

# Minimal model of strategy switching in the plus-maze navigation task

Denis Sheynikhovich<sup>1</sup>, Laurent Dollé<sup>2</sup>, Ricardo Chavarriaga<sup>3</sup>, Angelo Arleo<sup>1</sup>

<sup>1</sup>Laboratoire de Neurobiologie des Processus Adaptatifs UPMC-Paris 6  
CNRS UMR 7102, 9 quai St. Bernard, 75005 Paris-France

<sup>2</sup> Institut des Systèmes Intelligents et de Robotique UPMC-Paris 6  
CNRS UMR 7222, 4 place Jussieu, 75252 Paris cedex 05-France

<sup>3</sup> CNBI-EPFL, CH-1015 Lausanne, Switzerland

**Abstract.** Prefrontal cortex (PFC) has been implicated in the ability to switch behavioral strategies in response to changes in reward contingencies. A recent experimental study has shown that separate subpopulations of neurons in the prefrontal cortex were activated when rats switched between allocentric place strategies and egocentric response strategies in the plus maze. In this paper we propose a simple neural-network model of strategy switching, in which the learning of the two strategies as well as learning to select between those strategies is governed by the same temporal-difference (TD) learning algorithm. We show that the model reproduces the experimental data on both behavioral and neural levels. On the basis of our results we derive testable prediction concerning a spatial dynamics of the phasic dopamine signal in the PFC, which is thought to encode reward-prediction error in the TD-learning theory.

## 1 Introduction

The ability to switch between different navigational strategies for reaching the goal is crucial for adaptive spatial behavior. Large body of animal studies suggest a large variety of navigational strategies, which depend on sensory input of different modalities [1]. Existing lines of experimental research focus on (*i*) how the different strategies are implemented in the brain and what memory systems support them [2–5]; and (*ii*) what is the mechanism of selection between different strategies and corresponding memory systems during ongoing behavior [6–8]. In particular, a series of animal studies in the plus maze (Figure 1A) provided an insight into the role of hippocampus (HPC), dorsolateral striatum (DLS) and prefrontal cortex (PFC) in learning the tasks in which changing reward contingency forced the animals to use either a place strategy or a response strategy. In the plus maze, the place strategies can be defined as approaching a particular place associated with reward, e.g. the end of the East arm in the case of strategy 'go East' (Figure 1C). The response strategies can be defined as making a particular egocentric turn at the center, e.g. 'turn left' strategy (Figure 1D). Note that if animals start their trial in the North position, both strategies lead

to the same location. In contrast, the two strategies lead to different locations if the starting position is South, permitting the dissociation between the two strategies.

Experimental studies in the plus maze have shown that rats with inactivated HPC were strongly biased towards response strategies, while they were biased towards place strategies when DLS was inactivated [4, 5]. In other experimental environments a double dissociation between HPC and DLS has been demonstrated in place-learning and response-learning tasks, respectively [2, 3]. Moreover, recent electrophysiological recordings of PFC neurons while rats were navigating in the plus maze have demonstrated that different neuronal populations were active depending on which strategy is currently in use [8]. When rats had learned that the correct strategy was a place-based one, a subpopulation of neurons were highly active. When the reward contingency changed so that the response-based behavior was the only valid strategy, another subpopulation became active, while neurons in the first subpopulation became silent. In addition to being a suitable experimental setup for studying strategy learning and switching, the plus maze is also attractive from the modeling point of view. This is because the analysis of strategy switching and its neural correlates can be performed at a well defined location (the cross point) where behavior is crucial for successful performance (in contrast to, e.g., water maze where no such location can be identified).

The experimental data reviewed above suggest a two-level hierarchical organization of spatial behavior in the plus maze. On the first level, different neuronal networks (located in e.g. HPC or DLS) learn corresponding navigational strategies taking a particular sensory modality as an input. On the second level, a neural network (located in the PFC) learns to choose which strategy is more successful and gives the control of the behavior to that strategy [9]. In a simple way, a behavioral strategy can be considered as a rule that determines which of the available motor actions should be performed when a particular sensory input is observed. Similarly, the strategy switching can be described as the choice of one of the available strategies, given the sensory input. From a theoretical point of view, these two learning tasks are identical, assuming that available motor actions (first level) or available strategies (second level) are encoded by the activities of neural populations.

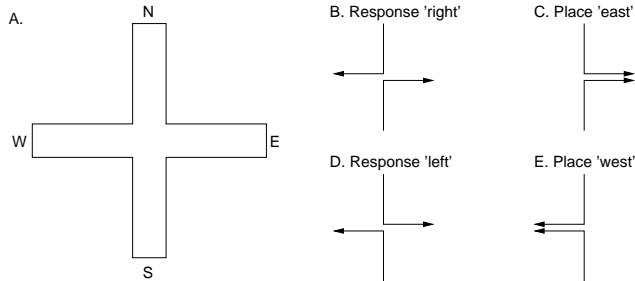
Here we were interested to see how far we can go in explaining the behavioral and neural data by suggesting the learning algorithm is identical on both levels. The learning algorithm we use is a standard temporal-difference (TD) learning rule called Q-learning [10]. At the core of TD-learning algorithms is the minimization of a so-called reward-prediction error which is proposed to be coded by the phasic activity of dopaminergic (DA) neurons in the ventral tegmental area (VTA) [11]. We have analyzed the learning dynamics of the reward-prediction error in our model and compared the error propagation on the two levels. Our results suggest that a two-level organization of behavioral control with TD-learning on each level is compatible with principal experimental results [4, 8]. In addition,

our model generated testable predictions concerning differences in DA signaling between HPC and PFC.

## 2 Model

### 2.1 Experimental setup

The simulated plus maze (Figure 1A) had distance  $L$  between the ends of the opposite arms (in arbitrary units, see all parameter values in Table 1). Arms S and N were used as starting positions, while arms W and E as rewarded (goal) locations. Simulated rats were initially trained to learn either one of the two response tasks (Figures 1B,D) or one of the place tasks (Figures 1C,E). After a certain number of trials, the task contingencies changed leading to strategy reversals or strategy switches. A strategy reversal was defined as the change from one response task to another (e.g. from ‘turn right at the cross point’ to ‘turn left at the cross point’), or from one place task to another (e.g. from ‘go east from the cross point’ to ‘go west from the cross point’). A strategy switch occurred when a response task changed to a place task, or a place task changed to a response task. On each trial, a simulated rat was released from either South or North arms in a pseudo-random order. When starting from the South (North) arm, entrance to the North (South) arm was blocked at the center. Upon reaching the goal arm, the rat was given reward  $R_t = R$  (see Equation 3 below) and a new trial started. An attempt to backtrack, or reaching the arm opposite to the goal arm resulted in  $R_t = 0$  and starting a new trial (backtracks were not counted as trials). For wall hits  $R_t = 0$ .



**Fig. 1.** Experimental setup. A. Plus maze. B-E. Learning tasks testing response strategies (B,D) and place strategies (C,E). In (B,C) and (D,E) lower paths are consistent, while upper paths are inconsistent.

### 2.2 Learning equations

Basic learning equations implemented standard Q-learning algorithm [10] in which states and actions were encoded in the firing rates of activities of arti-

ficial neurons [12, 13]. More precisely, a state  $s_t$  at time  $t$  was represented by the activities of state units  $r_i^{\text{state}}(t)$ , which projected via connections with weights  $w_{ij}$  to action units with activities  $r_i^{\text{action}}$ :

$$r_i^{\text{action}}(t) = \sum_j w_{ij} r_j^{\text{state}}(t) \quad (1)$$

Each action unit  $i$  represented an action  $a_i$  available in the state  $s_t$ . Activity of an action cell (Equation 1) was interpreted as the Q-value  $Q(s_t, a_t)$  of performing action  $a_t$  at state  $s_t$ . In the present notation, symbol  $a_i$  denotes  $i$ -th action available at the current state, while  $a_t$  denotes the action actually chosen at time  $t$  from the available actions, thus  $\forall t \exists i : a_t = a_i$ .

According to the reinforcement learning theory, true action values on subsequent time steps should be related as  $Q(s_t, a_t) = R_t + \gamma Q(s_{t+1}, a_{t+1})$ , where  $\gamma$  is the discount factor for future rewards. During learning, the weights  $w_{ij}$  were adjusted on each time step so as to enforce this relationship:

$$\Delta w_{ij}(t) = \eta \delta(t) e_{ij}(t) \quad (2)$$

where  $\eta$  is the learning rate,

$$\delta(t) = R_t + \gamma Q^*(s_t, a_t) - Q(s_{t-1}, a_{t-1}) \quad (3)$$

is the reward prediction error and  $e_{ij}(t)$  is the eligibility trace that represents the memory of past actions. The eligibility trace of a synapse [10] was increased each time the synapse had participated in generating a movement, i.e.  $e_{ij}(t+1) = e_{ij}(t) + r_j^{\text{state}}$  if action  $a_i$  was performed at time  $t$ . All eligibility traces decayed with time according to  $e_{ij}(t+1) = \gamma \lambda e_{ij}(t)$ , where  $0 < \gamma, \lambda < 1$ . In Equation 3,  $Q^*(s_t, a_t) = \max_i Q(s_t, a_i)$  is the Q-value of the optimal action at state  $s_t$ .

Outcomes of different actions at different states need to be explored in order to learn action values. Exploration was ensured by choosing actions stochastically in each state using softmax algorithm, i.e. the probability of choosing action  $a_i$  at time  $t$  was calculated according to  $p(a_t = a_i) = \exp(\beta a_i) / \sum_i \exp(\beta a_i)$  with  $\beta$  as a constant parameter.

Maze length $L$	7.0
Size of input populations $N_{\text{pc}}, N_{\text{ac}}$	13
Q-learning parameters $\gamma, \lambda$	0.9
Reward for reaching the goal $R$	10.0
Learning rate $\eta$	0.05
Softmax parameter $\beta$ (selection network/strategy networks)	1.0 / 4.0
Place field width $\sigma$	0.4

**Table 1.** Model parameters.

### 2.3 Place learning

Place strategy learning was implemented as follows. States  $s_t^P$  were encoded in the activities of  $N_{pc}$  place cells with Gaussian spatial receptive fields, centers of which were distributed uniformly over the environment. More precisely, the activity of place cell  $j$  was calculated as  $r_j^P = \exp(-d_j^2/2\sigma^2)$ , where  $\sigma$  is the width of receptive field and  $d_j$  is the distance between the position of the simulated animal in the maze and the center of the  $j$ -th receptive field. Place cells projected to place-action cells  $a_i^P$  with weights  $w_{ij}^P$ . Four place-action cells coded for movements in allocentric directions East, West, North and South. Activity of a place-action cell was calculated according to Equation 1, and the weights were updated using Equation 2 on each time step.

### 2.4 Response learning

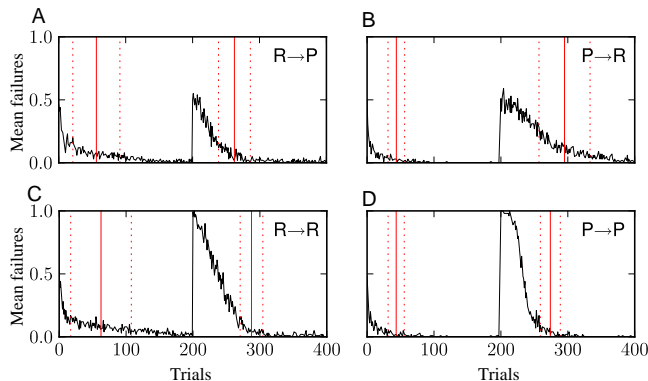
In order to implement the response strategy, states  $s_t^R$  were encoded in the activities of  $N_{sc}$  sensory cells that coded for the presence of walls around the simulated rat. All sensory cells were divided into four subpopulations coding for the space in front of the rat, left from the rat, right from the rat and behind the rat, respectively. Presence or absence of a wall at a specified egocentric direction from the rat (and hence the ability to move in this direction) was encoded by setting activities  $r_j^{sc}$  of the corresponding sensory cells to 0 or 1, respectively. Sensory cells projected to response-action cells  $a_i^R$  with weights  $w_{ij}^R$ . Four response-action cells coded for movements in egocentric directions Forward, Leftward, Rightward and Backward. Activity of a response-action cell was calculated according to Equation 1, and the weights were updated using Equation 2 on each time step.

### 2.5 Strategy-selection learning

In order to implement the selection between strategies, states  $s_t^S$  were encoded in the combined activities of  $N_{pc} + N_{sc}$  input cells from place and response strategy networks. An intuition for this particular choice of input is the following: given all available information at the current state, the selection network decides which strategy should be chosen to take control over behavior. Input cells projected to strategy-selective (action) cells  $a_i^S$  with weights  $w_{ij}^S$ . Two strategy-selective cells coded for place and response strategies, respectively. Activity of a strategy-selective cell was calculated according to Equation 1, and the weights were updated using Equation 2 on each time step.

## 3 Results

Learning performance was estimated by calculating the mean number of failures across training trials, for strategy switches and for reversals (Figure 2). A trial was defined as a failure if the simulated rat did not reach the goal arm at the



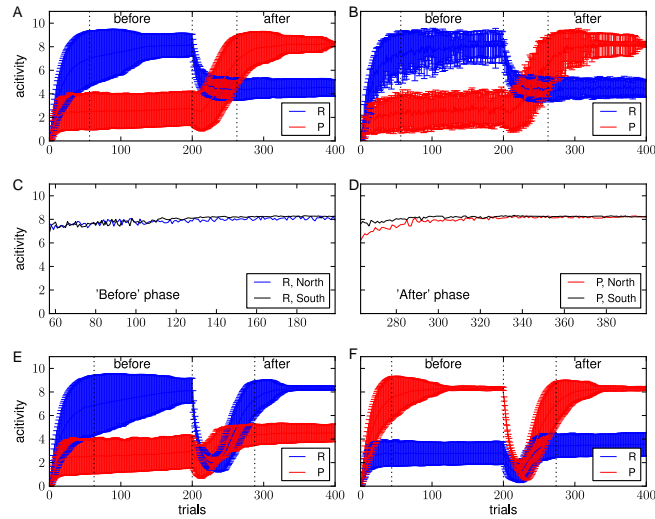
**Fig. 2.** Model performance during strategy switches (A,B) and reversals (C,D). Black lines show the mean number of failures as a function of trials. Red solid and dotted lines show mean trial number  $\pm$  SD at which the criterion was reached (see text).

end of trial. Criterion of 80% of correct trials in the last 40 trials [8] was reached within less than 100 trials for switches and reversals (red solid lines in Figure 2). For the purposes of subsequent analysis, we defined ‘before’ phase as the period between the trial in which the criterion was reached for the first time and the trial where task contingencies changed; ‘after’ phase corresponded to the period after the trial when the criterion was reached after the contingency change and until the end of training (see Figure 2).

We analyzed the activity of strategy-selective neurons during behavior. The focus was on the activity at the choice point of the simulated maze, since only at this position the differences in the activities is crucial for successful performance. During switches from response to place strategy, the response-strategy neuron was significantly more active than the place-strategy neuron during ‘before’ phase, while the activity pattern reversed during the ‘after’ phase (Figure 3A). In the model, this is a consequence of the fact that for a reward contingency corresponding to response strategy, place strategy is able to correctly predict reward only in approximately half of the trials. Conversely, when the contingency corresponds to the place strategy, reward predictability of the response strategy is only about half of that of the place strategy.

The activity of the strategy selective neurons in our model resembled the persistently changing cells recorded by Rich & Shapiro (2009) [8] ( $\approx 24\%$  of all cells recorded from PFC during strategy switches). Persistently changing cells were defined as cells that fired at significantly different rates during the ‘before’ and ‘after’ phases.

The advantage of the experimental setup used by Rich & Shapiro (2009) is that the selectivity of neural activity to a particular strategy can be dissociated from the selectivity to other behavioral variables. To clarify this idea, consider all trials started in e.g., North position (see Figure 1). When switching from,



**Fig. 3.** Activity of the response-strategy (blue) and place-strategy (red) cells during switches (A-D) and reversals (E,F). A. Cell activity during reversal from response ('left') to place ('east') task. B. The same data as in A, but only successful trials from position N are shown (i.e., only for consistent paths). C,D. Activity of the response-strategy neuron during the 'before' phase (C) and of the place strategy neuron during the 'after' phase (D). The neural activities are grouped according to starting positions. E,F. Cell activity during response (E) and place (F) reversal.

e.g., response strategy 'turn left' to place strategy 'go east', the 'north-to-east' path remains to be a valid (or *consistent*) path, in spite of being controlled by another strategy. In contrast, the 'south-to-west' path which was a valid path before the switch, becomes invalid (or *inconsistent*) after the switch. Therefore, in order to show that putative strategy-selective neurons respond specifically to strategy switches, but not to other behavioral and motor variables, one has to demonstrate that these neurons change their activity while the rat runs along the consistent path before and after the switch. This is true for strategy-selective cells in our model. In spite of the fact that on the consistent paths the simulated rats made the same turns before and after the switch, these turns were controlled by the response strategy before the switch and by the place strategy after the switch. The change in strategy is reflected by the corresponding change in the firing rate of strategy-selective cells in our model (Figure 3B).

Conversely, putative strategy-selective neurons should fire with similar rates when strategy remains the same but behavior changes (e.g. when starting from the opposite arms of the maze). This is so in our model, as demonstrated by similar rates of response-strategy cell before the switch, when starting from different starting positions (Figure 3C). The same is true for the place-strategy cell after the switch (Figure 3D).

Finally, putative strategy-selective neurons should be sensitive to changes in strategy, but not to changes in reward contingency when the strategy remains the same (as in reversals, [8]). Thus, if paths and contingency change, but strategy remains the same, as in reversals, strategy-selective neurons should fire with similar rates after the change in contingency. This is indeed so in our model, since the response-strategy cell fired with similar rates during the ‘before’ and ‘after’ phases in response reversal (Figure 3E); similarly, the place-strategy cell fired with similar rates before and after spatial reversal (Figure 3F). This is in contrast to the case when contingency, paths *and* strategy change, where neurons change significantly their rates as shown in Figure 3A (the figure is similar in the case of switch from place to response strategy, not shown).

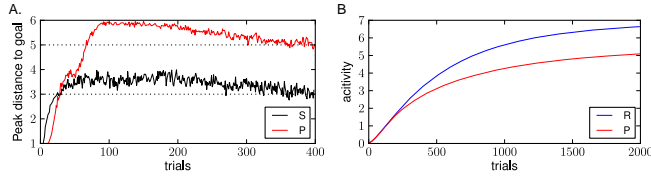
These results suggest that different strategies are indeed encoded in the activities of the strategy-selective cells in our model, in a manner resembling the strategy encoding by the rat PFC neurons [8]. It is interesting to see what predictions can be derived from this simple model. The principal feature of the TD learning algorithm that is used in our model to learn strategies and strategy selection is the reliance on the reward-prediction error  $\delta$  (Equation 3). Reinforcement learning theory predicts that in the beginning of learning the reward prediction error is maximal near the goal, while it propagates to states further from the goal with training. Schultz (1998) suggested that DA neurons in the VTA encode reward prediction error and has shown that the DA signal is maximal at the reward presentation in the beginning of training, while it is maximal at the presentation of a goal-predicting cue at the end of training [11]. We thus looked at the differences in reward propagation dynamics in different levels in our model.

In order to describe the reward propagation dynamics, we measured the evolution across trials of the mean distance between the location where  $\delta$  was maximal and the goal location for the place-strategy network and selection network during stable performance session (i.e., 400 trials with the goal always located in the same arm). According to the theory of reward signaling by DA, the distance should increase with learning and reach, by the end of training, the states farthest from the goal. In agreement with this theory, we found that the peak of reward-prediction error in the place-strategy network propagated to the starting position with training (Figure 4A, red line). In contrast, error propagation in the selection network was limited by the choice point of the plus maze (Figure 4A, black line), due to the influence of the response strategy. Thus, the general testable prediction derived from our model is that propagation of the peak DA signal in the prefrontal cortex will be limited by the choice point of the maze.

## 4 Discussion

We presented a simple model of strategy switching in the plus maze in which place and response strategies, as well as selection between those strategies, are learned using standard Q-learning TD algorithm. The architecture of the model





**Fig. 4.** A. Evolution of the distance between peak  $\delta$  location and the goal across trials for place-strategy network (red) and strategy-selection network (black). B. Evolution of Q-values of the response strategy (blue) and place strategy (red) across trials. For this simulation, we decreased the learning rate to  $\eta = 0.001$  in order to minimize the influence of noise. The starting position (south) and goal location (north) were constant for all trials.

can be directly mapped to the architecture of neuronal networks implicated in navigation. In spite of its simplicity, the model has been able to learn successfully spatial reversals, response reversals and strategy switches in the simulated plus maze task. The learning time was comparable to the learning time of real rats in behavioral experiments [8]. Moreover, activities of strategy-selective cells in the selection model were similar to those of strategy-selective biological neurons recorded from the PFC of behaving rats. In addition, a testable prediction concerning DA signaling in the PFC has been derived from our modeling results.

There are several limitations of the presented model in relation to available experimental data concerning the role of prefrontal cortex in behavior. First, it is known that strategy switches and reversals in the plus maze are mediated by different subareas of the prefrontal cortex, medial PFC (mPFC) and orbito-frontal (OFC) cortex, respectively [6, 7]. Moreover, there is experimental evidence suggesting hierarchical organization of these two subareas [14]. In our model, both switches and reversals are learned within a single network which can be considered as a combined model of mPFC/OFC. We will address the differential roles of mPFC and OFC in future versions of the model. Second, it has been shown that when rats were trained to perform multiple strategy switches, lesions to mPFC did not impair switching performance after the third switch. These data suggest that by the third switch rats acquired a mPFC-independent strategy (possibly depending on reactive or habitual mechanisms) which allowed them to immediately switch strategy when reward was not obtained in the predicted location. Our model can not explain these data, since from the point of view of the selection network all switches are identical. Finally, our model is only partially consistent with the data suggesting that place strategy is preferred early in training, while this preference shifts towards response strategy with prolonged training [4]. In Figure 4B, we plot the evolution of the Q-values for the response and place strategies (which are equal to the activities of strategy-selective cells in the selection network) across trials. The Q-value for the response strategy at the asymptote of training is higher than that for the place strategy, consistent with the preference for the response strategy after overtraining. However, this

result is the consequence of the particular model of the sensory input that we chosen for the response-strategy network and hence is not general enough to provide a valid explanation for the experimental data. A more detailed model of sensory (see, e.g. [13]) input may be required to make a stronger claim.

The architecture of the model presented here is similar to a model proposed previously [15, 16]. However, the learning algorithm is substantially simplified and is identical in the strategy learning networks and in strategy selection network, greatly simplifying the analysis of the model. In addition, we provide a comparison between strategy selective neurons in our model with new data from experimentally recorded mPFC neurons.

## References

1. Arleo, A., Rondi-Reig, L.: Multimodal sensory integration and concurrent navigation strategies for spatial cognition in real and artificial organisms. *J Integr Neurosci* **6** (2007) 327–366
2. White, N.M., McDonald, R.J.: Multiple parallel memory systems in the brain of the rat. *Neurobiol Learn Mem.* **77** (2002) 125–184
3. Packard, M.G., Hirsh, R., White, N.M.: Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: evidence for multiple memory systems. *J Neurosci* **9** (1989) 1465–72
4. Packard, M.G., McGaugh, J.L.: Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiol Learn Mem.* **65** (1996) 65–72
5. Chang, Q., Gold, P.E.: Intra-hippocampal lidocaine injections impair acquisition of a place task and facilitate acquisition of a response task in rats. *Behav Brain Res* **144** (2003) 19–24
6. Ragozzino, M.E., Detrick, S., Kesner, R.P.: Involvement of the Prelimbic-Infralimbic areas of the rodent prefrontal cortex in behavioral flexibility for place and response learning. *J Neurosci.* **19** (1999) 4585–4594
7. Rich, E.L., Shapiro, M.L.: Prelimbic/infralimbic inactivation impairs memory for multiple task switches, but not flexible selection of familiar tasks. *J Neurosci.* **27** (2007) 4747–55.
8. Rich, E.L., Shapiro, M.: Rat prefrontal cortical neurons selectively code strategy switches. *J Neurosci.* **29** (2009) 7208–19.
9. Miller, E.K., Cohen, J.D.: An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24** (2001) 167–202
10. Sutton, R., Barto, A.G.: *Reinforcement Learning - An Introduction*. MIT Press, Cambridge, MA (1998)
11. Schultz, W.: Predictive reward signal of dopamine neurons. *J Neurophysiol.* **80** (1998) 1–27
12. Arleo, A., Gerstner, W.: Spatial cognition and neuro-mimetic navigation: A model of hippocampal place cell activity. *Biol Cybern* **83** (2000) 287–299
13. Sheynikhovich, D., Chavarriaga, R., Strösslin, T., Arleo, A., Gerstner, W.: Is there a geometric module for spatial orientation? Insights from a rodent navigation model. *Psychol Rev.* **116** (2009) 540–566
14. Young, J.J., Shapiro, M.L.: Double dissociation and hierarchical organization of strategy switches and reversals in the rat PFC. *Behav Neurosci.* **123** (2009) 1028–1035

15. Chavarriaga, R., Strösslin, T., Sheynikhovich, D., Gerstner, W.: A computational model of parallel navigation systems in rodents. *Neuroinformatics* **3** (2005) 223–242
16. Dollé, L., Khamassi, M., Girard, B., Guillot, A., Chavarriaga, R.: Analyzing interactions between navigation strategies using a computational model of action selection. Volume Lecture Notes in Computer Science: Spatial Cognition VI., Springer (2008) 71–86