Teaching Robots to Perceive Time: A Twofold Learning Approach

Inês Lourenço
Division of Decision and Control Systems
KTH Royal Institute of Technology
Stockholm, Sweden
ineslo@kth.se

Rodrigo Ventura
Institute for Systems and Robotics
Instituto Superior Técnico
Lisbon, Portugal
rodrigo.ventura@isr.tecnico.ulisboa.pt

Bo Wahlberg
Division of Decision and Control Systems
KTH Royal Institute of Technology
Stockholm, Sweden
bo@kth.se

***Abstract*—The concept of *time perception* is used to describe the phenomenological experience of time. There is strong evidence that dopaminergic neurons are involved in the timing mechanisms responsible for time perception. The phasic activity of these neurons resembles the behavior of the reward prediction error in temporal-difference learning models. Therefore, these models are used to replicate the neuronal behaviour of the dopamine system and corresponding timing mechanisms. However, time perception has also been shown to be shaped by time estimation mechanisms from external stimuli. In this paper we propose a framework that combines these two principles, in order to provide temporal cognition abilities to intelligent systems such as robots. A time estimator based on observed environmental stimuli is combined with a reinforcement learning approach, using a feature representation called *Microstimuli* to replicate dopaminergic behaviour. The elapsed time perceived by the robot is estimated by modeling sensor measurements as Gaussian processes to capture the second-order statistics of the natural environment. The proposed framework is evaluated on a simulated robot that performs a temporal discrimination task originally performed by mice. The ability of the robot to replicate the timing mechanisms of the mice is demonstrated by the fact that both exhibit the same ability to classify the duration of intervals.**

***Index Terms*—time perception, robotics, reinforcement learning, Gaussian processes, microstimuli**

I. Introduction

Although fully understanding the brain is a task still in its genesis, recent advances in robotics and machine learning research have enabled the reproduction of neuroscientific theories *in silico* [1]. In this work, we focus on the mechanisms responsible for *time perception*, which is the ability to perceive the passage of time and duration of events [2]. Several areas of the human brain are responsible for temporal cognition, such as the basal ganglia, the cerebellum, and the cerebral cortex [3], [4]. It is also influenced by several context parameters, such as the emotional state, the level of attention, and the task difficulty [5], [6]. Furthermore, the ability to perceive time based on context is not restricted to humans, having been found in groups of animals such as fish, birds, dogs and mice [7], [8]. Temporal perception is considered to be of utmost importance in collaborative activities and social interactions.

On the other hand, cyber-physical agents, such as robots, perform tasks based on state transitions that occur according to linear clock ticks, and lack the ability to perceive time in a context-dependent way [9]. The general problem we are interested in this paper is:

How can biologically-inspired mechanisms of time perception be replicated in a cyber-physical system?

This problem has been motivated in [10], where the authors stress the urge to incorporate temporal cognition as an intrinsic part of robots' decision-making process. In [11], it is stated that the equipment of artificial agents with human-like time perception capacities remains largely unexplored and is a prerequisite for bringing robotic cognition closer to human intelligence. It would give robots the ability to experience the flow of time, perceive synchronization and understand duration, thus improving their ability to make plans, recall experiences and communicate. This problem has a vast number of applications. For example, for the area of speech, having a perception of time would enable robots to learn to adapt their pauses in conversations to the situation and persons involved.

Our paper introduces a novel framework to enable artificial intelligent systems to estimate and use the passage of time in a biologically-inspired and, therefore, context-dependent way. This is achieved by emulating the time perception mechanisms of the brain algorithmically and using them in combination with the agent's decision-making algorithms. Many different models have attempted to explain how time is perceived in the brain. In this work, we consider that time stems from two different sources explained next: *external (environmental) stimuli* and *internal neuronal processes* [12].

The former, which we denote *External Timing* (ET), concerns how external stimuli influence the perception of time, such as why time intervals are overestimated when a movie is seen in twice the natural speed [13]. To replicate the timing mechanisms involved in this process, we model the data collected from the robot's sensors as Gaussian processes and estimate time intervals from it. This not only validates the idea that external stimuli influence the perception of time, but also the possibility of providing context-dependent timing mechanisms to non-biological agents.

The latter source of time perception, which we denote *Internal Timing* (IT), is related to how internal biological mechanisms might affect, or enable, the perception of time. It is believed that time is encoded in the spiking activity (firing rate) of neurons, namely dopaminergic neurons, where this activity changes when a task is performed at different speeds [3]. To replicate these timing mechanisms, we thus replicate dopaminergic behaviour principles in an agent.

By combining the two sources of timing mechanisms, ET and IT, we capacitate robots with their own time perception mechanism, that can be used to perform tasks that require temporal cognition. A biologically-inspired decision-making algorithm is obtained by combining them in a reinforcement learning framework. In summary, the main contributions of this paper are as follows:

- Estimation of the elapsed time from environmental sensor data collected by a robot, using a Gaussian process based estimator;
- Capacitate the agent with the ability to perform time-aware actions, using the time estimated from external stimuli and a reinforcement learning framework that replicates internal neuronal timing mechanisms;
- Validation of the proposed framework in numerical simulations, which qualitatively show similarities between the ability of the robot and that of mice to estimate the duration of intervals.

This paper is organized as follows. Section II presents the framework and formulates the problem. Section III describes the proposed method to estimate time from environmental data (ET), and Section IV presents the reinforcement learning decision-making framework (IT) based on the time estimate. The complete framework is validated and evaluated in Section V, and the main results are highlighted. Finally, in Section VI, conclusions are taken and future extensions are outlined.

II. Problem formulation

In this section, we define the notation and introduce the two components of the proposed framework: we first explain the reinforcement learning setting considered, and then introduce the component responsible for estimating time from the environment. Finally, we formalize the problems studied throughout the rest of the paper and describe the experimental setup.

A. Notation

All vectors are column vectors, unless transposed. At time step t, the ith element of vector v is $v_t(i)$. Matrices are denoted by capital letters, and $p(\cdot)$ denotes a probability density. We define the *elapsed time*, τ, to be the time difference between two events. It is also referred to as *interval duration*, and is used in *interval timing* tasks.

B. Preliminaries

In this paper we consider a discrete episodic reinforcement learning setting, modeled as a Markov Decision Process [14] and represented in Figure 1. The environment is described by a sequence of states $s_t \in \mathcal{S}$, where $\mathcal{S}$ is the state-space and t represents a time step. The agent can perform an action, $a_t \in \mathcal{A}$, in the environment at each time step, and, based on the *value* of that action, it receives a reward r_t. The *value* of an action is measured by its contribution to maximizing the expected sum of future rewards. Since we are interested in the value of the actions performed by the agent, we consider a reinforcement learning setup with action selection using Q-values. For large state-action spaces, features can be used to generalise the estimation of the value of a state-action pair to similar pairs. This technique is called *function approximation*. In this work, we compute the Q-values of state-action pairs, $Q(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ using a linear weighted combination of the features $x(s,a)$:

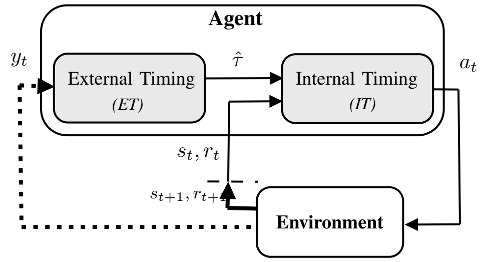

Fig. 1. Framework used to replicate the biological estimation of time as a combination of external environmental stimuli and internal neuronal mechanisms as follows. **ET)** The agent computes an estimate of the elapsed time $\hat{\tau}$ from environmental observations y_t. **IT)** This estimate is employed in a temporal-difference learning algorithm that replicates internal timing mechanisms: based on the elapsed time estimate $\hat{\tau}$ and the state s_t of the environment, the agent computes the Q-values (1) of each state-action pair, performs a corresponding action a_t (2), and receives a reward r_t.

$$Q_t(s,a) = w_t^T x_t(s,a) := \sum_{j=1}^{D} w_t(j) x_t(j). \tag{1}$$

The features are chosen as to replicate internal neuronal mechanisms (IT), and will be defined in Section IV. The goal of the agent is to learn the feature weights, $w_t \in \mathbb{R}^D$, that better reflect the importance of the features to the different state-action pairs. Based on the Q-values obtained from (1), the agent's actions are chosen following an ε-greedy policy:

$$a_t = \begin{cases} \arg\max\limits_a Q(s_t, a), & \text{with probability } 1 - \varepsilon_t, \\ \text{random action}, & \text{with probability } \varepsilon_t. \end{cases} \tag{2}$$

Here, ε_t is an exploration parameter that decays according to $\varepsilon_t = \rho \varepsilon_{t-1}$, with decay parameter $\rho \in [0,1]$.

To this standard reinforcement learning approach we add an external timing (ET) component, as represented in Figure 1. In this component the agent computes its own estimate of the elapsed time, $\hat{\tau}$, from data $\mathcal{O} = \{y_t(i)\}_{i=1}^M$, that it collects from the environment at each time instant t. The data is collected by M sensors and, from each one, N observations are uniformly taken at the time instances $t_1, \dots, t_N$, throughout the interval $[0, \tau]$. The sensor data is represented as the N-dimensional vector $y_t(i) = [y_{t_1}(i), \dots, y_{t_N}(i)]^T$, and $\tau \in \mathbb{R}^+$ is the *elapsed time*. The estimated elapsed time is then used as a parameter of the RL algorithm.

C. Problem formulation and experimental setup

To replicate biologically-inspired time perception mechanisms in a robot, we create a twofold approach that combines internal and external timing mechanisms.

A temporal discrimination episodic task is used to evaluate the proposed framework. It consists of an agent navigating around the environment and collecting data, $\mathcal{O}$, from its sensors. During each episode, the agent receives two stimuli separated by a certain time interval. Its task is to estimate the duration of the interval. We formulate the ET problem of estimating elapsed time from sensor data as:

Problem 1 (Timing from external stimuli)**.** *A set of observations, $\mathcal{O}$, is collected during a certain interval of length τ. Given these observations, estimate the interval duration; that is, estimate τ given $\mathcal{O}$.*

A solution to Problem 1 is presented in Section III.

The next step consists of reproducing internal neuronal timing mechanisms that are influenced by the obtained estimate $\hat{\tau}$ of the time perceived from external stimuli. This originates the complete decision-making framework for the agent to perform actions based on both external and internal timing (ET and IT) mechanisms, illustrated in Figure 1. How the estimate of the interval length can be used to perform interval timing tasks leads us to the second problem:

Problem 2 (Timing from internal and external mechanisms)**.** *How can internal neuronal mechanisms be replicated from the behaviour of the dopamine system and combined with the time estimate $\hat{\tau}$ obtained from the environment?*

Problem 2 is addressed in Section IV.

By solving these two problems, we obtain a framework that allows an agent to correctly perform an interval timing task based on the two timing mechanisms present in the brain. More specifically, we introduce the time perceived from external stimuli in the agent's decision-making process by using the perceived interval length, $\hat{\tau}$, to condition the features x_t used to compute the Q-values in (1). Since these features are also chosen to reproduce the behaviour of the dopaminergic system that is believed to regulate timing mechanisms, we establish a framework for performing actions, a_t, that considers both internal and external timing mechanisms.

III. ET: Timing from external stimuli

In this section, we present the method used in the ET component to solve Problem 1 by computing the elapsed time from environmental data. We start by presenting background information and then our proposed solution.

A. Related work

In [12], the following Bayesian framework is used to estimate the elapsed time τ given the data $\mathcal{O}$:

$$p(\tau|\mathcal{O}) \propto p(\mathcal{O}|\tau)p(\tau). \tag{3}$$

The peak of the posterior distribution $p(\tau|\mathcal{O})$ is the Maximum *a Posteriori* (MAP) estimate and is considered to be the estimate of elapsed time, due to being the most likely time interval to have passed given the data. The observations $\mathcal{O}$, collected from the environment, provide therefore a sensor-based estimate of the passage of time [15], [16]. They can be obtained from multiple sources of sensor data, such as images from a camera or time-series from a LIDAR.

The question is then *how can this data be modelled?* One should take into account that environmental information does not change completely randomly. It shows patterns of high correlation in both space and time [17]. Further, to avoid handling excessive amounts of data, the observations are usually treated in a low-dimensional representation. As an example, in [12] it was studied how external stimuli introduce a bias on the perceived time, and considered the estimate of the elapsed time as a probabilistic expectation of stimulus change in the environment, that can be inferred from its second-order statistical properties. These are characterized by the mean μ and correlation between observations (such as points in a natural time-varying image). The latter is represented by the kernel K and expresses how much the process changes from one time step to the next, corresponding to the rhythm of change of the natural environment. It was also shown that the power spectrum of the observations can be approximated by that of the Ornstein-Uhlenbeck (OU) function, which is a process of Brownian motion with friction [18]. If the statistical properties remain constant in time, the process is stationary and thus the observations are modelled as stationary Gaussian processes with an OU kernel [19].

B. Replicating External Timing

In the ET component from Figure 1 we provide a sense of the passage of time from environmental information to a cyber-physical agent by adopting the method of [12] to approach Problem 1: given observations $\mathcal{O}$, compute the MAP estimate of (3) to find the perceived elapsed time $\hat{\tau}$.

Assumption 1 (Uniform prior)**.** *The prior distribution $p(\tau)$ is considered to be uniformly distributed since we do not model the brain's a priori information about the elapsed time.*[1]

Under uniform prior, the MAP in (3) coincides with the maximum likelihood estimate [20], which means that the estimate of the elapsed time is the maximum of $p(\mathcal{O}|\tau)$. This corresponds to the probability of having observed $\mathcal{O} = \{y_t(i)\}_{i=1}^{M}$ during the interval τ, and is modelled as a zero-mean joint Gaussian distribution over the N observations of all M independent sensors. Considering that $t_1 = 0$ and $t_N = \tau$, this refers to $y_0(i), \ldots, y_\tau(i)$, for $i = 1, \ldots, M$:

$$p(y_t(i)|\tau) = \mathcal{N}(y_t(i); 0, K_\theta) = \frac{e^{-\frac{1}{2}y_t(i)K_\theta^{-1}y_t^T(i)}}{\sqrt{\det(2\pi K_\theta)}}. \tag{4}$$

This joint distribution has an unknown kernel function K_θ parametrized by θ.

To solve Problem 1 we first extend the work by [12] with a parameter estimation step to obtain the hyperparameters of the model from data: *use Bayesian model selection to find the hyperparameters of the model* (4)*, θ.* For this, we maximize

[1] Instead, in Section IV, we focus on directly replicating the neural mechanisms that would generate this information.

the logarithm of the likelihood in (4) with respect to θ, which involves computing the respective derivatives:

$$\frac{\partial}{\partial\theta_j}\log p(y_t(i)|\theta) = -\frac{1}{2}tr(\phi\phi^T - K_\theta^{-1})\frac{\partial K_\theta}{\partial\theta_j}, \qquad (5)$$

where $\phi = K^{-1}y_t(i)$.

Assumption 2 (Model selection)**.** *Assume that, as justified in Section III-A, the data comes from a Gaussian process with Ornstein-Uhlenbeck covariance.*

In this case, each element of the $N \times N$ OU kernel $K_\theta(\tau)$, where τ is the difference between time indexes, is given by:

$$K_{\lambda,\sigma}(\tau) = e^{-\lambda|\tau|} + \sigma^2\psi(\tau). \qquad (6)$$

Here, $\psi(0) = 1$ and $\psi(\tau) = 0$ for $\tau \neq 0$, and $\theta = [\lambda, \sigma]$ are the hyperparameters of the model. Hence, Problem 1 is solved by identifying the appropriate values for λ and σ, in such a way that the properties of the Gaussian process approximate the ones of the data, and with these maximize (4). The MAP is the robot's estimate of the elapsed time between the two stimuli, $\hat{\tau}$.

IV. IT: Timing from internal neural mechanisms

This section presents the algorithm used in the IT step from Figure 1 to answer Problem 2. It applies results from neuroscience to reinforement learning, resulting in a biologically-inspired Temporal-Difference *(TD) learning* algorithm. As in Section III, we start by presenting background information and then the proposed method.

A. Related work

For long, dopamine neurons have been know to be involved in action selection and reward prediction mechanisms [21]. Besides these, they are also believed to be responsible for interval timing mechanisms, that is, to have the ability to encode the passage of time [22]. This is due to their phasic activity encoding a reward prediction error signal, which is a biologically-inspired signal that reflects the difference between the received and the expected reward on the basis of previous learning. This signal is present in TD learning models, and is referred to as *TD error*. These models were introduced by [23] to account for the need of a learning system that tries to predict the value of future events from the patterns of stimuli and rewards. In TD learning algorithms with *function approximation*, each state s_t, introduced in Section II-B, is described by a set of D features, $x_t(1), \ldots, x_t(D)$, that encode cognitive and sensory experiences (e.g., environmental stimuli) of an animal at each time step t. The two most accepted theories nowadays to represent stimuli, given the need of consistency with what happens in the basal ganglia, are the *Complete Serial Compound* [23], [24] and the *Microstimuli* [25]. Since the former presents inconsistencies in representing certain characteristics of the dopamine system, the latter is, to the best of our knowledge, the most realistic one and is therefore used to solve Problem 2 in the next section.

B. Replicating Internal Timing using TD learning

We now present the *TD learning* algorithm considered. Recall that, at each time step, the agent performs the action with the highest Q-value (see (2)), and that the Q-values are computed as $Q(s_t, a_t) = \sum_{j=1}^{D} w_t(j)x_t(j)$ (see (1)), by multiplying the weights w_t by features x_t. Eligibility traces, e_t, are an essential attribute of reward learning that, when multiplied by the TD error δ_t, expand the influence of the presence of a state through time [26]. This error is the difference between the expected and received reward, r_t, at each time step. The update equations of these parameters are:

$$\delta_t = r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t), \qquad (7)$$

$$w_{t+1}(j) = w_t(j) + \alpha\delta_t e_t(j), \qquad (8)$$

$$e_{t+1}(j) = \gamma\eta e_t(j) + x_t(j), \qquad (9)$$

where γ is the discount rate, α the learning rate, and η the decay parameter that determines the plasticity window of recent stimuli.

To replicate the behaviour of dopaminergic neurons, we use the *Microstimuli* framework presented in Figure 2 of [25] to represent the features x_t. In the *Microstimuli* framework, both cues and rewards deploy their own set of m microstimuli. If there are ζ cues and rewards per episode, there are a total of $m\zeta = D$ microstimuli. Each feature $x_t(1), \ldots, x_t(D)$ represents the level of each microstimulus at time t. This level is computed from the product of the exponentially decaying trace height h_t,

$$h_t = \exp\{-(1-\xi)t\}, \qquad (10)$$

with decay parameter ξ, by Gaussian basis functions f with center ν and width β,

$$f(h, \nu, \beta) = \frac{1}{\sqrt{2\pi}}\exp\Big\{-\frac{(h-\nu)^2}{2\beta^2}\Big\}, \qquad (11)$$

according to:

$$x_t(j) = h_t f\left(h_t, \frac{j}{m}, \beta\right), \quad \text{for } j = 1, \ldots, D. \qquad (12)$$

Knowing how much a microstimulus has decayed due to its slowly decaying memory trace can be seen as a basis for the elapsed time. The weights $w_t(1), ..., w_t(D)$ from (8) that are multiplied by the Microstimuli features from (12), represent the strengths of the corticostriatal synapses and indicate how important each one is for each state and action.

C. Time perception as a combination of internal and external timing mechanisms

We design a robot that, following the TD learning framework from Section IV-B, can use the environmental estimate of elapsed time from Section III-B to correctly perform a sequence of actions. Correctness entails similarity to the actions performed by an agent with temporal cognition.

Following the setup from Section II-C, a set of microstimuli features x_t are deployed when the agent receives the first stimulus, and another set is deployed $\hat{\tau}$ time steps after. This

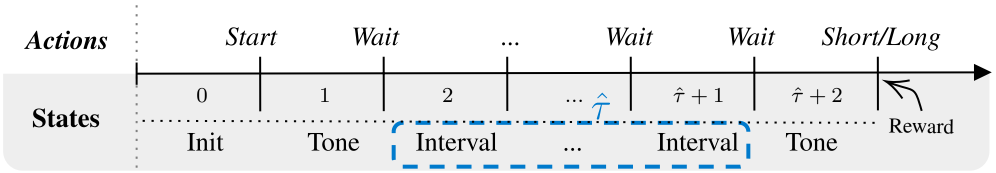

Fig. 2. Desired state transition, obtained when the optimal action (on the top row) is chosen. After pressing the *Start* button, the state of the environment changes to *Tone* and the number of *Interval* states between the next *Tone* state is uniformly sampled from the maximum interval length, which is a design variable for each experiment. The agent estimates the number of time steps spent in the *Interval* state, $\hat{\tau}$, and, after the second *Tone* state, chooses the action *Short* or *Long* that corresponds to its estimate. If the correct action is chosen, a positive reward is given to the agent.

Algorithm 1 Temporal discrimination task

```
Initialize $Q(s_0, a_0) = 0$, for all $s \in \mathcal{S}$, $a \in \mathcal{A}$, and $w(1), \ldots, w(D)$ randomly (e.g. $w(j) \in [0,1]$)
for each episode do
    Initialize $s_0$
    for each time step $t$ do
        if first stimulus == 1 then
            Update $x_t(1), \ldots, x_t(D)$ according to (12)
        else if has received stimulus 1 but not 2 then
            Collect data $y_t(1), \ldots, y_t(M)$
        else if second stimulus == 2 then
            Estimate the elapsed time, $\hat{\tau}$, by maximizing (4)
            Update $x_{\hat{\tau}}(1), \ldots, x_{\hat{\tau}}(D)$, according to (12)
        end if
        Compute the Q-values according to (1) and choose $a_t$ according to (2). Take action $a_t$, observe $r_t$, $s_{t+1}$
        $\delta_t \leftarrow r_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)$
        $w_{t+1}(j) \leftarrow w_t(j) + \alpha\delta_t e_t(j)$,    for $j = 1, \ldots, D$
        $e_{t+1}(j) \leftarrow \gamma\eta e_t(j) + x_t(j)$,    for $j = 1, \ldots, D$
        $s_t \leftarrow s_{t+1}$
    end for
    Until $s_t$ is terminal
end for
```

means that the time step at which the second set of stimuli is deployed depends on the agent's estimate of elapsed time, $\hat{\tau}$, computed in Section III. The features influence the Q-value of the state-action pair according to (1), and therefore the action a_t chosen by the agent. The complete framework from Figure 1 is summarized in Algorithm 1. Hence, by solving Problems 1 and 2 we replicate internal and external timing mechanisms. In the next section, the framework is evaluated by analysing the sequence of actions chosen by the agent in a temporal discrimination task. If the actions are similar to those of an animal, we claim that the agent attained temporal cognition.

V. Numerical Results

In this section, we evaluate the proposed framework described in Section IV-C in an interval timing task, and compare the behavior of an agent using Algorithm 1 with that of mice.

A. Background

In [3], a temporal discrimination episodic experiment was performed with mice. There are three available buttons: "Start", "Short" or "Long". When the mouse presses the former, two auditory tones are presented. These are separated by a time interval that varies between episodes and that can be classified as either short or long. Based on how much time the animal estimates to have passed between both cues, the animal presses the button corresponding to its estimated interval ("Short" or "Long"). If the action is correct, the animal is rewarded with water or food.

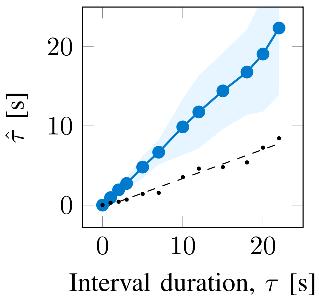

Fig. 3. Estimated $\hat{\tau}$ for each interval duration. The mean of the estimated interval almost perfectly matches the ground truth (blue dots), while the standard deviation (faded blue) increases linearly with the interval length. Its linear regression is shown in black. This illustrates the scalar property, a trend also exhibited by humans and animals.

We replicate this experiment in a simulated robot. The state of the environment is given by $\mathcal{S}$ ={Init, Tone, Interval}, and the action space of the agent by $\mathcal{A}$ ={Start, Wait, Short, Long}. The schematic representation of an episode where the optimal sequence of actions is performed is shown in Figure 2. We define the *interval duration* of an episode, τ, to be its number of "Interval" states. This number is a realization of a discrete uniform random variable $\mathcal{I} \sim \text{unif}\{1, L\}$, where $L \in \mathbb{N}$ is the maximum interval duration of the experiment. The interval duration is classified as

$$\begin{cases} \text{“Short”}, & \text{if } \tau \in [1, 2, \ldots, \lfloor L/2 \rfloor], \\ \text{“Long”}, & \text{if } \tau \in [\lceil L/2+1 \rceil, \ldots, L], \end{cases}$$

if τ is even, and $\{L/2, L/2+1\}$ is the classification boundary. In the results presented next, we chose the maximum duration to be $L = 3$ sec as in the real experiment, which corresponds to $L = 8$ time steps. So "Short" $:= \tau \in [1, 2, 3, 4]\}$ and "Long" $:= \tau \in [5, 6, 7, 8]\}$. The problem has increased complexity since the agent cannot perform the RL problem by counting the number of 'Interval" states. It has to use the $\hat{\tau}$ obtained from the data when solving Problem 1 from Section III.

B. Results and discussion

In this section we present the main numerical results obtained by our proposed solution to Problems 1 and 2.

Numerical result 1 (ET)**.** *The elapsed time is accurately estimated from sensor data.*

The observations $y_t(i)$ are considered to be the values of the ith angle of the simulated robot's LIDAR at time t, collected while the robot does the "Wait" action between tones. From (5) we estimated the maximum likelihood model parameters λ and θ of the collected data and used these to estimate the elapsed time τ. Figure 3 shows the estimated $\hat{\tau}$ for different intervals, from which it can be concluded that our estimate is accurate for this range of intervals. This shows that Problem 1 is therefore correctly solved. It can be concluded that the average estimated duration is not affected by the length of the interval to be estimated, but its standard deviation increases approximately linearly with the interval length. This is called the scalar property [27], and is one of the most important properties of time perception. It represents consistency with what happens in the brain, since the longer the interval, the harder it is for humans and animals to estimate it [28].

Numerical result 2 (ET and IT)**.** *The actions of the agent in the temporal discrimination task are similar to those of the mice, showing a similar ability to classify interval durations.*

We start by presenting two figures that provide insights about the framework. Firstly, Figure 4 shows three completed episodes with the same interval duration sampled at different phases of the training. Since we chose episodes where $\tau = 2$, the second tone happens at $t = 4$ and the reward is given after. It can be seen that as learning occurs, the TD error from (7): *i)* increases at cue onset, *ii)* decreases at reward delivery. This indicates that the model learns to predict the reward from earlier stimuli, as explained in [21]. These results match empirical data, and the second tone acts as a conditional stimulus from *classical conditioning*.

Figure 5 shows the Q-values computed from (1), after training, for two episodes with different interval durations. The action with the highest Q-value at each time step is marked with a dot and it is the one chosen if there is no exploration. In both cases the resulting sequence of actions corresponds to the optimal one presented in Figure 2, which means that the agent learns how to correctly act based on its time estimate.

Results like the ones from Figures 4 and 5 can be a premonition of the success of the framework for social interactions. Rather than waiting a predetermined interval between receiving a question and answering it, it can enable agents to adjust this interval to the situation and people involved.

The next results are presented using seconds instead of time steps to quantify the interval duration, for simplicity of comparison with the original experiment. After the exploration phase, the agent learns to perform the correct sequence of actions until hearing the second tone. However, to receive the reward it also needs to choose the button that correctly classifies the interval duration of that episode. Figure 6 shows that the number of misclassified interval durations is higher for those closer to the boundary between "Short" and "Long" (around $\hat{\tau} = \{1.5, 1.8\}$s – note that this is not the case of the episodes in Figure 5). This happens regardless of the maximum interval length, and is a trend also exhibited by humans and animals [4]. Figure 7 shows the corresponding psychometric curve, representing the empirical probability of the intervals being classified as "Long". The orange curve is a logistic function fit to the average performance of a mouse during 10 experiments, from [3], and shows a qualitatively similar behaviour to the one of our agent, in blue.

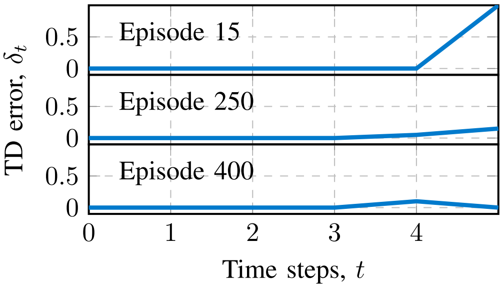

Fig. 4. Analysis of the TD error in different phases of learning. In the three episodes chosen, $\tau = 2$ time steps. As learning occurs, the agent starts expecting a reward after its correct classification of the interval ($t = 4$). Therefore, the TD error at the end of the episode decreases.

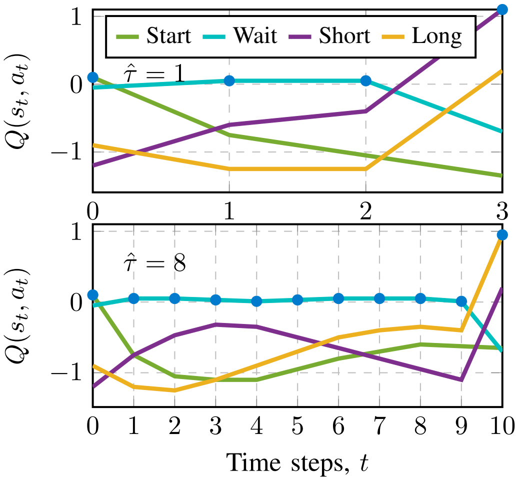

Fig. 5. Evolution of the Q-values with the interval duration, after training. The time step values correspond to the state numbers from Figure 2, and each line to the Q-value of each action from the same figure. In the top figure, $\tau = 1$ (short interval), and in the bottom one, $\tau = 8$ time steps (long interval).

In summary, the similarity between the timing mechanisms of the robot and mice is demonstrated by the following results:

- The elapsed time was successfully computed from environmental data;
- Uncertainty in the estimation of the interval duration from data increases with the length of the interval, exhibiting the scalar property;
- The error of the TD model replicates the firing rate of dopamine neurons, decreasing with reward expectancy;

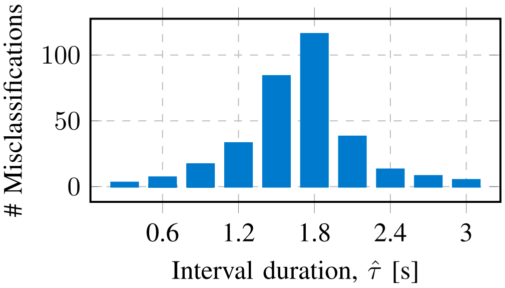

Fig. 6. Intervals misclassified based on their corresponding duration. The total number of misclassifications is 327, being the average 1.65 s and the median 1.8 s. As is the case in humans and animals, the intervals in the boundary between classes are the ones more commonly wrongly classified.

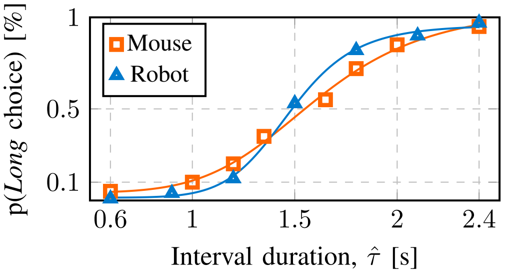

Fig. 7. Psychometric curves corresponding to the empirical probability of an interval being classified as *Long*. The psychometric curve of the RL agent closely matches that of the mouse.

- Uncertainty in the classification of intervals is higher on the boundary between classes;
- The psychometric curve of our agent closely matches the one of mice.

VI. Conclusion

In this paper, we have proposed an algorithm for providing temporal cognition to a robot. This algorithm combines two timing sources: estimation mechanisms from external stimuli, and internal neuronal mechanisms. To replicate the former, we exploited results from Gaussian processes. For the latter, we employed a temporal difference learning feature representation called *Microstimuli* to replicate dopaminergic behaviour. In numerical simulations, we showed that an agent using the proposed algorithm is able to succeed in time-dependent tasks, due to the ability to perceive the passage of time similar to the one that humans and animals have.

In the future, the framework proposed shall be implemented in a real robot. One direction for future work is to use deep learning to analyse the non-parametric distribution of the environmental data instead of parametrized Gaussian processes. Another direction involves studying how the framework behaves for other time scales and how it can be adapted, since dopaminergic neurons are believed to only control temporal judgments on a time scale of seconds.

References

[1] X. Fan and H. Markram, “A brief history of simulation neuroscience,” *Frontiers in neuroinformatics*, vol. 13, p. 32, 2019.

[2] L. G. Allan, “The perception of time,” *Perception & Psychophysics*, vol. 26, no. 5, pp. 340–354, 1979.

[3] S. Soares, B. Atallah, and J. Paton, “Midbrain dopamine neurons control judgment of time,” *Science*, vol. 354, no. 6317, pp. 1273–1277, 2016.

[4] T. Gouvêa, T. Monteiro, A. Motiwala, S. Soares, C. Machens, and J. Paton, “Striatal dynamics explain duration judgments,” *Elife*, vol. 4, 2015.

[5] S. Droit-Volet and S. Gil, “The Time–Emotion paradox,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1525, pp. 1943–1953, 2009.

[6] S. Droit-Volet and H. Meck, “How emotions colour our perception of time,” *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 504–513, 2007.

[7] M. R. Drew, B. Zupan, A. Cooke, P. Couvillon, and P. D. Balsam, “Temporal control of conditioned responding in goldfish.,” *Journal of Experimental Psychology: Animal Behavior Processes*, vol. 31, no. 1, p. 31, 2005.

[8] K. Healy, L. McNally, G. D. Ruxton, N. Cooper, and A. L. Jackson, “Metabolic rate and body size are linked with perception of temporal information,” *Animal Behaviour*, vol. 86, no. 4, pp. 685–696, 2013.

[9] M. Maniadakis, P. Trahanias, and J. Tani, “Explorations on artificial time perception,” *Neural Networks*, vol. 22, no. 5-6, pp. 509–517, 2009.

[10] M. Maniadakis and P. Trahanias, “Temporal cognition: A key ingredient of intelligent systems,” *Frontiers in Neurorobotics*, vol. 5, p. 2, 2011.

[11] M. Maniadakis and P. Trahanias, “Time in consciousness, memory and human-robot interaction,” in *International Conference on Simulation of Adaptive Behavior*, pp. 11–20, Springer, 2014.

[12] M. B. Ahrens and M. Sahani, “Observers exploit stochastic models of sensory change to help judge the passage of time,” *Current Biology*, vol. 21, no. 3, pp. 200–206, 2011.

[13] S. W. Brown, “Time, change, and motion: The effects of stimulus movement on temporal perception,” *Perception & psychophysics*, vol. 57, pp. 105–116, 1995.

[14] V. Krishnamurthy, *Partially Observed Markov Decision Processes*. Cambridge University Press, 2016.

[15] D. Eagleman, “Time perception is distorted during slow motion sequences in movies,” *Journal of Vision*, vol. 4, no. 8, pp. 491–491, 2004.

[16] S. W. Brown, “Time, change, and motion: The effects of stimulus movement on temporal perception,” *Perception & Psychophysics*, vol. 57, pp. 105–116, 1995.

[17] D. W. Dong and J. J. Atick, “Statistics of natural time-varying images,” *Network: Comp. in Neural Systems*, vol. 6, no. 3, pp. 345–358, 1995.

[18] G. E. Uhlenbeck and L. S. Ornstein, “On the theory of the Brownian motion,” *Physical Review*, vol. 36, no. 5, p. 823, 1930.

[19] C. B. Do, “Gaussian processes,” *Stanford University, Stanford, CA*, vol. 5, p. 2017, 2007.

[20] L. Ljung, *System Identification: Theory for the User*. Prentrice-Hall, New Jersey, 1987.

[21] P. W. Glimcher, “Understanding dopamine and reinforcement learning: The dopamine reward-prediction error hypothesis,” *Proc. of the National Academy of Sci.*, vol. 108, no. Supplement 3, pp. 15647–15654, 2011.

[22] S. J. Gershman, A. A. Moustafa, and E. A. Ludvig, “Time representation in reinforcement learning models of the basal ganglia,” *Frontiers in Computational Neuroscience*, vol. 7, p. 194, 2014.

[23] R. S. Sutton and A. G. Barto, “Time-derivative models of Pavlovian reinforcement,” *Learning and Computational Neuroscience: Foundations of adaptive networks*, pp. 497–537, 1990.

[24] P. R. Montague, P. Dayan, and T. J. Sejnowski, “A framework for mesencephalic dopamine systems based on predictive Hebbian learning,” *Journal of neuroscience*, vol. 16, no. 5, pp. 1936–1947, 1996.

[25] E. A. Ludvig, R. S. Sutton, and E. J. Kehoe, “Stimulus representation and the timing of reward-prediction errors in models of the dopamine system,” *Neural computation*, vol. 20, no. 12, pp. 3034–3054, 2008.

[26] S. P. Singh and R. S. Sutton, “Reinforcement learning with replacing eligibility traces,” *Machine learning*, vol. 22, no. 1-3, pp. 123–158, 1996.

[27] H. Lejeune and J. Wearden, “Scalar properties in animal timing: Conformity and violations,” *Quarterly Journal of Experimental Psychology*, vol. 59, no. 11, pp. 1875–1908, 2006.

[28] M. Sucala, B. Scheckner, and D. David, “Psychological time: interval length judgments and subjective passage of time judgments,” *Current psychology letters. Behaviour, brain & cognition*, vol. 26, no. 2, 2011.