://dx.doi.org/10.1016/j.neuron.2012.03.037.

Sesack, S. R., & Pickel, V. M. (1992). Prefrontal cortical efferents in the rat synapse on unlabeled neuronal targets of catecholamine terminals in the nucleus accumbens septi and on dopamine neurons in the ventral tegmental area. *The Journal of Comparative Neurology*, *320*(2), 145–160. http://dx.doi.org/10.1002/cne.903200202.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the Rhesus monkey. *Journal of Neurophysiology*, *86*(4), 1916–1936. http://dx.doi.org/10.1152/jn.2001.86.4.1916.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., et al. (2014). Optimal behavioral hierarchy. *PLoS Computational Biology*, *10*(8), http://dx.doi.org/10.1371/journal.pcbi.1003779.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. http://dx.doi.org/10.1111/j.1467-7687.2007.00569.x.

Starkweather, C. K., & Uchida, N. (2021). Dopamine signals as temporal difference errors: Recent advances. *Current Opinion in Neurobiology*, *67*, 95–105. http://dx.doi.org/10.1016/j.conb.2020.08.014.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). The MIT Press..

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, *7*(8), 887–893. http://dx.doi.org/10.1038/nn1279.

Tanaka, S. C., Schweighofer, N., Asahi, S., Shishida, K., Okamoto, Y., Yamawaki, S., et al. (2007). Serotonin differentially regulates short- and long-term prediction of rewards in the ventral and dorsal striatum. *PLoS One*, *2*(12), http://dx.doi.org/10.1371/journal.pone.0001333.

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., et al. (2007). Schemas and memory consolidation. *Science*, *e1333*(5821), 76–82. http://dx.doi.org/10.1126/science.1135935.

Tsutsui, K.-I., Grabenhorst, F., Kobayashi, S., & Schultz, W. (2016). A dynamic code for economic object valuation in prefrontal cortex neurons. *Nature Communications*, *7*(1), 12554. http://dx.doi.org/10.1038/ncomms12554.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, *283*(5401), 549–554. http://dx.doi.org/10.1126/science.283.5401.549.

Verbruggen, F., Aron, A. R., Band, G. P., Beste, C., Bissett, P. G., Brockett, A. T., et al. (2019). A consensus guide to capturing the ability to inhibit actions and impulsive behaviors in the stop-signal task. *ELife*, *8*(e46323), http://dx.doi.org/10.7554/eLife.46323.

Wang, J. X. (2020). Meta-learning in natural and artificial intelligence. arXiv:2011.13464 [Cs]. http://arxiv.org/abs/2011.13464.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., et al. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. http://dx.doi.org/10.1038/s41593-018-0147-8.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., et al. (2017). Learning to reinforcement learn. arXiv:1611.05763 [Cs, Stat], http://arxiv.org/abs/1611.05763.

Wessel, J. R., & Aron, A. R. (2017). On the globality of motor suppression: Unexpected events and their influence on behavior and cognition. *Neuron*, *93*(2), 259–280. http://dx.doi.org/10.1016/j.neuron.2016.12.013.

Wickens, J. (1990). Striatal dopamine in motor activation and reward-mediated learning: Steps towards a unifying model. *Journal of Neural Transmission. General Section*, *80*(1), 9–31. http://dx.doi.org/10.1007/BF01245020.

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review*, *120*(2), 329–355. http://dx.doi.org/10.1037/a0031542.

Williams, B. R., Ponesse, J. S., Schachar, R. J., Logan, G. D., & Tannock, R. (1999). Development of inhibitory control across the life span. *Developmental Psychology*, *35*(1), 205–213. http://dx.doi.org/10.1037/0012-1649.35.1.205.

Winstanley, C. A., Theobald, D. E. H., Dalley, J. W., & Robbins, T. W. (2005). Interactions between serotonin and dopamine in the control of impulsive choice in rats: Therapeutic implications for impulse control disorders. *Neuropsychopharmacology*, *30*(4), 669–682. http://dx.doi.org/10.1038/sj.npp.1300610.

Wise, R. A., & Rompre, P. P. (1989). Brain dopamine and reward. *Annual Review of Psychology*, *40*(1), 191–225. http://dx.doi.org/10.1146/annurev.ps.40.020189.001203.

Xu, Z., van Hasselt, H., Hessel, M., Oh, J., Singh, S., & Silver, D. (2020). Meta-gradient reinforcement learning with an objective discovered. Online. arXiv:2007.08433 [Cs, Stat]. http://arxiv.org/abs/2007.08433.

Xu, Z., Hasselt, H. van., & Silver, D. (2018). Meta-gradient reinforcement learning. arXiv:1805.09801 [Cs, Stat], http://arxiv.org/abs/1805.09801.

Ye, Z., Altena, E., Nombela, C., Housden, C. R., Maxwell, H., Rittman, T., et al. (2014). Selective serotonin reuptake inhibition modulates response inhibition in parkinson's disease. *Brain*, *137*(4), 1145–1155. http://dx.doi.org/10.1093/brain/awu032.

Yu, A. J., & Dayan, P. (2002). Acetylcholine in cortical inference. *Neural Networks*, *15*(4–6), 719–730. http://dx.doi.org/10.1016/S0893-6080(02)00058-8.

Zhou, F.-M., Liang, Y., Salas, R., Zhang, L., De Biasi, M., & Dani, J. A. (2005). Corelease of dopamine and serotonin from striatal dopamine terminals. *Neuron*, *46*(1), 65–74. http://dx.doi.org/10.1016/j.neuron.2005.02.010.