Supplementary Information for

Turning the body into a clock: accurate timing is facilitated by simple stereotyped interactions with the environment

Mostafa Safaie, Maria-Teresa Jurado-Parras, et. al.

Corresponding Author: David Robbe
E-mail: david.robbe@inserm.fr

This PDF file includes:

- Supplementary text
- Figs. S1 to S5
- Legends for Movies S1 to S3
- SI References

Other supplementary materials for this manuscript include the following:

- Movies S1 to S3

www.pnas.org/cgi/doi/10.1073/pnas.1921226117

Supporting Information Text

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. The 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 and defined the limit of the reward area. In each trial, 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.

Behavior.

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.

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 $\sim$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 onset. 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. The animals' entrance time in the reward area (ET, detected by the first interruption of the infrared beam in each trial after 1.5 s) 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 S1b). Trials in which animals entered the reward area before the GT were classified as error ($1.5 \leq ET < 7$, Figure S1c). If in 15 s an animal had not interrupted the infrared beam, the trial ended and was classified as omission (Figure S1d). Additionally, the exact value of the ET determined a reward/punishment ratio. The volume of the sucrose solution delivered, increased linearly for ET values between 1.5 s (no reward) and GT (maximal reward, i.e., $\sim$80 μL) and decreased again between GT and 15 s ($\sim$30 μL for ETs approaching 15 s). During training, to progressively encourage the animals to enter the reward area after the GT, partial reward was also delivered for error trials with $ET > ET_0$, where ET_0 denotes the minimum ET value delivering a drop of sucrose solution. The size of this partial reward increased linearly from zero for $ET = ET_0$, to its maximum volume for $ET = GT$. ET_0 was raised across sessions, according to each animal's performance, until it reached the GT (Figure S1b, inset). In the first session of training, $ET_0 = 1.5$ s. For each session (except the first one), ET_0 was raised to the value of median ETs of the previous sessions. During training, ET_0 was never decreased. Once ET_0 reached the GT, it was not updated anymore (late training reward profile in Figure S1b, inset). Finally, a penalty period of extra running started when the animals erroneously crossed the infrared beam before GT ($1.5 \leq ET < 7$) and its duration varied between 10 s and 1 s, according to the error magnitude (Figure S1c, inset). This running penalty was applied for all sessions.

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 4). 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 4d,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.

Data Analysis. Position information derived from video tracking (sampling rate 25 fps) was scaled to the treadmill length, and smoothed (Gaussian kernel, $\sigma = 0.3$ s).

Motor Routine Definition. We quantified the percentage of trials in which animals performed the front-back-front trajectory (wait-and-run motor routine). Trials were considered *routine* if all the following three conditions were met: 1) the animal started the trial in the front (initial position $< 30cm$); 2) the animal reached the rear portion of the treadmill after trial onset (maximum trial position $> 50cm$); 3) the animal completed the trial (i.e., they crossed the infrared beam). The same criteria were applied to the median trajectories after training (session #30) to classify animals into two groups: those that used the front-back-front trajectory and those that did not (Figure S3).

Statistics. All statistical comparisons were performed using resampling methods (permutation test and bootstrapping). These non-parametric methods alleviate many concerns in traditional 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.

We used the permutation test to compare the performance of two groups of animals during training on a session-by-session basis, such as in Figure 2b, and Figure 3b. To simplify the description (see (1) for more details), let's assume, as in Figure 2b, we have $\mathbf{X} = [X_1, X_2, ..., X_n]$, where X_i is the set of ETs of all the animals in session i. Similarly, we have $\mathbf{Y}$ that contains ETs from another experimental 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 $\sigma = 0.05$) average of $\mathbf{X}$ and $\mathbf{Y}$ for each session i: $D_0(i)$. We then generated one set of surrogate data by assigning ET of each animal in session i to either X_i or Y_i, randomly. For each set of surrogate data, the test statistic was similarly 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 (1)). Moreover, to compensate for the issue of multiple comparisons, we defined global bands of significant differences along the session index dimension (1)). From 10,000 sets of surrogate data, a 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 $\mathbf{Y}$.

A similar permutation test was also used when comparing only two sets of unpaired data points (such as in Figure 4e, comparing control vs. short goal time groups). 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).

For paired comparisons (such as in Figure 2f), we generated the bootstrap distribution of mean differences ($n = 10000$ with replacement). Significance was reported (yellow asterisks) if 95% Confidence Interval (CI) of the pairwise differences differed from zero (i.e., zero was not within the CI) (2). For example, in Figure 2f, right, the 95% CI of pairwise differences is $(19, 27)$%. Since this interval does not contain zero, it is reported significant, whereas in Figure 4e, the CI of the comparison between normal and sharp short goal time is $(-0.17, 0.01)$ which includes zero, and hence is reported non-significant.

Exceptionally, for the comparison in Figure 4h, even though it is not paired, we used bootstrapping, because we did not have enough data points to perform the permutation test. In this case, the resampled distribution ($n = 10000$ with replacement) for each group was calculated, and it was reported significant, since the distributions did not overlap at 95% CI.

In Figure 5f, we used repeated measures correlation implemented in the Pingouin package (3). This technique relaxes the assumption of independent data points, since each animal contributes more than one.

Reinforcement Learning Models. 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 ((4), 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 in addition, benefited from the information on the elapsed time since the beginning of the trial. Thus, each state was represented as $s_t = \{p_t, w_t, t\}$.

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 on 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:

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

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:

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

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

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

Updates in Equation 2 can be proved to converge to the optimal Q-value for each pair $\{s, a\}$ (4). 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 Organization and Availability. Data from each session was stored in separate text files, containing position information, entrance times, treadmill speeds, and all the task parameters. 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).

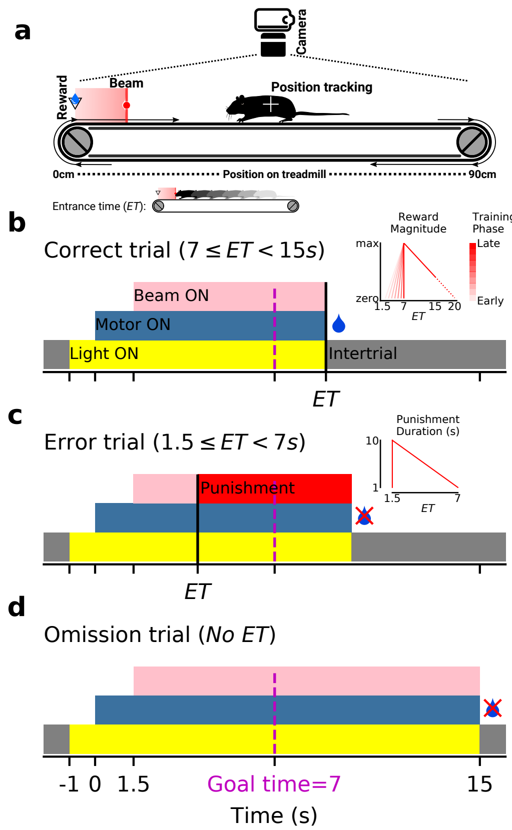

Fig. S1. Treadmill task and trial types. a) Rats were enclosed on a motorized treadmill. The infrared beam placed at 10 cm of the reward port marked the beginning of the reward area (pink shaded area). During each trial, the belt pushed the animals away from the reward area and the first infrared beam 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 SI Appendix, Methods). **c)** Schematic description of an error trial. Early ETs triggered an extra-running penalty and an audio noise. *Inset*: the duration of the penalty period was 10 s for the shortest ETs and fell linearly to 1 s for ETs approaching 7 s. **d)** Schematic description of an omission trial (no beam crossing between 1.5 and 15 s). **(b-d)** Note that ETs started to be detected 1.5 s after the motor start.

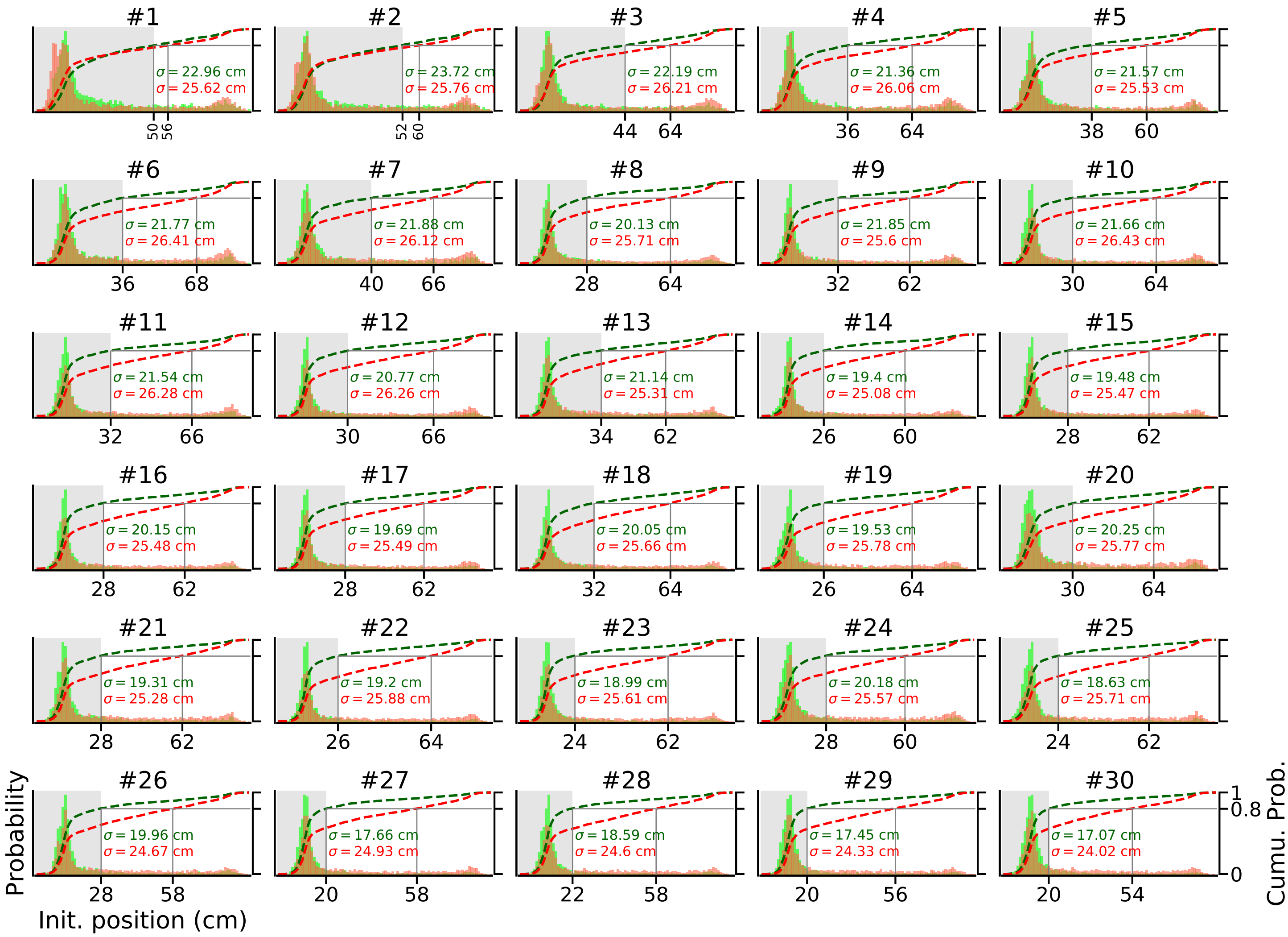

Fig. S2. Initial position distributions for correct and error trials diverged progressively during training. Similar to Figure 1e, each panel shows PDF of the initial position of the animals for correct (green) and incorrect (red) trials, but plotted separately for each training session (#1 to #30). 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 ($n = 54$).

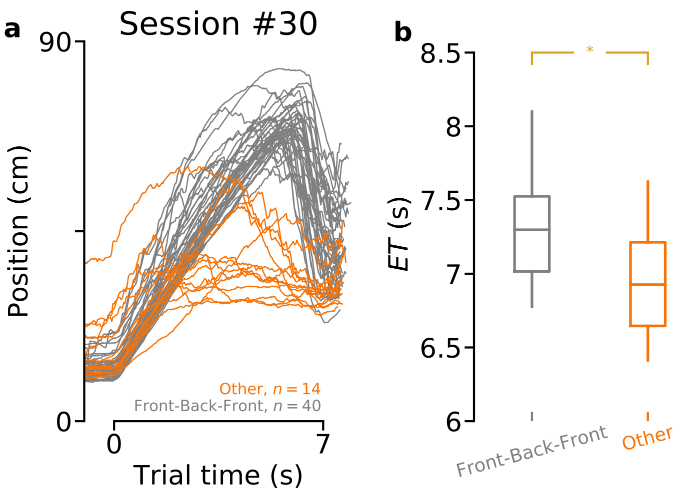

Fig. S3. Task proficiency according to the type of trajectory performed by animals. a) Same as Figure 1, panel c, right, but the animals were divided in two groups according to whether they performed the front-back-front trajectory (gray) or not (other, orange). **b)** Entrance times (ETs). $p = 0.0066$ (permutation test). **c)** SD of ET. $p = 0.03$ (permutation test). **d)** Percentage of correct trials. $p = 0.01$ (permutation test). For panels b, c, d, same color code as in panel a. Data from sessions # ≥ 20 were averaged for each animal.

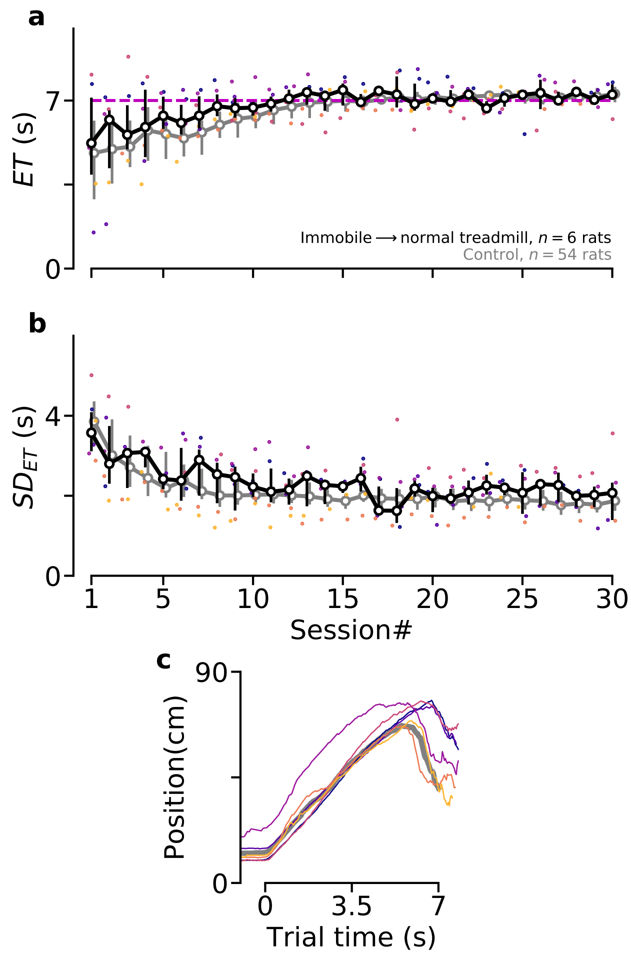

Fig. S4. Lack of temporal knowledge transfer across task protocols. After extensive training on the immobile treadmill, animals were trained under normal conditions (GT$=7$ s, treadmill speed$=10$ cm/s). **a)** Median ET across sessions in control condition. **b)** Similar to panel a, for the standard 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.

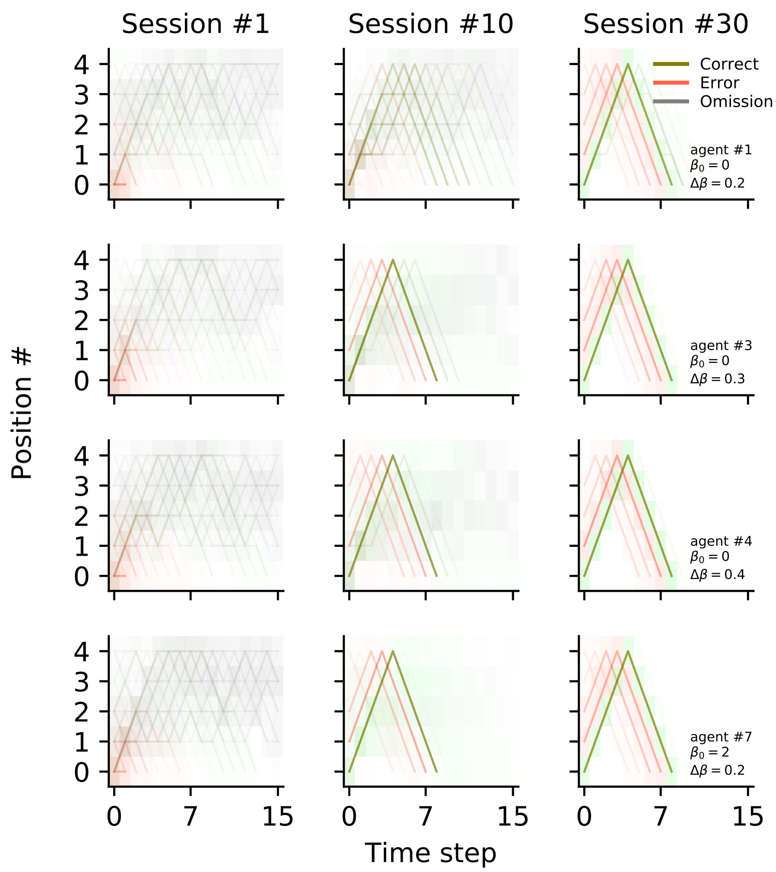

Fig. S5. Final trajectories performed by agents are identical regardless of exploitation/exploration parameters. Similar to Figure 6c but for four different agents (differences among agents are determined by the values of the exploitation/exploration parameters β_0 and $\Delta\beta$; see Methods). Even if agents displayed different trajectories during learning (sessions #1 and #10), all of them performed the same trajectory at session #30.

Movie S1. 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.

Movie S2. Same as Video 1 for a well-trained animal performing the task in control condition.

Movie S3. Same as Video 2 for an animal performing the task in the immobile treadmill condition.

References

1. S Fujisawa, A Amarasingham, MT Harrison, G Buzsáki, Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nat. neuroscience* **11**, 823 (2008).
2. B Efron, RJ Tibshirani, *An introduction to the bootstrap.* (CRC press), (1994).
3. R Vallat, Pingouin: statistics in python. *J. Open Source Softw.* **3**, 1026 (2018).
4. RS Sutton, AG Barto, , et al., *Introduction to reinforcement learning.* (MIT press Cambridge) Vol. 135, (1998).