The Embodied Nature of Well-Timed Behavior

Mostafa Safaie[1,2*], Maria-Teresa Jurado-Parras[1,2*], Stefania Sarno[1,2,3,4], Jordane Louis[1,2], Corane Karoutchi[1,2], Ludovic F. Petit[1,2], Matthieu O. Pasquet[1,2], Christophe Eloy[3,4], David Robbe[1,2,4†]

[1] Institut de Neurobiologie de la Méditerranée (INMED), Marseille, France.
[2] Aix Marseille Université, Marseille, France.
[3] Aix Marseille Univ, CNRS, Centrale Marseille, IRPHE, Marseille, France.
[4] Turing Centre for Living Systems (CENTURI), Marseille, France.

[*] Co-first authors.

[†] Corresponding author, email: david.robbe@inserm.fr

Abstract

It has been proposed that animals rely on internal representation of time to measure supra-second long intervals and adapt their behavior accordingly. If this was the case, accurate timing should be largely independent of variations in external factors (e.g., changes in the environment). Here, we used a task in which rats, freely moving on a motorized treadmill, could obtain a reward if they approached it after a fixed interval. By manipulating several task parameters, such as the speed of the treadmill, we found that the animals accurately timed their approaches only if they could perform a stereotyped motor sequence inside the confined space of the treadmill. Reinforcement learning-based models further suggested that the strategy of the animals was incongruent with that of agents using an internal representation of time. We conclude that the emergence of well-timed behaviors critically depends on the ability of animals to develop motor routines adapted to their environment.

Introduction

For humans and other animals, an adaptive behavior is defined as selecting and executing, in a given context, the action that maximizes immediate or future rewards while minimizing energetic expenditure. In some situations, the appearance of a sensory cue will not only determine which action should be chosen but also how long after the cue appeared this action must be initiated [1,2]. For instance, athletes performing sprint races learn by experience that the go command signalling them to start running will be given no later than two seconds after the set command. The athletes' accurate estimation of the average time interval between set and go commands is demonstrated by their tendency to anticipate the go command, which in some cases results in false starts. More generally, the ability of animals to adapt their behavior to temporal regularities in the environment is critical for survival, as the appearance of a sensory cue can predict the timing of food availability, predator attack or mating opportunity [3,4].

Understanding how animals adapt their behavior to time intervals lasting for a few seconds is challenging because, unlike sensory modalities (vision, olfaction, audition, etc.), time is not a material entity and animals are not equipped with a sense organ for time perception. One influential idea in the field of systems neuroscience, commonly referred to as the internal clock hypothesis, states that complex nervous systems can encode time in their spiking activity and that animals use such representation to estimate if the duration of a given time interval is similar to (or different from) a previously experienced interval [5–9]. Irrespective of its exact neural implementation, the internal clock works according to the following principles. Once a cue appears to signal the beginning of a time interval (e.g., the "set" command in sprint races), neuronal time quanta are accumulated at a specific rate. Time interval estimation consists in comparing the magnitude of this ongoing accumulation process with a stored value determined through experience (e.g., multiple exposures to time intervals between "set" and "go" commands). Errors in counting neuronal time quanta will accumulate with time. Consequently, such a mechanism predicts that the variance of time estimation report should increase with the duration of the time interval, which has been verified in humans and animals [10–17].

The internal clock hypothesis is appealing due to its simplicity and its compliance with the aforementioned scalar property of variance. But considering the variety of neuronal correlates of passage of time or time intervals (e.g., duration-selective neurons, ramping firing rate, neuronal population, network state) expressed by individual neurons or neuronal ensembles of various sizes recorded in multiple brain regions [18], it is unlikely that a single internal clock can be isolated from the multitude of temporally-ordered neuronal activities in the brain. How animals could know which internal clock they should rely on remains unclear. This difficulty, combined with the fact that time representations are: 1) relevant to various behavioral conditions engaging several brain functions (sensory detection, memory, motor control, attention, response to threats) and 2) intrinsic to the activity of ensembles of neurons in the brain (i.e., neuronal activity covaries with time), has led to the proposal that time estimation ultimately relies on emergent properties of interacting neuronal networks rather than a pure time-dedicated internal clock [18,19].

Understanding the perception of time is also complicated by the common observation that it is not isomorphic with the time measured by human-made clocks and strongly depends on contextual factors affecting the subjects' attentional, motivational or emotional state [20–24]. It is believed that this volatility of time perception is primarily caused by alterations in its underlying

neural mechanisms (e.g., increase or decrease in the speed of the dedicated or emergent population clocks) [6, 18]. Alternatively, the strong influence of contextual factors on time perception may indicate that mechanisms beyond internal neural activity play a primary role in how humans and animals perceive time and adapt their behavior accordingly. The embodiment framework proposes that important aspects of cognition are determined by physical features of the body, which in turn influence how humans or animals interact with their environment [25]. This framework seems particularly relevant for time perception as humans naturally tend to estimate time intervals through rhythmical movements of body parts such as flicking fingers, moving arms, nodding heads or tapping feet [26]. Although humans can estimate time by counting in their heads, it is difficult to separate this mental faculty from the repeated experiences of counting the passing seconds aloud in everyday life, which is primarily a sequence of coordinated movements across respiratory, laryngeal and supraglottal articulatory systems. It is also noteworthy that humans have created devices that indicate time by moving objects in space (clocks) and extensively use metaphors containing movement and space references when speaking of time ("holidays are approaching", "time flies") [27,28]. Finally, it has been recently argued that explicit perception of time may be constructed implicitly through association between the duration of an interval and its sensorimotor content [29]. Thus, animals (and humans) may primarily rely on the movement of their body in space to estimate time [30], which incidentally explains why brain regions involved in movement control, such as the basal ganglia, have consistently been associated with time perception [29].

In contrast to the wealth of neurophysiological evidence supporting the internal clock hypotheses (dedicated and emergent), studies that have directly examined the contribution of embodied strategies to well-timed behaviors are scarce. In tasks in which fixed time intervals must be respected to obtain a reward, rats do not stay immobile but take advantage of the structure of the environment to develop stereotyped motor routines whose durations match these task-specific temporal constraints [31–33]. Crucially, whether animals can accurately respect a temporal constraint without relying on such motor routines is unknown.

Here, we challenged rats in a task taking place on a 90 cm-long motorized treadmill. Animals had to respect a 7 s long time interval before approaching a reward port located at one extremity of the treadmill. We observed that rats took advantage of their ability to move in the confined space of the treadmill to develop a simple motor routine (a spatial trajectory within the treadmill boundaries) whose execution duration matched the 7 s long time constraint. By manipulating the time interval and the speed of the treadmill (its magnitude and reliability across trials) or interfering with the initiation of the motor routine, we created conditions that either facilitated or prevented the development or usage of this simple motor routine. The results suggest that rats, to respect time intervals, naturally try to develop motor routines whose completion leads to performing the rewarding action (approaching the reward port) at the right time. Reinforcement learning-based models also suggested that the animals' behavioral strategy we observed in our task is incongruent with that of an agent using an internal representation of time.

Results

To investigate whether animals rely preferentially on an inner or an embodied sense of time, we challenged Long-Evans rats in a time estimation task taking place on a motorized treadmill. In brief,

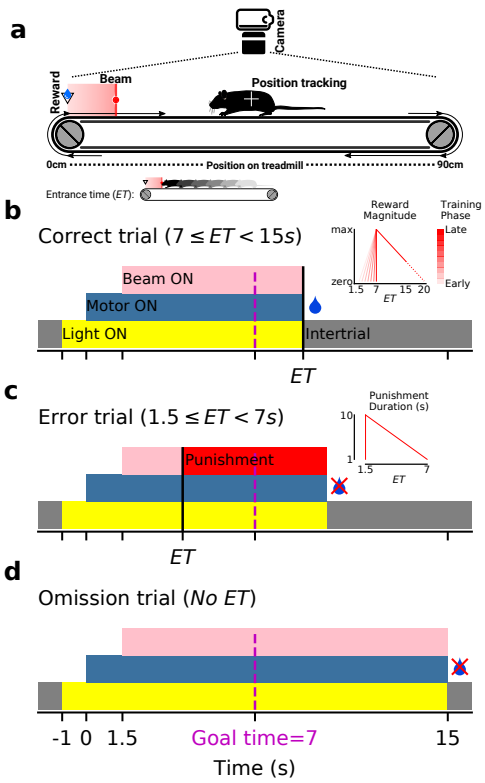


Figure 1: Treadmill task and trial types. a) Rats were enclosed on a motorized treadmill. The infrared beam placed at 10 cm of the reward port marked the beginning of the reward area (pink shaded area). During each trial, the belt pushed the animals away from the reward area and the first infrared beam interruption defined the reward area entrance time (*ET*). During trials and intertrials, the animals' position was tracked via a ceiling-mounted video camera. **b)** Schematic description of a rewarded correct trial. *Inset*: the magnitude of the delivered reward dropped linearly as *ET* increased (maximum reward at goal time, GT = 7 s). In early stages of training, smaller rewards were delivered for trials with $ET < 7$ s. However, the smallest *ET* value that triggered reward delivery was progressively raised during learning (see Methods). **c)** Schematic description of an error trial. Early *ET*s triggered an extra-running penalty and an audio noise. *Inset*: the duration of the penalty period was 10 s for the shortest *ET*s and fell linearly to 1 s for *ET*s approaching 7 s. **d)** Schematic description of an omission trial (no beam crossing between 1.5 and 15 s). **(b-d)** Note that *ET*s started to be detected 1.5 s after the motor start.

to obtain a reward, animals had to reach one of the extremities of the treadmill after a fixed time interval.

The treadmill had a length of 90 cm and was surrounded by 4 walls. The front wall was equipped with a device delivering drops of a sucrose solution. A ceiling-mounted video camera continuously recorded the position of the animal on the treadmill. An infrared beam, located at 10 cm of the front wall, defined the limit of a "reward area" (Figure 1a). Animals were first familiarized with the apparatus and trained to lick drops of the sucrose solution delivered every minute while the treadmill was immobile (see Methods). Then, rats were trained once a day (Mondays to Fridays) for 55 minutes in the time estimation task. Each daily session contained ∼130 trials interleaved with resting periods of 15 s (intertrial, motor and ambient light off). Each trial started by turning the treadmill's motor on at a fixed speed of 10 cm/s. The conveyor belt moved toward the rear of the

treadmill (Figure 1a). One second before motor onset, an ambient light was turned on (to warn the animals of the imminence of the belt movement, Figure 1). The animals' entrance time in the reward area (ET, detected by the first interruption of the infrared beam in each trial) relative to a goal time (GT, 7 s after motor onset) defined 3 types of trials. Trials in which animals entered the reward area after the GT were classified as correct ($7 \leq ET < 15$, Figure 1b). Trials in which animals entered the reward area before the GT were classified as error ($1.5 \leq ET < 7$, Figure 1c). If in 15 s an animal had not interrupted the infrared beam, the trial ended and was classified as omission (Figure 1d). Interruptions that occurred during the first 1.5 s ($ET < 1.5$) were ignored to give the opportunity to the animals to leave (passively or actively) the reward area at the beginning of each trial. Additionally, the exact value of the ET determined a reward/punishment ratio. The reward was a drop of sucrose solution and the punishment was a period of extra running. The volume of the sucrose solution delivered, increased linearly for ET values between 1.5 s (minimal reward) and GT (maximal reward) and decreased again between GT and 15 s (end of trial). The running penalty started when the animals erroneously crossed the infrared detector before GT (error trial) and its duration varied between 10 s and 1 s, according to the error magnitude (Figure 1c, inset). In addition, to progressively encourage the animals to enter the reward area just after the GT, the smallest ET value that triggered reward delivery was raised across sessions, according to each animal's performance, until it reached the GT (Figure 1b, inset and see Methods for details). Thus, to maximize reward collection and minimize running time, animals should cross the infrared beam just after the GT.

During the first training sessions, animals started most trials in the front of the treadmill, mostly ran in the reward area and interrupted the infrared beam before the GT (Video 1, Figure 2a, top, d, left). Progressively, across training sessions, animals waited longer and after $\sim$15 sessions, they reliably entered the reward area just after the GT (Figure 2b,c). Interestingly, for a large majority of animals, this ability of precisely waiting 7 s before entering the reward area was associated with the performance of a stereotyped motor sequence on the treadmill (Video 2, Figure 2a, bottom, and d, right). First, animals began each trial in the reward area. Then, when the treadmill was turned on, they remained largely still while being pushed away from the reward area until they reached the rear wall. Finally, after reaching the rear wall, they ran across the treadmill, without pause, and crossed the infrared beam. Even though a strong preference for the reward area was observed for both correct and error trials, the probability to start a trial in the frontal portion of the treadmill was higher for correct trials compared to error trials (Figure 2e), a tendency that developed progressively during training (Figure S1). In addition, if an animal started a trial in the frontal portion of the treadmill, the probability of reaching the rear portion of the treadmill was higher in correct trials than in error trials (Figure 2f), confirming that correct trials were associated with the animals performing a front-back-front trajectory. However, a significant fraction of the animals (14/54) did not develop such a strategy (Figure 2d right, Figure S2a). Based on their median trajectory after training, these animals were divided into a group that ran mostly just behind the infrared beam ("front" group, $n = 9$ rats), and another group that performed an intermediate form of the front-back-front trajectory and started running forward in the middle portion of the treadmill ("middle" group, $n = 5$ rats, Figure S2a). Compared to animals from the middle and front groups, those performing the full front-back-front sequence entered the reward area later, displayed reduced variability and an increased percentage of correct trials (Figure S2b-d). While we cannot exclude that animals from the middle and front groups

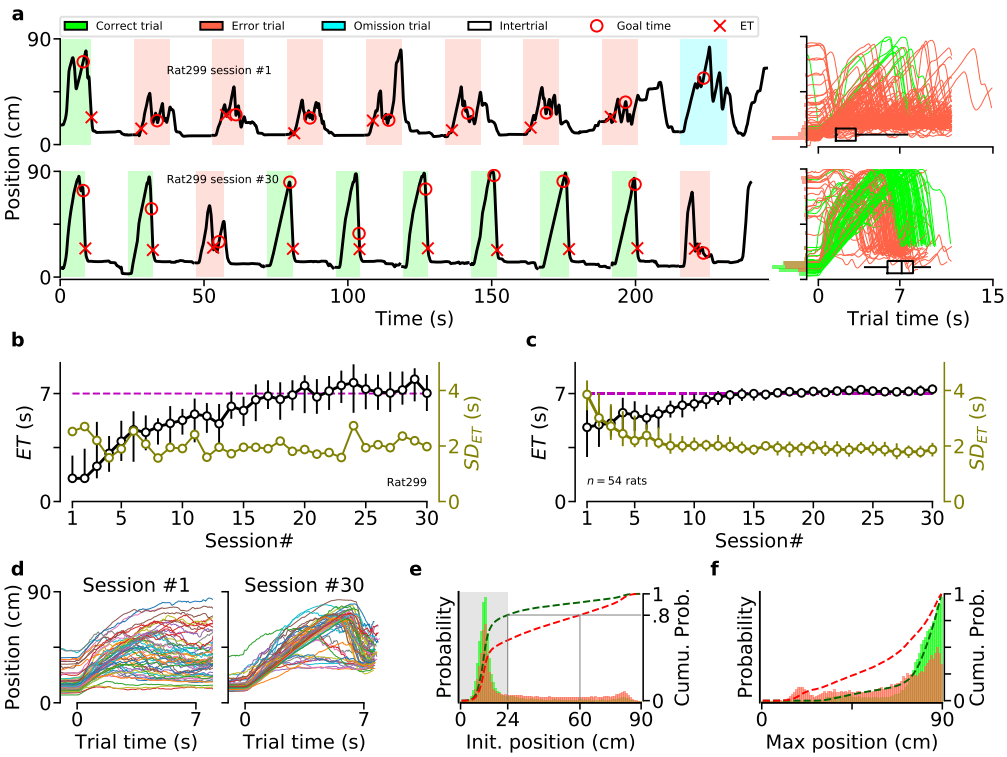


Figure 2: Most animals developed a unique stereotyped motor sequence. a) *Left*: illustration of an animal's trajectory on the treadmill during 9 consecutive trials of the 1st (*top*) and 30th (*bottom*) training sessions. On the y-axis, 0 and 90 indicate the treadmill's front (reward port) and rear wall, respectively. *Right*: trajectories for all trials during the 1st (*top*) and 30th (*bottom*) sessions (same animal as left panels). Distributions of initial positions for correct (green) and error (red) trials are shown on the y-axis. Black horizontal boxplots depict entrance time range (center line, median; box, 25th and 75th percentiles; whiskers, 5th and 95th percentiles). **b)** Median entrance time (ET) in the reward area for the first 30 daily training sessions (same animal as in panel a). Error bars show ET range (25th and 75th percentiles). The dashed magenta line shows the goal time (7 s). The right y-axis shows standard deviation (SD) of ET. **c)** Similar to panel b, but for all the animals. Circles indicate group median and error bars, the median range (25th and 75th percentiles) across animals for ET and SD of ET (SD_{ET}) values. **d)** Median trajectory of all the trials for the 1st (*left*) and 30th (*right*) training sessions. Each line represents a single animal ($n = 54$). **e)** Probability distribution function (PDF) of the position of the animals at the beginning of each correct (green) and error (red) trial, from sessions #20 to #30. Dashed lines represent cumulative distribution functions (right y-axis). The gray area indicates that in trained animals, 80% of correct trials began with the animal located near the front of the treadmill. **f)** PDF of the maximum position along the treadmill reached by animals before crossing the beam ($= ET$). Only trials in which animals were initially located in the front of the treadmill (gray area in panel e) were included.

also used a more subtle stereotyped motor sequence not captured by tracking the average body positions along the treadmill length, the above results suggest that performing the front-back-front trajectory through a "wait-and-run" motor sequence is the most reliable strategy to accurately enter the reward area just after 7 s.

It could be argued that a combination of the task parameters (length of the treadmill, its speed and direction, possibility to start trials in the reward area, position of the infrared beam) and the length of the rats' body (from head to tail) favored the development of this stereotyped strategy. Indeed, it takes about 6 s for the animals to passively reach the back of the treadmill (Figure 2a,d)

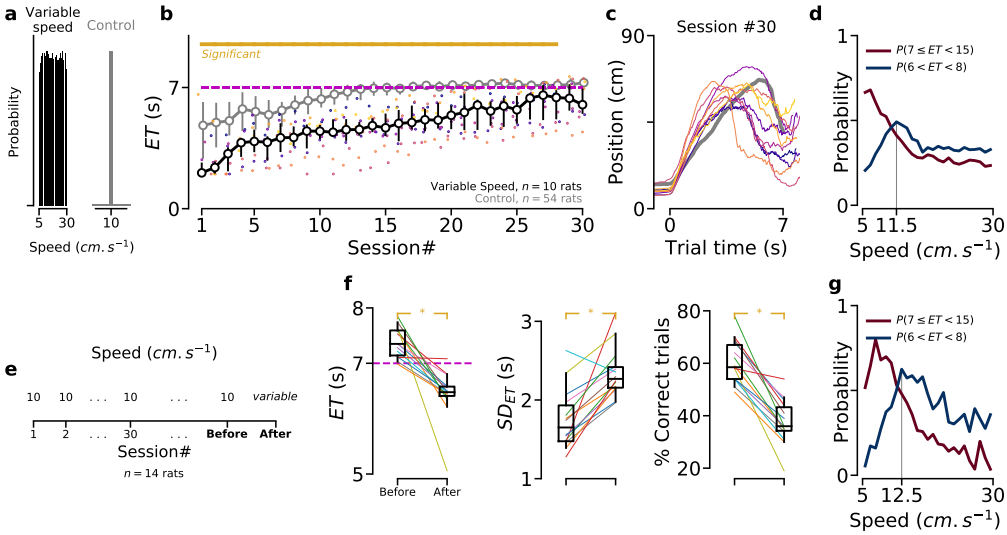


Figure 3: Decreased temporal accuracy when the treadmill speed changes across trials. a) For each trial, treadmill speed was either fixed at 10 cm/s (control condition, same data as in Figure 2), or randomly selected from a uniform distribution between 5 and 30 cm/s (variable speed condition). **b)** Median ET for animals trained in the variable speed (black), and control (gray) conditions. Colored dots indicate individual performance for "variable speed" animals. Yellow line shows statistically significant difference between groups (see Methods). **c)** Median trajectory of "variable speed" animals in session #30 (same colors as in panel b). **d)** Probability of correct ($7 \leq ET < 15$ s) and precise ($6 < ET < 8$ s) trial, given the treadmill speed, for "variable speed" animals (session # $\geq$20). **e)** After extensive training in control condition, animals ($n = 14$) were tested in a probe session with variable speed. **f)** Median ETs (*left*), SD of ETs (*middle*) and percentage of correct trials (*right*) in the sessions immediately before and after the change in speed condition. Each line represents a single animal. Yellow star indicates significant difference ($p < 0.0001$ for all three comparisons, see Methods). **g)** Similar to panel d, for the data collected from the probe session.

after which they can start running toward the reward area without the need to estimate time. Thus, in the following experiments, we examined how accurately animals respected the GT, when distinct task parameters were modified such as to prevent the use of this simple wait-and-run motor sequence. First, we trained a new group of rats in a version of the task in which, for each trial, the speed of the treadmill was selected randomly from a uniform distribution between 5 and 30 cm/s (Figure 3a). We found that, during the course of training, these animals consistently failed to wait as long as the animals trained in the control version of the task ("control" group, Figure 3b). Still, the average trajectories of animals extensively trained in this "variable speed" condition revealed that they performed a front-back-front trajectory (Figure 3c). Accordingly, the probability of performing a correct trial, given different speeds, fell rapidly from 5 to ~15 cm/s and was lowest for the fastest treadmill speeds (Figure 3d). Indeed, when the treadmill speed was fast, performing the wait-and-run strategy result in error trials, as animals reach the back region of the treadmill earlier than when the treadmill speed was slow. We also found that the probability of entering the reward area at the GT$\pm$1 s sharply peaked for a treadmill speed (11.5 cm/s) that is suitable to perform the wait-and-run motor sequence (Figure 3d). Finally, when rats extensively trained in the control version of the task underwent a single probe session with variable speeds (Figure 3e), all measures of performance dropped significantly (Figure 3f). Examining the probability of correct trials and accurate ET (7 ± 1 s)

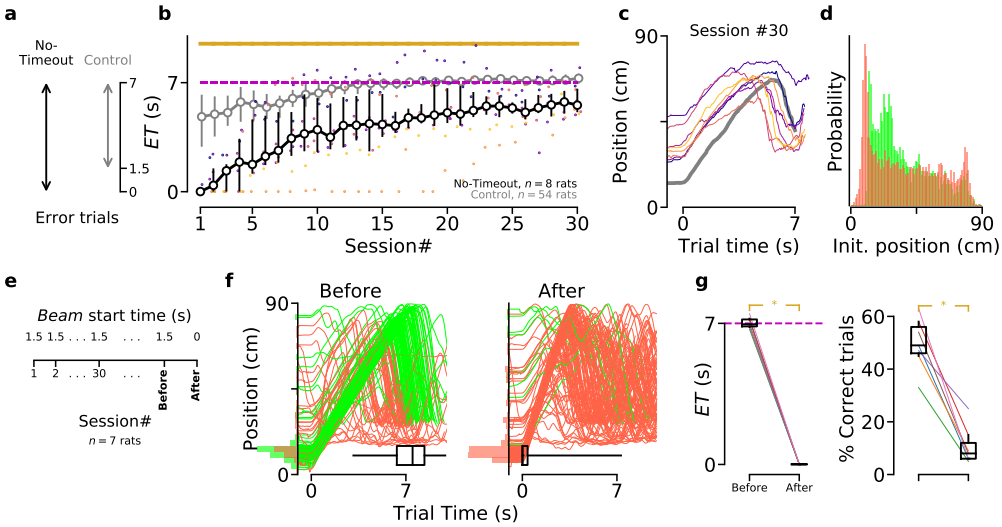


Figure 4: Decreased temporal accuracy when animals are penalized for starting trials in the reward area. **a)** In control condition, animals had a 1.5 s timeout period to leave the reward area after motor onset. In "no-timeout" condition, crossing the infrared beam any time before 7 s is considered as an error. **b)** Median *ET* for animals trained in the no-timeout (black), and control (gray) conditions. Colored dots indicate performance for individual "no-timeout" animals. **c)** Median trajectory of no-timeout animals (same colors as in panel b) in session #30. **d)** PDF of the no-timeout animals' positions at the beginning of each trial, from sessions #20 to #30. **e)** After extensive training in control condition, animals ($n = 7$) were tested in a no-timeout probe session, in which the beam started at the beginning of the trial, rather than 1.5 s later. **f)** Trajectories of a representative animal in the last "control" session (*left*), and the probe session (*right*). **g)** Median *ET*s (*left*), and percentage of correct trials (*right*) in the sessions immediately before and after the change in beam start time. Each line represents a single animal. $p < 0.0001$ for both comparisons.

given the treadmill speed suggested that, animals kept performing the front-back-front trajectory they previously learned in the control condition (compare Figure 3g and Figure 3d).

In the control condition, ∼80% of correct trials started while animals were in the reward area (Figure 2e). If rats relied on an internal clock-based algorithm to accurately time their entrance in the reward area, they should adapt relatively easily to a perturbation of their initial starting position. To test this prediction, we trained a group of rats in a modified version of the task that penalized them when they started the trials in the front region of the treadmill. This was done by activating the infrared beam as soon as the motor was turned on (in the control condition, the infrared beam was inactive during a *timeout* period that lasted 1.5 s after treadmill onset). In this "no-timeout" condition, error trials corresponded to *ET*s occurring between 0 and 7 s after motor onset (Figure 4a). Animals trained in this condition never reached the level of timing accuracy displayed by animals in the control condition (Figure 4b). Still, no-timeout animals performed a front-back-front trajectory (Figure 4c) and correct trials were associated with the animals starting the trials just behind the infrared beam (Figure 4d). The stereotyped reliance on the wait-and-run strategy was also demonstrated by the fact that rats extensively trained in control condition kept performing the exact same trajectory when tested in a single probe session under no-timeout condition, leading to a sharp decrease in performance (Figure 4e-g).

We next examined how animals behaved when the goal time (GT) was set to 3.5 s (Figure 5),

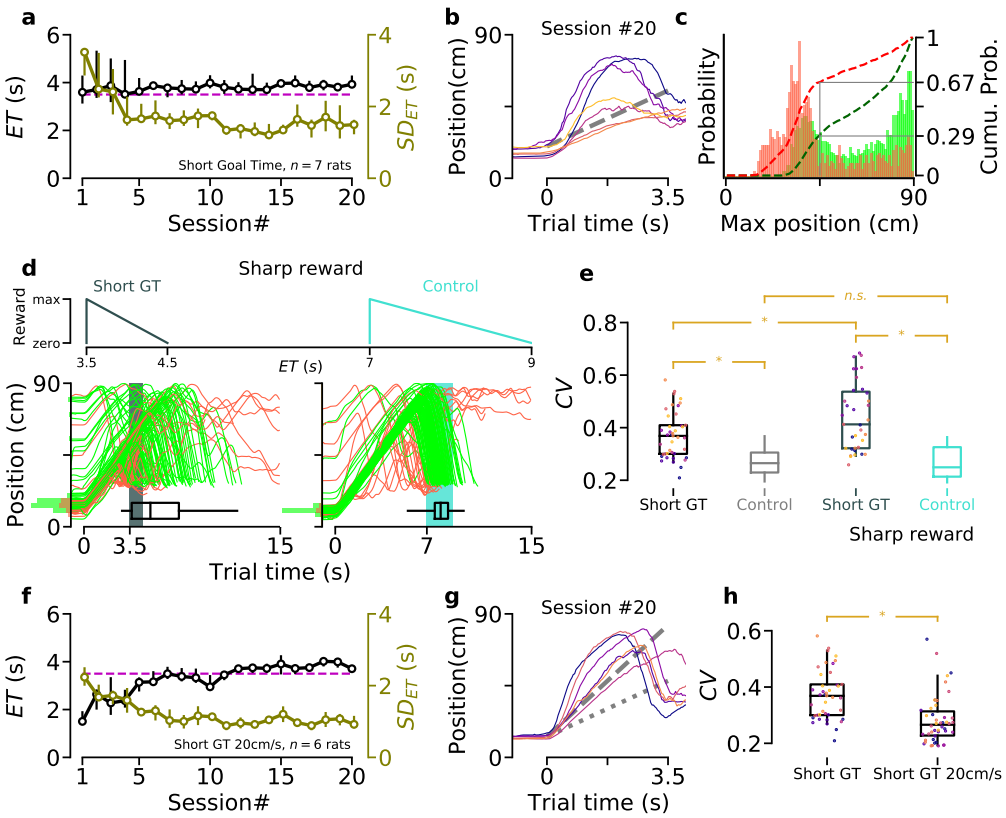


Figure 5: Decreased temporal accuracy when the goal time is shortened. **a)** Median entrance time (ET) during training. The dashed magenta line shows the goal time ($GT = 3.5$ s). The right y-axis shows standard deviation (SD) of ET. **b)** Median trajectory of "short GT" animals after training. Colored lines indicate performance of individual animals. Dashed line's slope shows the treadmill speed (10 cm/s). **c)** PDF of the maximum position reached by short GT animals before ET for correct (green) and incorrect (red) trials. Dashed lines represent cumulative distribution functions (right y-axis). Data collected from session# ≥ 15. **d)** Sharp reward condition applied to short GT and control experiments. *Top*: reward profiles in the sharp condition. *Bottom*: trajectories of 2 illustrative sessions after training in sharp condition (*left*, short GT; *right*, control). Highlighted areas indicate the reward window. **e)** Coefficient of variation (CV) for short GT and control experiments with normal (first two boxes), and sharp (last two boxes) reward profiles. Data collected once performance plateaued (after session #15 for short GT, between session #20 to #30 for control, and last 5 sessions for the sharp condition experiments). Short GT ($n = 41$) vs. Control ($n = 540$): $p < 0.0001$; Sharp short GT ($n = 35$) vs. Sharp control ($n = 40$): $p < 0.0001$; Short GT vs. Sharp short GT: $p = 0.0068$; Control vs. Sharp control: $p = 0.7$. **f)** Similar to panel a, for another group of animals that were trained to wait for 3.5 s while the speed of the treadmill was 20 cm/s. **g)** Similar to panel b, for animals trained in short goal time 20 cm/s condition (panel f). Dashed line's slope shows the treadmill speed (20 cm/s). Dotted line's slope indicates control treadmill speed (10 cm/s). **h)** CV for short GT and short GT 20 cm/s conditions (same colors as in panel b,g). Data collected once performance plateaued (after session #15). Short GT ($n = 41$) vs. Short GT 20 cm/s ($n = 46$): $p < 0.0001$.

a condition in which the performance of the wait-and-run strategy would lead to late ETs (and smaller rewards) because it takes ∼6 s for the animals to passively travel from the front to the rear portion of the treadmill. Animals successfully entered the reward area after 3.5 s and reduced their variability across training sessions (Figure 5a) but as a group, they displayed an increased

ET variability compared to animals trained in the control condition, with GT set 7 s (Figure 5e). From the averaged trajectories of "short GT" animals measured once their performance plateaued, it appeared that 3 subjects out 7 performed the front-back-front trajectory by running toward the rear portion of the treadmill. The other 4 animals remained still when the treadmill was turned on and tried to run forward before reaching the rear wall (Figure 5b). Interestingly, after training, in 67% of the error trials, the rats started running forward before reaching the middle of the treadmill (Figure 5c, compare with red histogram in Figure 2f). Conversely, after initiating a trial in the reward area, the probability of visiting a deeper portion of the treadmill was much stronger in correct than error trials, reinforcing the idea that accurate timing was accomplished by exploiting the most salient physical features of the environment (Figure 5c). Accordingly, the 3 rats that performed the front-back-front running sequence were less variable than those that passively stayed still before running toward the reward area from the middle of the treadmill (Figure 5e, same color code as in panel b). In addition, among animals trained in the short goal time condition, we found that the magnitude of the backward displacement on the treadmill was negatively correlated with ET variability ($r = -0.49, p = 2.7 \times 10^{-3}$, Pearson's correlation). In the short GT condition, animals become proficient more rapidly than in the control condition (compare Figure 5a with Figure 2c). Could the increased ET variance when the GT is 3.5 s be explained by the fact that the task is easier in this condition and that animals do not need to be very precise? To test this possibility, we increased the penalty for early ETs and decreased reward size for late ETs. In this "sharp reward" condition, the performance of the animals trained in the short GT was even more variable (Figure 5d-e). This result confirms that under short GT condition animals can not accurately time their entrance in the reward area. Finally, another group of animals was trained with GT set to 3.5 s and treadmill speed at 20 cm/s (i.e., twice as fast, such as performing the front-back-front trajectory through the wait-and-run motor sequence would lead to ETs close to the GT, Figure 5f). These animals displayed reduced ET variability compared to animals trained at 10 cm/s, and after treadmill onset they stayed immobile until reaching the end of the treadmill, similar to animals trained in the control condition (Figure 5g,h).

The above results suggest that, in a time interval task, the possibility to perform a stereotypical motor sequence adapted to salient features of the environment (here, taking advantage of the full treadmill length and its physical boundaries) critically determines temporal accuracy. To further investigate this idea, we trained a group of animals in a version of the task in which the treadmill was never turned on (trial onset was signalled by turning the ambient light on). In this condition, animals displayed a strong impairment in respecting the GT, compared to animals trained in the control condition (Figure 6a,b). On average, animals reached the reward area later and later across sessions but displayed a constant high variability in ET (Figure 6c). We also noticed that correct trials preferentially occurred when animals crossed the treadmill from the rear wall to the reward area (Video 3, Figure 6a,d,e). Accordingly, after extensive training, a robust correlation was observed on a session by session basis between the percentage of correct trials and displacement of the animal on the treadmill (Figure 6f). Moreover, the probability of a correct trial increased for higher values of displacement (Figure 6g).

Lastly, animals trained in the immobile treadmill condition during several weeks were challenged in the control condition (i.e., by simply setting the treadmill speed at 10 cm/s). These animals improved their behavior at the same pace and with the same front-back-front strategy as naïve

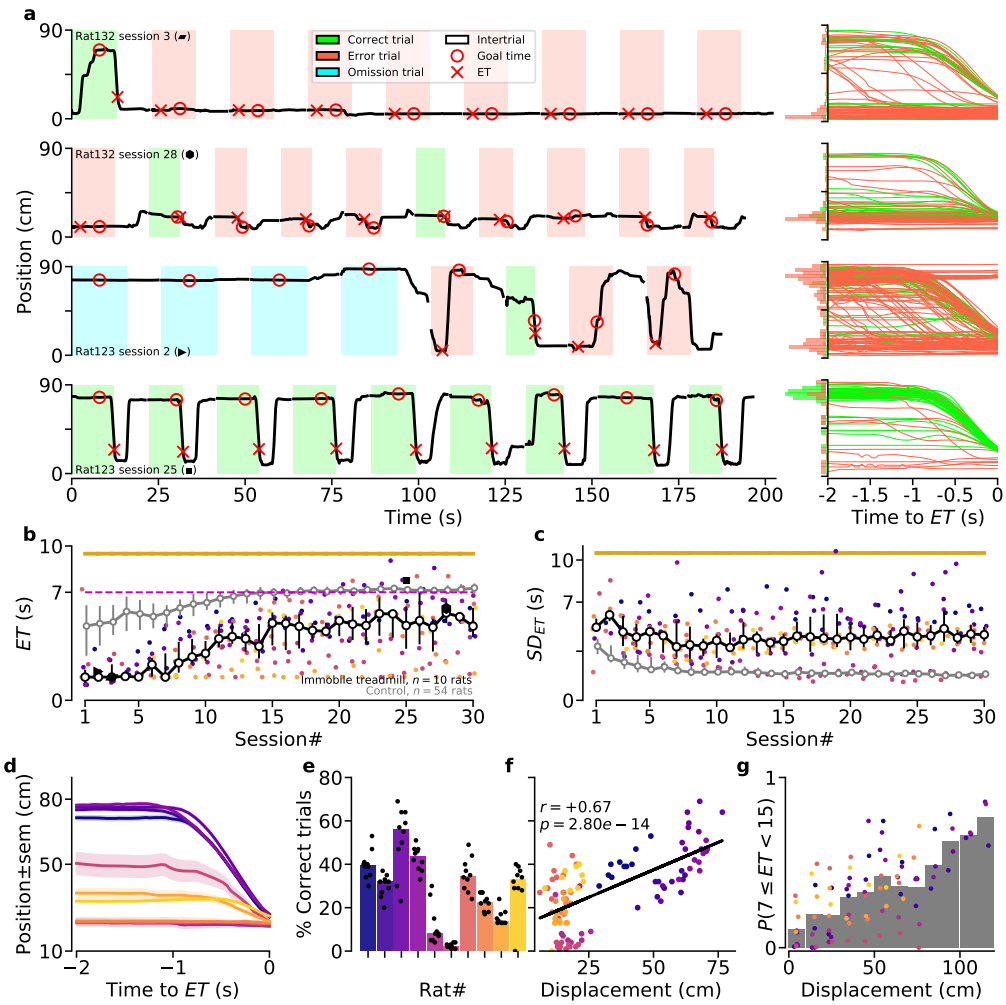


Figure 6: Performance of animals trained while the treadmill remained immobile. a) *Left*: illustrations of the positions of two animals on the immobile treadmill for 9 consecutive trials, early (*1st row:* Rat #132-session #3, *3rd row:* Rat #123-session #2) and late (*2nd row:* Rat #132-session #28, *4th row:* Rat #123-session #25) during training. *Right*: trajectories for all the trials of the corresponding sessions on the left, aligned to the *ET*. Distributions of positions 2 s before *ET*, for correct (green) and error (red) trials are shown on the y-axis. **b)** Median *ET* across sessions for "immobile treadmill" animals. Filled black markers correspond to the sessions illustrated in panel a. **c)** Similar to panel b, for the standard deviation of entrance times(SD_{ET}). **d)** Median trajectory aligned to *ET* of each "immobile treadmill" animal (only correct trials from sessions #20 to #30 are considered; shaded regions denote standard error). **e)** Median percentage of correct trials for each immobile treadmill animal (same sessions as in panel d). Each dot represents one session. **f)** Pearson's correlation between percentage of correct trials and average displacement during a session. Each dot represents one session. **g)** PDF of a correct trial, given the displacement of an animal. Each dot represents the average probability for an individual animal, during a single session. **(e-g)** Analyses based on the same sessions as in panel d. Individual animal color code is preserved in panels b-g.

animals (Figure 7a-c). Thus, animals that previously learned to wait in one version of the task did not learn faster than naïve animals when challenged in a second version of the task with distinct movement requirement but identical GT, demonstrating again that task proficiency relied primarily on the acquisition of a motor sequence rather than an internal knowledge of time.

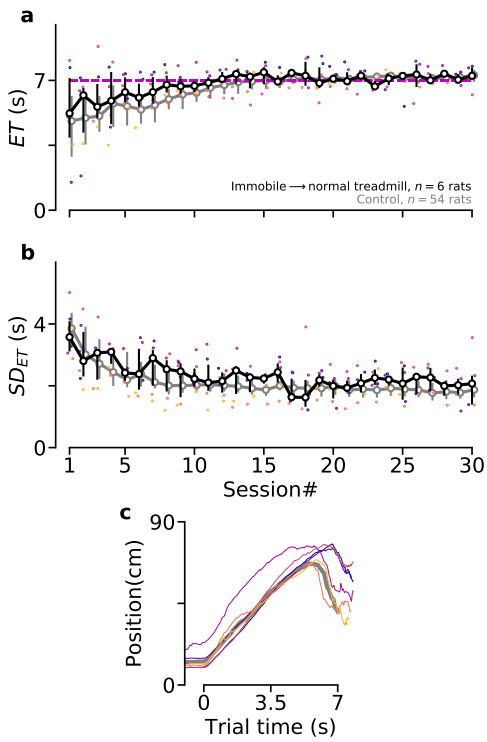


Figure 7: Lack of temporal knowledge transfer across task protocols. After extensive training on the immobile treadmill, animals were trained under normal conditions (GT= 7 s, treadmill speed= 10 cm/s). **a)** Median ET across sessions in control condition. **b)** Similar to panel a, for the standard deviation of entrance times (SD_{ET}). **c)** Median trajectory of the individual animals after relearning the task in the control condition. **a-c)** Individual animal color code is preserved in all panels.

Altogether, our results suggest that rats respected the 7 s time constraint of the task by taking advantage of the space available on the treadmill to converge to the front-back-front motor sequence as a recipe of task proficiency. We assumed that animals developed this strategy through a trial-and-error learning process without any help of a disembodied sense of time. To assess the validity of this assumption, we developed a reinforcement learning algorithm used by artificial agents exposed to a simplified version of our task (see Methods for details). Briefly, the treadmill was divided into 5 regions (with the most frontal region corresponding to the reward area) and time was also discretized. At each time step, the agent had to choose one of the three following actions: stay immobile to move one region backward, run at the treadmill speed to remain in the same region, or run at twice the treadmill speed to move forward (Figure 8a, top). The choice of the action entirely depended on the current state of the agent. Critically, each region corresponded to two different states, depending on whether the agent has reached the back of the treadmill or not (Figure 8a, bottom). The position of the agent at the beginning of each trial was random with a bias toward the reward area, to mimic the natural tendency of rats to stay close to the reward area during intertrials. We found that, across training sessions, an artificial agent waited longer and longer and reduced its entrance time variability by performing the front-back-front trajectory (Figure 8b,c). The same strategy was developed by agents endowed with different learning parameters and exploration/exploitation rates

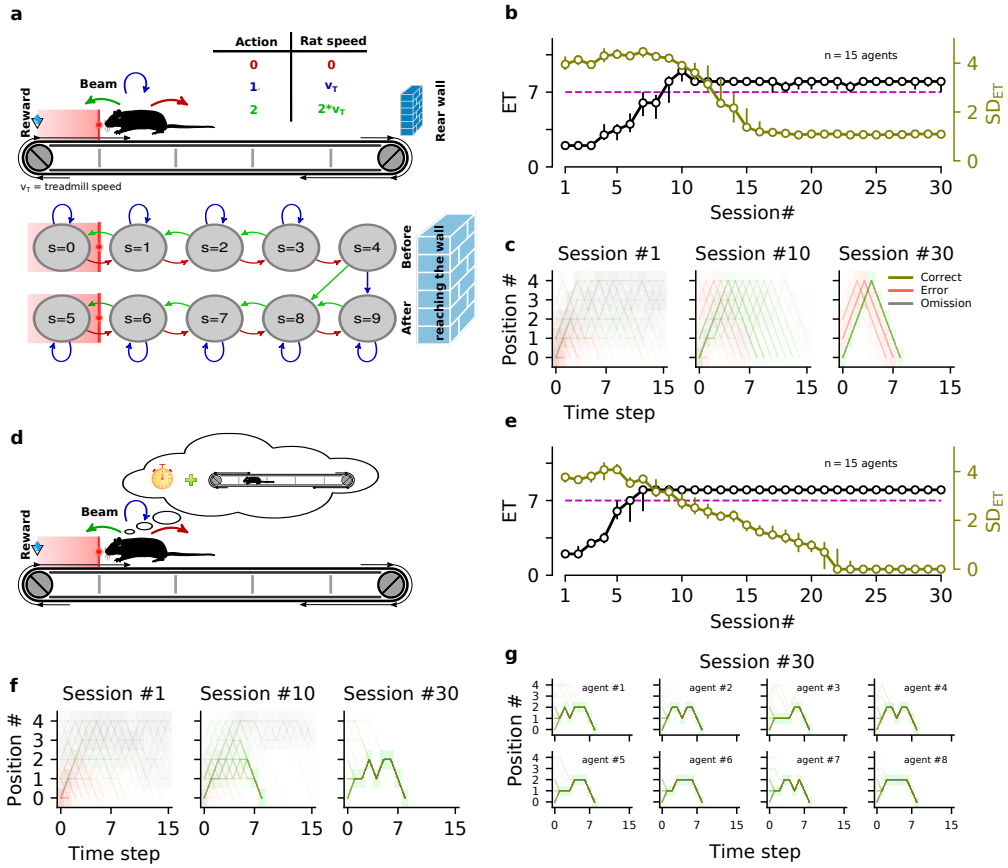


Figure 8: Artificial agents without time representation performed similarly to animals. **a)** Schematic representation of the state-action space and transitions used to model the treadmill task for agents that did not use time information. Space and time were discretized. The same physical location corresponded to 2 different internal states before and after reaching the rear wall. At each time step, agents could choose one of the three following actions: staying still (red arrow, action $a = 0$); moving at the treadmill speed v_T (blue arrow, $a = 1$); or moving at twice v_T (green arrow, $a = 2$, see Methods). **b)** Average learning profile for several agents with different learning parameters (see Methods). Median entrance time (ET) for the first 30 sessions. Error bars show ET range (25th and 75th percentiles). The dashed magenta line shows the goal time (7 time steps). The right y-axis shows standard deviation (SD) of ET. **c)** Trajectories of three sessions at different stages of learning. Each session contained 100 trials. For each session, the space-time occupancy matrix was normalized to its maximum value, for correct (green), error (red), and omission (gray) trials. **d)** Same as in panel a, when agents have access to the time elapsed since trial onset. **e-f)** Same as panels b-c, respectively, but for agents following the model sketched in panel d. **g)** Trajectories after training for 8 different agents that accessed time information.

(Figure S3). After learning, agents performed the front-back-front routine independently of their initial positions. Thus, they arrived in the reward area too early when their initial position was near the middle of the treadmill, in a striking similarity with the behavior of well-trained rats. Next, we examined how artificial agents would perform the same task if they had information not only about their position, but also about the time elapsed since trial onset (Figure 8d). Such agents also learned progressively to enter in the reward area at the right time and reduced their variability (Figure 8e). However, after learning, they did not perform the front-back-front trajectory. Rather, they mainly

stayed behind the infrared beam and were capable of respecting the GT independently of their initial position (Figure 8f). Finally, different agents developed idiosyncratic position trajectories (even if they remained close to the reward area during most of the trial duration), due to stochastic variations in their learning dynamics (Figure 8g). Altogether, these results support the hypothesis that the behavior developed by most rats in our task is consistent with the usage of a motor sequence, rather than a disembodied internal clock.

Discussion

In this study, we used a treadmill-based behavioral assay that required rats to wait for a specific amount of time before approaching a reward. Animals may accurately time their approaches using either one of the following two algorithms. First, they may use an internal knowledge of time to learn how long they should wait and decide when to approach the reward port. The second algorithm is based on the animals' ability to develop a motor routine adapted to the environment, whose completion would take them to the reward at the right time. The reliance on either one or the other algorithm can be directly inferred from behavioral experiments in which critical task parameters are manipulated. Indeed, if animals were to use an *internal* representation of time, behavioral accuracy should be largely independent of variations in *external* factors (e.g., the speed of the treadmill, the animals position on the treadmill at the beginning of the trial,...). In contrast, if rats were to rely on a motor sequence adapted to their environment, accuracy in timing will heavily depend on task-specific features of the environment or the order of the elements composing the motor sequence. The results of our behavioral experiments clearly favor the latter embodied (and situated) strategy. Using two distinct reinforcement learning-based algorithms that either incorporated or lacked time representation, we confirmed that the behavior of our animals is incongruent with them accessing an internal knowledge of elapsed time. We find this result particularly interesting because it questions the algorithmic meaning of the many neuronal correlates of elapsed time observed in humans [34–36] and animals [33, 37–40].

To accurately wait 7 seconds before approaching the reward port, most rats took advantage of the possibility to start each trial in the front region of the treadmill, the constant speed and direction of the treadmill (which slowly carried them away from the reward port after trial onset), and the length of the treadmill, to develop the following stereotyped sequence of actions. First, rats waited for the beginning of each trial close to the reward area. Then, upon trial onset, they stayed relatively still until the treadmill carried them to the rear wall of the treadmill. Finally, as soon as they reached the back of the treadmill, they ran straight to the reward port. In this experimental "control" condition, the accuracy of the animals reached its peak after 15 to 20 training sessions. However, even for proficient animals, the probability of performing a correct trial was almost null when they started a trial in the back region of the treadmill. In addition, when animals started a trial in the reward area, performing a correct trial was almost exclusively associated with the animals reaching the back portion of the treadmill. Finally, following extensive training in the control condition, when we modified the task parameters to penalize the stereotyped performance of the front-back-front trajectory, the behavioral proficiency and accuracy of the animals dropped dramatically. These results are in good agreement with the hypothesis that animals do not use a time representation independent of variations in their motor performance (be it the exact position of an animal when a

trial just started, or the precise time at which they reached the back of the treadmill).

It could be argued that the animals' tendency to develop this front-back-front trajectory resulted from the structure of the task that provided an easy solution that animals used instead of estimating time while continuously running just behind the infrared beam. However, several additional experiments confirmed that rats have limited ability to use an internal representation of time when the task parameters are set such as to prevent the usage of a stereotyped motor sequence to solve the task. First, we trained a group of animals while the treadmill speed randomly changed across trials. Compared to animals trained in the control condition, those trained with variable speed were less accurate. Additionally, these animals attempted to use the front-back-front trajectory as shown by an increased probability of correct trials when the treadmill speed allowed it. Second, we trained rats in a version of the task that penalized them when they started the trials in the front region of the treadmill. In this condition, solving the task is not possible using stereotyped middle-back-front or a back-front trajectories (which would generate early entrance times in the reward area). Rats trained in this condition displayed strong accuracy impairment. Moreover, they kept trying to develop a modified front-back-front trajectory and started the trials as close as possible to the infrared beam. In all the above experiments, during trials, the treadmill pushed the animals away from the reward area which may favor using the front-back-front trajectory. To avoid this possible bias, in our last experiment, we trained a group of rats on an immobile treadmill. Rats' performance was poor in this condition, with some animals failing to show any signs of learning. Furthermore, animals that did eventually learn, performed a motor sequence during which they ran to the back, performed some movements in the back of the treadmill (that we could not quantify with our video tracking system) and then rushed back to the front. Altogether, we conclude from this set of experiments that rats, forced to wait for several seconds before doing a certain action, did not seem capable of using an internal disembodied representation of time, but always attempted to develop a motor sequence in the confined space of the treadmill, a sequence whose execution duration amounted to the time they needed to wait.

One particularly interesting set of results concerns unexpected changes in the variability of entrance time in the reward area between task conditions, which apparently violated the classical scalar property of variance. First, entrance times were more variable when animals were trained with the short goal time (3.5 s), compared to the long goal time (7 s). How can we explain such results? For short goal times, the wait-and-run front-back-front strategy is not optimal, as animals would enter the reward area too late. Thus, the increased variability might be explained by the difficulty for the rats to "self-estimate" when to start running forward without the help of a clear sensory cue (such as touching the back wall). In support of this idea, in 67% of the error trials, the rats started running forward before reaching the middle of the treadmill. In addition, a few animals trained in the short goal time condition developed a new stereotyped motor sequence (running from front to back and back to front) and their entrance times were less variable than animals that remained immobile after trial onset. Another violation of the scalar property of variance is the big difference in ET variability between control and immobile treadmill conditions, despite similar goal times. We noticed that the best rats in the immobile treadmill condition systematically ran to the back region of the treadmill where they performed a series of rearing and wall-touching movements, just before crossing the treadmill toward the reward area. Thus, the increased variability is likely to result from the fact that, when the treadmill is immobile, a motor sequence to fit in 7 s is more difficult

to reproduce reliably than in the control condition in which most of the sequence is a passive wait on the treadmill until the animal reached the rear wall. Thus, the scalar property of variance may primarily reflect a difficulty in performing long motor sequences, rather than long time itself. In a study in which a group of rats had to perform two lever presses interleaved by 700 ms, each animal developed an idiosyncratic motor sequence (e.g., 1# first press on the lever with the left paw; 2# touching the wall above the lever with the right paw; 3# second press on the lever with the left paw), lasting precisely 700 ms [32]. The large inter-individual variability reported in this study may arise from the multiple possibilities of simple action sequences that can be squeezed in such a short time interval and easily reproduced across trials. If the time interval was longer, all the animals might have developed the same motor sequence (e.g., running back and forth in the experimental cage between the two lever presses).

It is well known that temporal regularities in animal conditioning protocols favor the development of automatic motor sequences. Still, in one of the rare studies that carefully examined the full body dynamics of rats performing a time interval categorization task, it was reported that animals developed brief motor sequences triggered by the occurrence of the first cue [31]. Importantly, the choice of the animal could be predicted by examining the execution of the motor sequence. Thus, rats may also use an embodied strategy in tasks requiring them to categorize (short or long) the duration of time intervals.

Our conclusion that animals rely on an embodied strategy, rather than internal neuronal clocks (dedicated or emergent), does not mean that internal brain activity is irrelevant to time-related behavior. We do not question that representations of elapsed time have been observed in individual and population neuronal activity in various brain regions during time-constrained tasks or that perturbation of neuronal activity impairs timing accuracy or discrimination. However, it has been recently clarified that the algorithmic meaning of patterns of neuronal activity that correlate with certain aspects of behavior can remain obscure, even when complemented with perturbation experiments [41]. Our behavioral results are not easily compatible with the idea that neural representations of time are a signature of a clock-like algorithm for time estimation. In addition, we found that an artificial agent endowed with time representation solved a virtual implementation of our task in a different way from rats. Nevertheless, the embodied view of time perception is compatible with the idea that timing is associated with spiking dynamics of neuronal populations distributed across several brain regions, including motor-related cortical and subcortical regions [18, 19]. Our work specifically suggests that this population activity (neuronal ensemble trajectories) is constrained by the dynamics of movements and sensory signals. Thus, well-timed behavior emerges from reciprocal and repetitive interactions between the nervous system, the body (sensors and actuators) and the surrounding environment.

In our experimental setting, rats may decide when to run forward after gathering sufficient evidence (based on a mixture of visual information, sensory stimulation of their tail or their back). Thus, it may be relevant to consider the process governing when the rats will run forward as an accumulation of sensorimotor evidence. The dorsal striatum is critical for processing sensorimotor information [42] and has been proposed to contribute to the process of evidence accumulation during decision making [43]. Interestingly, it has been recently proposed that a competition between the direct and indirect basal ganglia pathways, tuned by dopaminergic modulation, may determine the speed of evidence accumulation toward decision taking [44, 45]. Such a model predicts that, in

the context of our task, an increase (decrease) in dopaminergic activity will speed up (slow down) the accumulation of sensorimotor information and will lead to an early (delayed) response. Such a scenario is in agreement with a recent study in mice performing an auditory duration discrimination task, showing that increased dopaminergic activity in the substantia nigra compacta was associated with the animals perceiving long tones as short ones [46]. Future work should further investigate how to mathematically capture the algorithms underlying temporally constrained actions and decisions and the possible contribution of the striatum to processing multimodal (sensorimotor, cognitive, emotional) contextual information together with dopamine-mediated feedback and motivational signals.

Methods

Subjects

Subjects were male Long-Evans rats. They were 12 weeks old at the beginning of the experiments, housed in groups of 4 rats in temperature-controlled ventilated racks and kept under 12 h–12 h light/dark cycle. All the experiments were performed during the light cycle. Food was available *ad libitum* in their homecage. Rats had restricted access to water while their body weights were regularly measured. A total of 111 rats were used in this study (the number of animals in each experimental condition is systematically shown in its respective figure). No animal was excluded from the analysis. All experimental procedures were conducted in accordance with standard ethical guidelines (European Communities Directive 86/60 - EEC) and were approved by the relevant national ethics committee (Ministère de l'enseignement supérieur et de la recherche, France, Authorizations #00172.01 and #16195).

Apparatus

Four identical treadmills were used for the experiments. Treadmills were 90 cm long and 14 cm wide, surrounded by plexiglass walls such that the animals were completely confined on top of the treadmill. Each treadmill was placed inside a sound-attenuating box. Treadmill belt covered the entire floor surface and was driven by a brushless digital motor (BGB 44 SI, Dunkermotoren). A reward delivery port was installed on the front (relative to the turning direction of the belt) wall of the treadmill and in case of a full reward, released a $\sim$80 μL drop of 10% sucrose water solution. An infrared beam was installed 10 cm from the reward port. The first interruption of the beam was registered as entrance time in the reward area (ET). A loudspeaker placed outside the treadmill was used to play an auditory noise (1.5 kHz, 65 db) to signal incorrect behavior (see below). Two strips of LED lights were installed on the ceiling along the treadmill to provide visible and infrared lighting during trials and intertrials, respectively (see below). The animals' position was tracked via a ceiling-mounted camera (Basler scout, 25 fps). A custom-made algorithm detected the animal's body and recorded its centroid as animal's position. The entire setup was fully automated by a custom-made program (LabVIEW, National Instruments). Experimenter was never present in the behavioral laboratory during the experiments.

Habituation

Animals were handled 30 m per day for 3 days, then habituated to the treadmill for 3 to 5 daily sessions of 30 min, while the treadmill's motor remained turned off and a drop of reward was delivered every minute. Habituation sessions resulted in systematic consumption of the reward upon delivery.

Behavioral Task (Normal Condition)

Treadmill Waiting Task

Training started after handling and habituation. Each animal was trained once a day, 5 times a week (no training on weekends). Each of the daily sessions lasted for 55 min and contained ∼130 trials. Trials were separated by intertrial periods lasting 15 s. During intertrials, the treadmill remained in the dark and infrared ceiling-mounted LEDs were turned on to enable video tracking of the animals. Position was not recorded during the last second of the intertrials to avoid buffer overflow of our tracking routine and allow for writing to the disk. The beginning of each trial was cued by turning on the ambient light, 1 s before motor onset. Since animals developed a preference to stay in the front (i.e., close to the reward port), the infrared beam was turned on 1.5 s after trial start. This *timeout* period was sufficient to let the animals be carried out of the reward area by the treadmill, provided they did not move forward. After the first 1.5 s, the first interruption of the beam was considered as ET. The outcome of the trial depended solely on the value of the ET, compared to the goal time (GT= 7 s, except for the short goal time condition). In a correct trial ($ET \geq GT$, see Figure 1), infrared beam crossing stopped the motor, turned off the ambient light, and triggered the delivery of reward. In an error trial ($ET < GT$, see Figure 1), there was an extended running penalty for a duration determined by the rule displayed in Figure 1c, inset. During the penalty, the motor kept running, the ambient light stayed on and an audio noise indicated an error trial. In omission trials wherein animals didn't cross the beam in 15 s since the motor start, trial stopped and reward was not delivered.

Reward Profile The magnitude of the reward was a function of the ET and animal's performance in previous sessions (only in early training, see Figure 1b, inset). Reward was maximal at $ET = GT$ and dropped linearly to a minimum (= 38% of the maximum) for ETs approaching 15 s (i.e., the maximum trial duration). Moreover, in the beginning of the training, partial reward was also delivered for error trials with $ET > ET_0$, where ET_0 denotes the minimum threshold for getting a reward. The magnitude of this additional reward increased linearly from zero for $ET = ET_0$, to its maximum volume for $ET = GT$. In the first session of training, $ET_0 = 1.5$ s and for every following session, it was updated to the maximum value of median ETs of the past sessions. Once ET_0 reached the GT, it was not updated anymore (late training reward profile in Figure 1b, inset).

Alternative Training and Testing Conditions

Variable Speed Condition

In this condition, for each trial, treadmill speed was pseudo-randomly drawn from a uniform distribution between 5 and 30 cm/s. During any given trial, the speed remained constant. We used

5 cm/s as the lowest treadmill speed. Lower speeds generated choppy movements of the conveyor belt. Also, velocities higher than 30 cm/s were not used, to avoid any physical harm to the animals.

No-timeout Condition

In the control condition, the infrared beam was not active during the first 1.5 s of the trials. This *timeout* period was sufficient to let the animals be carried out of the reward area by the treadmill, provided they did not move forward. In the "no-timeout" condition, the infrared beam was activated as soon as the trial started. Thus, in this condition, error trials corresponded to ETs between 0 and 7 s. Consequently, animals were penalized if they were in the reward area when the trial started (i.e., $ET = 0$ s).

Short Goal Time Condition

In this condition, the goal time (GT) was set to 3.5 s, half the value for the control condition. The reward profile in this condition followed the same rules as for the control condition, except that reward was maximal at $ET = GT = 3.5$ s. Two different groups of animals were trained in this condition, one with treadmill speed set to the normal value of 10 cm/s, and another with treadmill running twice as fast (20 cm/s, see Figure 5). In the short goal time condition, we also examined if the increased variability in ET could be attenuated when the penalty associated with early ET was increased and when reward magnitude was decreased for late ET. This was implemented by doubling the treadmill speed during the penalty period (from 10 cm/s to 20 cm/s), and the reward was delivered for a narrower window of ETs (maximal reward at $ET = GT = 3.5$ s, and no reward after $ET = 4.5$ s). For proper comparison, we also examined the behavior of rats trained with $GT = 7$ s when the running penalty was increased and the reward was decreased for late ETs (maximal reward at $ET = GT = 7$ s, and no reward after $ET = 9$ s, see Figure 5d,e).

Immobile Condition

In this condition, the treadmill's motor was never turned on. The ambient light was turned on during the trials and turned off during the intertrials. Error trials were penalized by an audio noise and extended exposure to the ambient light.

Statistical Test

All statistical comparisons were performed using a permutation test previously described in [47]. This non-parametric method alleviates many concerns in statistical hypothesis tests, such as distribution assumptions (e.g., normality assumption under analysis of variance), error inflation due to multiple comparisons, and sensitivity to unbalanced group size.

To simplify the description, let's assume, as in Figure 3b, we have $\mathbf{X} = [X_1, X_2, ..., X_n]$, where X_i is the set of ETs of session i. Similarly, we have $\mathbf{Y}$ that contains ETs of all the sessions from another condition. Here, the null hypothesis states that the assignment of each data point in X_i and Y_i to either $\mathbf{X}$ or $\mathbf{Y}$ is random, hence there is no difference between $\mathbf{X}$ and $\mathbf{Y}$.

In short, the test statistic was defined as the difference between smoothed (using Gaussian kernel with $\sigma = 0.05$) $\mathbf{X}$ and $\mathbf{Y}$ for each session i: $D_0(i)$. At this point, we generated one set of surrogate

data by assigning each ET of session i to either X_i or Y_i, randomly. For each set of surrogate data, the test statistic was calculated, i.e., $D_m(i)$. This process was repeated 10,000 times for all the statistical comparisons in this study, obtaining: $D_1(i), \ldots, D_{10000}(i)$.

At this step, two-tailed pointwise p-values could be directly calculated for each i, from the $D_m(i)$ quantiles (see [47]). Moreover, to compensate for the issue of multiple comparisons, we defined global bands of significant differences along the session index dimension. From 10,000 sets of surrogate data, band of the largest α-percentile was constructed, such that less than 5% of $D_m(i)$s broke the band at any given session i. This band (denoted as the *global band*) represents the threshold for significance, and any break-point by $D_0(i)$ at any i is a point of significant difference between $\mathbf{X}$ and $\mathbf{Y}$.

In cases of comparing only two sets of data points (such as in Figure 3f), the same algorithm was employed, having only one value for index i. If none of the $D_m(i)$s exceeded $D_0(i)$, the value $p < 0.0001$ was reported (i.e., less than one chance in 10,000).

Reinforcement Learning Models

Here, we used the Markov Decision Process (MDP) formalism to analyze how artificial agents learn to perform a simplified version of the treadmill task. According to the MDP formalism, at each time step, the agent occupies a state and selects an action. The probability to transition to a new state depends entirely on the previous action and state, and each transition is associated with a certain reward. The agent tries to maximize future rewards and, in our simulations, we used a simple Q-learning algorithm ([48], see below) to model the way the agent learned an optimal policy (i.e., which action to take for any possible state).

We modeled the treadmill task using a deterministic environment in which the time was discretized and the treadmill was divided in 5 regions of equal length. In this simplified setting, we simulated two types of agents that differed only by the type of the information available to them to select actions and analyzed how their behaviour varied.

The first type of agents did not use an explicit representation of time to perform the task. At each time step t, the state s_t (i.e., the information used to select actions) consisted in the agent's position p_t, in the treadmill and in a boolean variable w_t, whose value was equal to 1, if the agent had previously reached the rear wall during the trial and 0, otherwise. Given these assumptions, each state can be written as $s_t = \{p_t, w_t\}$ and the state space consisted of 5 pair of states (a total of 10 states).

The second type of agents instead, benefited of information on both its position on the treadmill and the elapsed time since the beginning of the trial. Thus, each state was represented as $s_t = \{p_t, t\}$.

For both types of agents, the task was simulated in an episodic manner and the initial position p_0 at the beginning of each trial was assigned randomly as follows: the probability $P(p_0)$ that the initial state corresponds to p_0 was proportional to $q(1-q)^{p_0}$ for $p_0 = 0, \ldots, 4$. We set the parameter $q = 0.5$ such as to account for the tendency of the rats to initiate trials in the reward area.

During the rest of the trial, at each time step t, agents occupied a state s_t, and could select one of three different actions that determined a transition to a new state s_{t+1}. Action $a_t = 0$ corresponded to remaining still and, considering that the treadmill was on, moving one position backward in the treadmill. Action $a_t = 1$ consisted in moving at the same speed of the treadmill (v_T), but in the opposite direction. Thus after performing this action, the agents remained at the same position on the treadmill. Finally, performing action $a_t = 2$, the agents moved at twice the treadmill speed which

made him move one position step forward. We also introduced two physical constraints that limited the action space at the extreme sides of the treadmill. In the front of the treadmill, the agents cannot move forward (i.e., when the position was $p = 0$ the action $a = 2$ was forbidden). In the rear of the treadmill the agents could not stay still, as otherwise it would hit the rear wall (i.e., when $p = 4$ the action $a = 0$ was not available).

After entering a new state at time $t + 1$, the agents received a reward $r_{t+1} = \bar{r}$. The value $\bar{r}$ varied depending on the position p_{t+1} and on the current time t. Similarly than in the real task, the agent had to reach the most frontal region of the treadmill (equivalent of the reward area) after 7 time steps (the minimum ET in the frontal region to obtain a reward is 8 time steps). We also created an equivalent of the time out period (see above in experimental method section), such as the agent was not penalized to start a trial in the reward area. Still, the agents had to leave the front of the treadmill (i.e., $p = 0$) within 2 time steps. Finally, agents had a maximum amount of time (15 time steps) to perform the task. More specifically, reward rules were as follows. The punishment associated with an early ET ($2 \leq ET < 8$) had a maximum (negative) value of $\bar{r} = -2$ and its absolute value decreased linearly between 2 and 7. Correct trials occurred when agents reached the frontal region of the treadmill between 8 and 15 time step ($8 \leq ET \leq 15$), which delivered a reward with a maximum value of $\bar{r} = +3$, that decreased linearly with ET. Omission trials (i.e., those trials in which the agent did not approach the front area within 15 time steps) were associated with the delivery of a small punishment $\bar{r} = -0.5$. We also modeled the cost of the passage of time while the treadmill was on, by adding a small punishment $\bar{r} = -0.1$ at each time step in all trial types.

Agents learned the value (expressed in terms of future rewards) of selecting a particular action in a specific internal state via the Q-learning algorithm. Specifically, for any state-action pair $\{s, a\}$, a state-action value function $Q(s, a)$ can be defined as follows:

$$Q(s, a) = E\Big[G_t \mid s_t = s, a_t = a\Big] \tag{1}$$

where $G_t = \sum_{i=0}^{T-t} \gamma^i \cdot r_{t+1+i}$ is the discounted sum of expected future rewards, and γ is the discount factor ($0 \leq \gamma \leq 1$). Equation 1 implies that each value $Q(s, a)$ is a measure of the future reward that the agent expects to receive after performing action a when its current state is s.

Following the Q-learning algorithm, after each time step t, the $Q(s_t, a_t)$ will change according to:

$$\Delta Q(s_t, a_t) = \alpha \left(r_{t+1} + \gamma \, max_{a'}\{Q(s_{t+1}, a')\} - Q(s_t, a_t)\right) \tag{2}$$

where the parameter α represents the learning rate.

These state-action values are then used to determine the policy π: a mapping from states to actions (i.e., the way agents acted in any possible state). In our model, the policy was stochastic and depended on the Q-values via a *softmax* distribution:

$$P(a \mid s_t) = \frac{exp(\beta Q(s_t, a))}{\sum_{a'} exp(\beta Q(s_t, a'))} \tag{3}$$

where the parameter β governs the exploitation/exploration trade-off (when $\beta \to 0$, the policy becomes more and more random).

Updates in Equation 2 can be proved to converge to the optimal Q-value for each pair $\{s, a\}$ [48]. Optimal value means the value (in terms of rewards) that action a assumes in state s, when the policy of agent across all the sequence of states and actions is such to maximize future rewards. Therefore

selecting actions with a probability that increases with the Q-values allows learning of the optimal behavior.

We used the formalism described above to simulate $n = 15$ agents of the first type and $n = 15$ of the second type. Each agent differed in the exploitation/exploration parameter (see below) and performed the task for 30 sessions of 100 trials each. The exploitation/exploration parameter started with an initial value β_0, and was increased after each session of training by an amount $\Delta\beta$ (i.e., the policy became more and more greedy), up to a maximum of $\beta_{max} = 10$. Different agents were represented by different values of β_0 and $\Delta\beta$. The agents of our simulations corresponded to all the possible combinations of $\beta_0 = \{0, 2, 2.5, 3, 4\}$ and $\Delta\beta = \{0.3, 0.35, 0.4\}$. In all the simulation, we set the parameters $\alpha = 0.1$, and $\gamma = 0.99$.

Data Analysis

Data from each session was stored in separate text files, containing position information, entrance times, treadmill speeds, and all the task parameters. Position information was scaled to the treadmill length, and smoothed (Gaussian kernel, $\sigma = 0.3$ s). The entire data processing pipeline was implemented in python, using open-source libraries and custom-made scripts. We used a series of Jupyter Notebooks to process, quantify, and visualize every aspect of behavior, to develop and run the reinforcement learning algorithms, and to generate all the figures in this manuscript. All the Jupyter Notebooks, as well as the raw data necessary for full replication of the figures and videos are publicly available via the Open Science Foundation (`https://osf.io/7s2r8/?view_only=7db3818dcf5e49e88d708b2597a21956`).

Author contributions

Mostafa Safaie: Conceptualization, Software, Visualization, Investigation, Formal analysis, Writing–original draft; Maria-Teresa Jurado-Parras: Conceptualization, Investigation, Writing–review & editing; Stefania Sarno: Software, Methodology, Formal analysis, Conceptualization, Visualization; Jordane Louis: Investigation; Corane Karoutchi: Investigation; Ludovic F. Petit: Methodology, Resources; Matthieu O. Pasquet: Methodology, Resources, Software; Christophe Eloy: Conceptualization, Methodology; David Robbe: Funding acquisition, Project administration, Supervision, Visualization, Writing–original draft, Conceptualization, Software.

Acknowledgements

We thank Drs. I. Bureau and J. Epsztein for critical reading of the manuscript. This work was supported by European Research Council (ERC-2013-CoG – 615699 – NeuroKinematics; D.R.) and a Centuri postdoctoral fellowship (S.S.).

Competing Interests

The authors declare no competing interests.

References

[1] P. D. Balsam and C. R. Gallistel, "Temporal maps and informativeness in associative learning," *Trends in neurosciences*, vol. 32, no. 2, pp. 73–78, 2009.

[2] A. C. Nobre and F. van Ede, "Anticipated moments: temporal structure in attention," *Nature Reviews Neuroscience*, vol. 19, no. 1, p. 34, 2018.

[3] A. Kacelnik and D. Brunner, "Timing and foraging: Gibbon's scalar expectancy theory and optimal patch exploitation," *Learning and Motivation*, vol. 33, no. 1, pp. 177–195, 2002.

[4] C. R. Gallistel, *The organization of learning.* The MIT Press, 1990.

[5] J. Gibbon, R. M. Church, W. H. Meck, *et al.*, "Scalar timing in memory," *Annals of the New York Academy of sciences*, vol. 423, no. 1, pp. 52–77, 1984.

[6] C. V. Buhusi and W. H. Meck, "What makes us tick? functional and neural mechanisms of interval timing," *Nature reviews neuroscience*, vol. 6, no. 10, p. 755, 2005.

[7] M. Wittmann, "The inner sense of time: how the brain creates a representation of duration," *Nature Reviews Neuroscience*, vol. 14, no. 3, p. 217, 2013.

[8] M. J. Allman, S. Teki, T. D. Griffiths, and W. H. Meck, "Properties of the internal clock: first-and second-order principles of subjective time," *Annual review of psychology*, vol. 65, pp. 743–771, 2014.

[9] T. W. Kononowicz, H. Van Rijn, and W. H. Meck, "Timing and time perception: A critical review of neural timing signatures before, during, and after the to-be-timed interval," *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, vol. 1, pp. 1–38, 2018.

[10] J. Gibbon, "Scalar expectancy theory and weber's law in animal timing," *Psychological review*, vol. 84, no. 3, p. 279, 1977.

[11] H. Lejeune and J. Wearden, "The comparative psychology of fixed-interval responding: Some quantitative analyses," *Learning and Motivation*, vol. 22, no. 1-2, pp. 84–111, 1991.

[12] B. C. Rakitin, J. Gibbon, T. B. Penney, C. Malapani, S. C. Hinton, and W. H. Meck, "Scalar expectancy theory and peak-interval timing in humans.," *Journal of Experimental Psychology: Animal Behavior Processes*, vol. 24, no. 1, p. 15, 1998.

[13] S. Whitaker, C. Lowe, and J. Wearden, "Multiple-interval timing in rats: Performance on two-valued mixed fixed-interval schedules," *Journal of Experimental Psychology: Animal Behavior Processes*, vol. 29, no. 4, p. 277, 2003.

[14] W. Zarco, H. Merchant, L. Prado, and J. C. Mendez, "Subsecond timing in primates: comparison of interval production between human subjects and rhesus monkeys," *Journal of neurophysiology*, vol. 102, no. 6, pp. 3191–3202, 2009.

[15] S. C. Hinton and W. H. Meck, "Frontal–striatal circuitry activated by human peak-interval timing in the supra-seconds range," *Cognitive Brain Research*, vol. 21, no. 2, pp. 171–182, 2004.

[16] J. Wang, D. Narain, E. A. Hosseini, and M. Jazayeri, "Flexible timing by temporal scaling of cortical responses," *Nature neuroscience*, vol. 21, no. 1, p. 102, 2018.

[17] C. Gallistel, A. King, and R. McDonald, "Sources of variability and systematic error in mouse timing behavior.," *Journal of Experimental Psychology: Animal Behavior Processes*, vol. 30, no. 1, p. 3, 2004.

[18] J. J. Paton and D. V. Buonomano, "The neural basis of timing: distributed mechanisms for diverse functions," *Neuron*, vol. 98, no. 4, pp. 687–705, 2018.

[19] A. Goel and D. V. Buonomano, "Timing as an intrinsic property of neural networks: evidence from in vivo and in vitro experiments," *Philosophical transactions of the Royal Society B: Biological sciences*, vol. 369, no. 1637, p. 20120460, 2014.

[20] J. T. Coull, F. Vidal, B. Nazarian, and F. Macar, "Functional anatomy of the attentional modulation of time estimation," *Science*, vol. 303, no. 5663, pp. 1506–1508, 2004.

[21] S. Droit-Volet and W. H. Meck, "How emotions colour our perception of time," *Trends in cognitive sciences*, vol. 11, no. 12, pp. 504–513, 2007.

[22] D. A. Effron, P. M. Niedenthal, S. Gil, and S. Droit-Volet, "Embodied temporal perception of emotion," *Emotion*, vol. 6, no. 1, p. 1, 2006.

[23] V. Pariyadath and D. Eagleman, "The effect of predictability on subjective duration," *PloS one*, vol. 2, no. 11, p. e1264, 2007.

[24] P. A. Gable and B. D. Poole, "Time flies when you're having approach-motivated fun: Effects of motivational intensity on time perception," *Psychological science*, vol. 23, no. 8, pp. 879–886, 2012.

[25] R. Pfeifer and J. Bongard, *How the body shapes the way we think: a new view of intelligence*. MIT press, 2006.

[26] B. Morillon and S. Baillet, "Motor origin of temporal predictions in auditory attention," *Proceedings of the National Academy of Sciences*, vol. 114, no. 42, pp. E8913–E8921, 2017.

[27] R. Núñez and K. Cooperrider, "The tangle of space and time in human cognition," *Trends in cognitive sciences*, vol. 17, no. 5, pp. 220–229, 2013.

[28] B. Winter, T. Marghetis, and T. Matlock, "Of magnitudes and metaphors: Explaining cognitive interactions between space, time, and number," *Cortex*, vol. 64, pp. 209–224, 2015.

[29] J. T. Coull and S. Droit-Volet, "Explicit understanding of duration develops implicitly through action," *Trends in cognitive sciences*, vol. 22, no. 10, pp. 923–937, 2018.

[30] S. Droit-Volet, S. Fayolle, M. Lamotte, and S. Gil, "Time, emotion and the embodiment of timing," *Timing & Time Perception*, vol. 1, no. 1, pp. 99–126, 2013.

[31] T. S. Gouvêa, T. Monteiro, S. Soares, B. V. Atallah, and J. J. Paton, "Ongoing behavior predicts perceptual report of interval duration," *Frontiers in neurorobotics*, vol. 8, p. 10, 2014.

[32] R. Kawai, T. Markman, R. Poddar, R. Ko, A. L. Fantana, A. K. Dhawale, A. R. Kampff, and B. P. Ölveczky, "Motor cortex is required for learning but not for executing a motor skill," *Neuron*, vol. 86, no. 3, pp. 800–812, 2015.

[33] P. E. Rueda-Orozco and D. Robbe, "The striatum multiplexes contextual and kinematic information to constrain motor habits execution," *Nature neuroscience*, vol. 18, no. 3, p. 453, 2015.

[34] E. Amenedo and C. Escera, "The accuracy of sound duration representation in the human brain determines the accuracy of behavioural perception," *European Journal of Neuroscience*, vol. 12, no. 7, pp. 2570–2574, 2000.

[35] V. Pouthas, N. George, J.-B. Poline, M. Pfeuty, P.-F. VandeMoorteele, L. Hugueville, A.-M. Ferrandez, S. Lehéricy, D. LeBihan, and B. Renault, "Neural network involved in time perception: an fmri study comparing long and short interval estimation," *Human brain mapping*, vol. 25, no. 4, pp. 433–441, 2005.

[36] L. Grabot, T. W. Kononowicz, T. D. La Tour, A. Gramfort, V. Doyère, and V. Van Wassenhove, "The strength of alpha–beta oscillatory coupling predicts motor timing precision," *Journal of Neuroscience*, vol. 39, no. 17, pp. 3277–3291, 2019.

[37] B. J. Kraus, R. J. Robinson II, J. A. White, H. Eichenbaum, and M. E. Hasselmo, "Hippocampal "time cells": time versus path integration," *Neuron*, vol. 78, no. 6, pp. 1090–1101, 2013.

[38] K. I. Bakhurin, V. Goudar, J. L. Shobe, L. D. Claar, D. V. Buonomano, and S. C. Masmanidis, "Differential encoding of time by prefrontal and striatal network dynamics," *Journal of Neuroscience*, vol. 37, no. 4, pp. 854–870, 2017.

[39] B.-M. Gu, K. Kukreja, and W. H. Meck, "Oscillation patterns of local field potentials in the dorsal striatum and sensorimotor cortex during the encoding, maintenance, and decision stages for the ordinal comparison of sub-and supra-second signal durations," *Neurobiology of learning and memory*, vol. 153, pp. 79–91, 2018.

[40] G. B. Mello, S. Soares, and J. J. Paton, "A scalable population code for time in the striatum," *Current Biology*, vol. 25, no. 9, pp. 1113–1122, 2015.

[41] J. W. Krakauer, A. A. Ghazanfar, A. Gomez-Marin, M. A. MacIver, and D. Poeppel, "Neuroscience needs behavior: correcting a reductionist bias," *Neuron*, vol. 93, no. 3, pp. 480–490, 2017.

[42] D. Robbe, "To move or to sense? incorporating somatosensory representation into striatal functions," *Current opinion in neurobiology*, vol. 52, pp. 123–130, 2018.

[43] M. M. Yartsev, T. D. Hanks, A. M. Yoon, and C. D. Brody, "Causal contribution and dynamical encoding in the striatum during evidence accumulation," *Elife*, vol. 7, p. e34929, 2018.

[44] K. Dunovan and T. Verstynen, "Believer-skeptic meets actor-critic: Rethinking the role of basal ganglia pathways during decision-making and reinforcement learning," *Frontiers in neuroscience*, vol. 10, p. 106, 2016.

[45] K. Dunovan and T. Verstynen, "Errors in action timing and inhibition facilitate learning by tuning distinct mechanisms in the underlying decision process," *Journal of Neuroscience*, vol. 39, no. 12, pp. 2251–2264, 2019.

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

[47] S. Fujisawa, A. Amarasingham, M. T. Harrison, and G. Buzsáki, "Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex," *Nature neuroscience*, vol. 11, no. 7, p. 823, 2008.

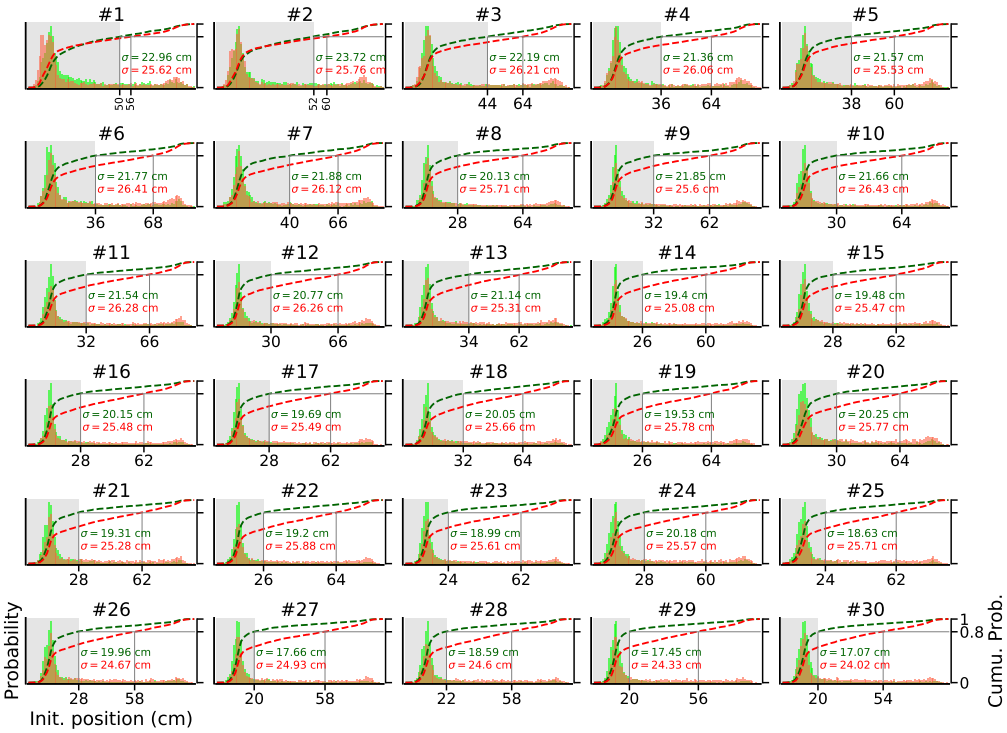
[48] R. S. Sutton, A. G. Barto, *et al.*, *Introduction to reinforcement learning*, vol. 2. MIT press Cambridge, 1998.

Supplementary Material

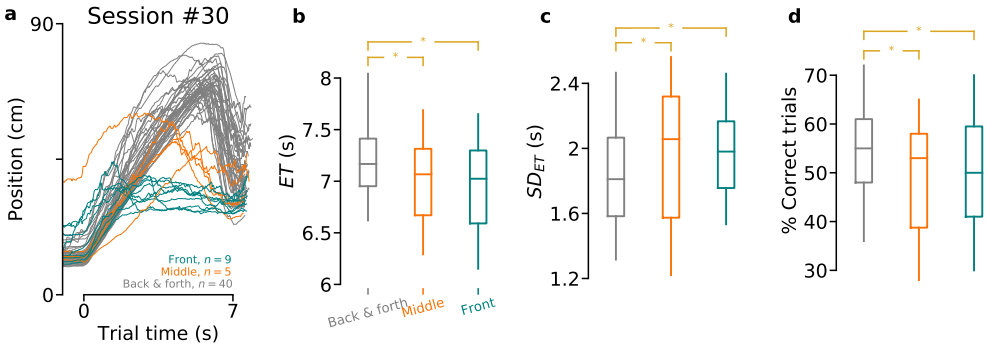
Video 1. Video clip showing several consecutive trials from an animal performing its first training session in control condition. Information about trial number, time since light on, GT, *ET*, and ongoing task status are given on the upper left corner.

Video 2. Same as Video 1 for a well-trained animal performing the task in control condition.

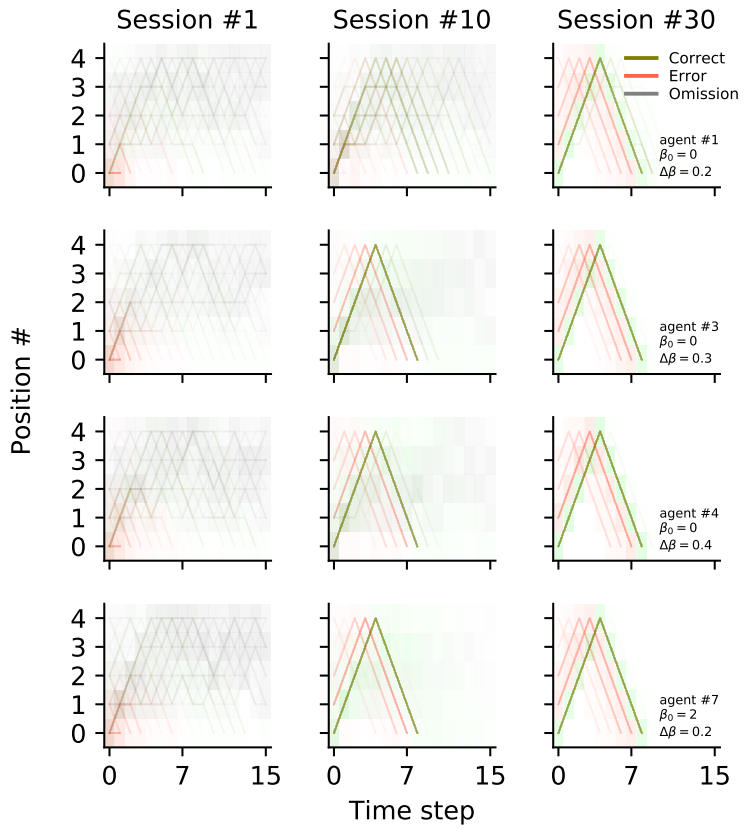
Video 3. Same as Video 2 for an animal performing the task in the immobile treadmill condition.



Supplementary Figure S1: Initial position distributions for correct and error trials diverged progressively during training. Similar to Figure 2e, each panel shows PDF of the initial position of the animals for correct (green) and incorrect (red) trials, but plotted separately for each training session (#1 to #30). Dashed lines represent cumulative distribution functions (right y-axis). For each PDF, σ values denote the standard deviation. Each PDF included pooled data from all the animals trained in the control condition ($n = 54$).



Supplementary Figure S2: Task proficiency according to the type of trajectory performed by animals. a) Same as Figure 2, panel d (right), with different colors representing the 3 main types of trajectories. **b)** Entrance times (ETs). $p < 0.0001$ for both comparisons. **c)** SD of ET. For back & forth vs. middle group: $p < 0.05$; Back & forth vs. front group: $p < 0.001$. **d)** Percentage of correct trials. $p < 0.0001$ for both comparisons. For panels b, c, d, animals were sorted according to their type of trajectory (same color code as in panel a). Only data from sessions # ≥ 20 were considered. Number of data points in each group: Back & forth= 536, Middle= 100, Front= 171.



Supplementary Figure S3: Final trajectories performed by agents are identical regardless of exploitation/exploration parameters. Similar to Figure 8c but for four different agents (differences among agents are determined by the values of the exploitation/exploration parameters β_0 and $\Delta\beta$; see Methods). Even if agents displayed different trajectories during learning (sessions #1 and #10), all of them performed the same trajectory at session #30.