Why and How Hippocampal Transition Cells Can Be Used in Reinforcement Learning

Julien Hirel, Philippe Gaussier, Mathias Quoy, and Jean-Paul Banquet

Neurocybernetic team, ETIS, CNRS - ENSEA - University of Cergy-Pontoise, 95000 Cergy-Pontoise, France
julien.hirel@ensea.fr

Final draft version, accepted for publication as:

Hirel, J., Gaussier, P., Quoy, M., Banquet, J.P.: Why and how hippocampal transition cells can be used in reinforcement learning. In Doncieux, S., Girard, B., Guillot, A., Hallam, J., Meyer, J.A., Mouret, J.B., eds.: From Animals to Animats 11. Volume 6226 of Lecture Notes in Computer Science., Springer Berlin / Heidelberg (2010) 359–369 10.1007/978-3-642-15193-4_34.

The original publication is available at www.springerlink.com

Abstract. In this paper we present a model of *reinforcement learning* (RL) which can be used to solve goal-oriented navigation tasks. Our model supposes that *transitions between places* are learned in the hippocampus (CA pyramidal cells) and associated with information coming from path-integration. The RL neural network acts as a bias on these transitions to perform action selection. RL originates in the basal ganglia and matches observations of reward-based activity in dopaminergic neurons. Experiments were conducted in a simulated environment. We show that our model using transitions and inspired by Q-learning performs more efficiently than traditional actor-critic models of the basal ganglia based on temporal difference (TD) learning and using static states.

Keywords: hippocampus, basal ganglia, navigation, reinforcement learning, Q-learning

1 Introduction

In previous papers, we proposed a model in which *"place cells"* [1] are not primary located in the hippocampus proper but in the entorhinal cortex. The activity recorded in the CA pyramidal cells would not primarily originate from *"place cells"* but from *"transition cells"* coding for the transient states from one place to the next [2,3]. The reason for this proposal arose from two experimental findings. First, experimental recording of our EC artificial visual place cells displayed large place fields allowing to reach a goal without the need to store a lot of places in the environment [4]. The merging of "What" and "Where" information about surrounding landmarks was sufficient to build a robust place code

that could be simply recognized in order to build place cells. Hence the need for a dense mapping of the environment was not justified in simple sensori-motor navigation tasks. Second, we faced the impossibility to connect directly a cognitive map made of place cells and coding for multiple goals and motivations with a motor control system [2]. As a matter of fact, an homonculus was necessary to read the gradient activity on the cognitive map in order to deduce that moving in a particular direction would induce a better satisfaction than taking another direction. It was then always necessary to simulate at each time step these back and forth movements between the current place and the next possible places.

The building of a cognitive map linking transition cells suppressed this problem since one transition is always associated with a single movement. Action selection would take place in the nucleus accumbens (ACC) where planning activity coming from the cognitive map, linked to the prefrontal and/or parietal cortices, could be used as a bias to select from the current static state the most interesting transition. In our model, we used the dentate gyrus and its granular cells as a way to store past activities using a spectral timing model [5]. Area CA3 of the hippocampus received information about current and past states from the entorhinal cortex and dentate gyrus respectively. An associative memory allowed the learning of existing transitions between places. According to our model, CA3 pyramidal cells should predict the next possible transitions. Recording such cells should induce a strong spatial activity correlated with the animal place (the reason why they are called place cells) but somehow anticipating the animal next place. New neurobiological results are in agreement with such a prediction [6] but it is not sufficient to convince all the neurobiologists to move from a place cell model to a transition cell model. The cognitive map uses latent learning and constitutes an efficient system for dealing with dynamically changing environments with multiple goals. Yet there is no proof that the rat builds a cognitive map. Most of the hippocampal models used for navigation are based on place-action associations through RL and succeed to display interesting navigation performances [7,8,9,10].

In this paper, we show how the learning of transitions in the hippocampus, required by the cognitive map for complex planning tasks, can also form the perfect basis for a RL model based on Q-learning, as transitions are analogous to state/action couples. RL can easily be added to allow both backward planning with latent learning using the cognitive map and motivations, and forward planning using reinforcement hints to select the current action. Moreover the model can account for anatomical and physiological data in both the hippocampus and basal ganglia. This work is part of a project aiming at modeling the interaction between the hippocampus, the prefrontal cortex and the basal ganglia. We will show how our model can be more efficient than actor-critic models based on TD learning in tasks with several goals and motivations. Finally we will demonstrate the performances of the model in goal-oriented tasks in a simulated environment.

2 Model

In RL the environment is usually described as a Markov Decision Process (MDP). The agent can be in a certain number of states in which it can choose between a certain set of actions to perform. Experiments have been made in simulation where the agent switched between finite states based on its location in a grid world [11] or relative to prominent landmarks [9,10]. Place cells, with their *place fields* defining particular locations of the environment, can be used to characterize the state of the agent in RL [7,8].

The *Temporal Difference* (TD) learning algorithm [12] aims at maximizing the sum of expected rewards. While in TD learning an estimation of that sum is learned as a function of states, Q-learning [13] creates an estimate as a function $Q(s, a)$ of state and action. After performing action a_1 to move from state s to state s', the Q value is adjusted with the following equation:

$$Q(s, a_1) \leftarrow Q(s, a_1) + \alpha(r + \gamma \max_a Q(s', a) - Q(s, a_1)) \quad (1)$$

where r is the reward obtained when in s', α is the learning rate and γ is a discount factor. The pair (s, a_1) can also be represented as a transition $s \rightarrow s'$

The discovery of the response of dopaminergic neurons in the substantia nigra pars compacta (SNc) and the ventral tegmental area (VTA) with their modulation of the basal ganglia neuronal activity, suggested the strong involvement of these structures in RL [14]. These neurons exhibit short bursts of firing just after the occurrence of an unexpected reward and go through a short period of depression when an expected reward is not received. The similarity of this behavior with the computation of the error on the prediction of expected rewards in TD learning has lead many researchers to build models of RL associated with the basal ganglia [15,16]. In the models the computation of the TD error made in the SNc matches the neurophysiological observations of dopaminergic neurons.

How the neural differentiator used to compute the difference between subsequent predictions for the TD error signal works is subject of debate. A hypothesis is that it originates from the direct and indirect connections between the striatum and the substantia nigra pars compacta (SNc)[15]. Direct inhibitory connections and indirect excitatory connections through the subthalamic sideloop would provide the desired signal. This model supposes different timings of spike propagation in the direct and indirect pathway. It is limited because of its reliance on the internal dynamics of synapses and neurons to account for the acceptable delay between subsequent predictions. Moreover the use of the temporal characteristics of the direct/indirect pathway as the neural substrate for the TD error computation seems to be inconsistent with the known neuroanatomy [17].

In addition, several RL models use delayed synaptic learning with an eligibility mechanism [18,8,7,9]. This mechanism assumes that a memory trace of past activity is present at the synaptic level. The current reward expectation is used to modify the synaptic weights selected by the eligibility trace corresponding to the last actions. The biological plausibility of the eligibility trace remains unclear. Houk and colleagues [15] gave an hypothesis as to how this learning can

happen in real synapses. Their model involves the spiny neurons in the striosomal compartments of the striatum. The properties of a protein (CaM PK II) and a cascade of intracellular signaling mechanisms are used to account for the delay of the synaptic strengthening. However, once again, the timing of the reward is highly dependent on the properties of the internal dynamics of the neuron. It cannot account for a large variability in the delay between the action and the occurrence of the reward signal.

The need for both temporal mechanisms arises from the unavailability of the corrected reward prediction ($r + \gamma \max_a Q(s', a)$ in eq. 1) when the action is performed. This value is available in the following moments when the agent is in the new state s' and has received an optional reward r. However the previous state is no more active and cannot be directly associated with the corrected estimation of its reward expectation value.

Taking inspiration from the actor-critic model, a neural implementation of the TD learning algorithm [18], we designed a neural network model of the Q-learning algorithm (Fig. 1). The model addresses the issues discussed in the previous paragraphs by the use of a 2-step learning mechanism, suppressing the need for both an eligibility trace and specific temporal dynamics in direct/indirect pathways between the striatum and the SNc.

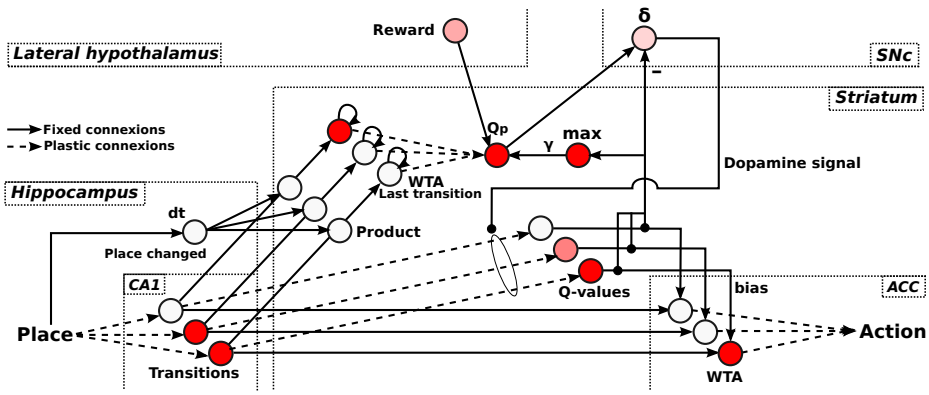

Fig. 1. Model of the Q-learning using the 2-step learning. The mapping of the various functions to cerebral structures is shown.

Step 1: A working memory in the striatum stores information about the last transition performed. When the representation of the new state is stabilized and reward predictions about available transitions arise, their maximum value is learned and associated with the value in the working memory. If a primary reward signal is received, it will also be learned. For any transition $s \rightarrow s'$ we learn to predict what the value of $t_j = r + \gamma \max_a Q(s', a)$ will be. The learning is made by a simple conditioning using the Widrow-Hoff Delta rule:

$$w_{ji} \leftarrow w_{ji} + \alpha(t_j - x_j^{Q_p}).x_i^{WTA} \quad (2)$$

where α is the learning rate. $x_j^{Q_p}$ and x_i^{WTA} are the activities of post- and pre-synaptic neurons respectively. All activities are rate-coded.

Step 2: Q values are learned in synaptic weights with transition cells as pre-synaptic neurons. Connections from the hippocampus (area CA1) to the striatum allow the propagation of transition activity to the RL system. When the agent starts to explore a new place, it begins to predict all the available transitions along with their Q values. The TD error signal, computed from the difference of current and predicted reward expectations, acts as a dopaminergic modulation of synaptic learning for transition Q-values. The learning equation used is :

$$w_{ji} \leftarrow w_{ji} + \alpha.\delta.x_i^{CA1} \quad (3)$$

where δ is the TD error signal and x_i^{CA1} the activity of the currently performed transition. Transition activity is as follows: if a transition is being performed (i.e. the agent switches from place A to B) then the only active transition is AB ($x_{AB} = 1$), otherwise if the agent explores place A (i.e. the place cell coding for A has the strongest activity) then all predicted transitions are active ($x_{AB} = 1, x_{AC} = 1$, etc.).

This system allows the simultaneous availability of the $Q(s, a_1)$ value learned in step 2 and the $r + \gamma \max_a Q(s', a)$ value learned in step 1. Hence the computation of the TD error signal does not require input pathways with different temporal properties. Only simple inhibitory and excitatory pathways are used. The trade-off for the absence of time-dependent local synaptic rules is a convergence speed for the neural network divided by 2.

The Q values for each predicted transition are used to bias the original activity of the transition cells. A WTA competition results in the optimal transition to be selected. The output of the competition is not a direct motor action but rather a motor transition, as opposed to hippocampal transitions which are perceptual. The transition then activates its corresponding learned action, which could range from complex behaviors to simple motor commands. Even though we chose to only take the optimal transition into account to select an action, secondary transitions are still predicted and provide their reward expectancies and actions as possible alternatives. In a model where actions are chosen from static states rather than transitions, a single state can correctly give a choice of actions along with their order of preference only if all the actions are coded in orthogonal patterns. If actions are coded as overlapping patterns, the connectivity of each transition with the action neurons allows the coding of well separate actions for each transition. Moreover in our case the learning of associations between actions and transitions is latent. It can happen at any time when navigating in the environment, even during an exploratory phase without any reward. Here the actions are represented by a direction to take and coded in a *neural field* [19]. Path integration information from the last place, computed from odometric input, is used to associate a direction with every transition performed. In the

model, the only actions considered in each state are based on what was learned to be possible, not a set of pre-programmed actions (e.g Go east, Go west, etc.) as it is often used in actor-critic models [7,8]. The architecture also distinguishes itself by merging the learning of state and action reward prediction into a single learning of state+action values.

The synaptic learning of predictions is modulated by particular events triggering transitory neuro-modulatory signals. The learning of Q-values through conditioning (3) happens when a transition is performed (i.e when the most active place cell changes). The learning of future predictions and rewards (2) is modulated by the delivery of the reward. A fixed time interval between place entry and reward delivery is fixed at the beginning of the experiments to allow extinction. The fixed delay is needed to provide the timing of expected rewards and produce negative reinforcement values in case an expected reward is not delivered. Future work will involve the use of a time spectrum architecture to learn reward timings and allow the delivery of rewards at any time.

3 Improvements over actor-critic models

In a simple experiment where the environment contains only one reward location, place and transition-based systems work in similar ways. In computational terms, in addition to the N place cells coding for states, the transition architecture requires the use of between $4N$ and $6N$ neurons in average to learn the transitions [3].

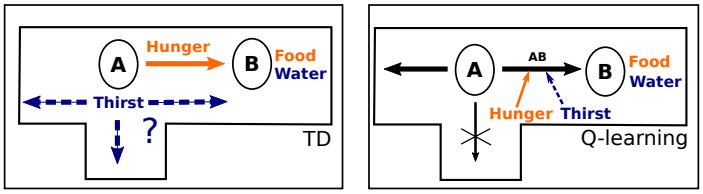

Fig. 2. Scenario with a food and water source intermittently located at place B. When the food source is found, the TD solution associates the A+hunger state with the action *Go east*. Further discovery of the water source eventually leads to the renewed slow learning of the action *Go east*, this time associated with the state A+thirst. With the transition solution, the agent learns the AB transition with the corresponding action *Go east*, independently of resource discovery. Further discovery of the food and water sources leads to the fast association of hunger and thirst with the existing AB transition. Moreover dead-end recognition could lead to a lower prediction value for the transition leading south, hence promoting the other transitions by default.

However the transition architecture shows its strength in complex tasks with multiple goals and motivations. The motivations could range from basic drives (e.g. hunger, thirst) to the need to satisfy various goals and sub-goals. Let us

consider a case where several types of resources (food, water etc.) are present in the environment. The corresponding K drives indicate the need for a particular resource. In TD learning, as a direct state-action association is created, the original model cannot learn to associate different actions to a particular state depending on the motivational context. A direct connexion from motivations to actions would indeed guide the behavior of the agent towards making always the same action when motivated, independently of the place it is in. An intermediate layer of $K * N$ neurons would need to be created to learn the association of state/drive couples with actions [7]. Actions learned in a state for one motivation would need to be learned again for other motivations event if they lead to the same direction (Fig. 2). If the action is coded as a direction vector, the learning of the movement between two place fields can take some time to converge to the vector between the two centroids (e.g. by averaging the directions taken each time to move from one place field to the other). On the other hand the Q-learning network would only need to associate the drives with existing transitions. The action associated with a transition is learned whenever the transition is made, independently from the motivational context. Transition prediction activity would however have to be initiated by the co-activation of place recognition in the hippocampus and learned drive associations. Figure 3 shows a comparison of the two architectures. The Q-learning system is the one shown in fig. 1 with a few modifications to allow multiple drives. Rewards are associated with a drive to detect different types of goals, the resulting signal is given as input to the 2-step learning RL system. In place of the modulation described in section 2, the dopaminergic neuromodulatory signal is used to modulate the learning of the Q values in synapses originating from the drive neurons. The bias used to select the next action is thus combined from current transition and drive activities.

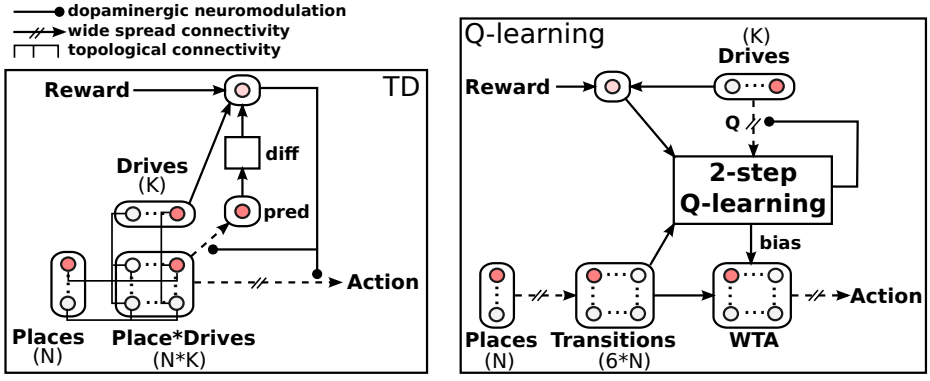

Fig. 3. Comparison of neural implementations of TD and Q-learning for a multiple drives scenario. When the number of places and drives increases, the transition based solution becomes less and less expensive as compared to the TD solution.

In our model, the number of neurons needed to encode transitions, states and actions is independent from the number of motivations. Due to the 2-step learning, a lot more neurons are needed for simple tasks with few motivations than in simple actor-critic networks. However these neurons can work with any number of motivations. The transitions model can use a direct bias of every new motivation on transition activity whereas actor-critic models would have to add extra place/drive neurons. In complex tasks with many goals and sub-goals this could lead to significant improvements in information compression, meaning more ecologically viable architectures. The trade-off is the need of wide-spread connectivity between places, drives, transitions and actions.

4 Experiments

The neural network has been tested in a simulated environment using the *Promethe* NN simulator [20]. The simulated environment is an open square environment with 20 perfectly identifiable landmarks equally spaced along the walls to simulate visual input. One food source is placed in the upper left corner. The speed of the agent is constant throughout the experiment except when avoiding walls. The passing of time in the simulation is discretized into a series of time steps. However the functioning of the architecture is not dependent on the fineness of this discretization. The simulation works with any time step (e.g. 50ms, 100ms, 500ms etc.), however too large time steps would lead to the agent "teleporting" itself and missing sensory input on the way, leading to a less reactive system and decreased performances. The results were obtained using 100ms time steps.

First the agent performs an exploratory session in order to map its environment. During this phase of the experiment, navigation is guided by a random exploration strategy. The direction of the agent is periodically changed, based on a Gaussian probability function centered on the current direction. Simulated ultrasound obstacle detection allows the agent to avoid hitting the walls. Place cells are learned based on a minimum activity threshold. As the agent moves from place to place, transitions between place cells are learned and associated with a direction. During this random exploratory phase, the agent is able to discover the food source and build its representation of optimal paths using transitions and RL (Fig. 4).

During the second phase of the experiment, the exploratory/exploitation phases are modulated by an internal motivational signal. When motivated, the agent will use the learned transition bias and corresponding actions to reach the food source. The delivery of the food reward then inhibits the motivation signal and an exploratory phase begins. The motivation is triggered again when the agent reaches an area comprising the eastern and southern extremities of the environment. A good level of performance in this task requires the ability to quickly reach the goal location from any starting position in this area.

Figure 4 shows example trajectories of motivated navigation using RL. As the agent follows the path given by a single winning transition, the trajectories roughly follow the edges of the transition graph and are thus not straight lines

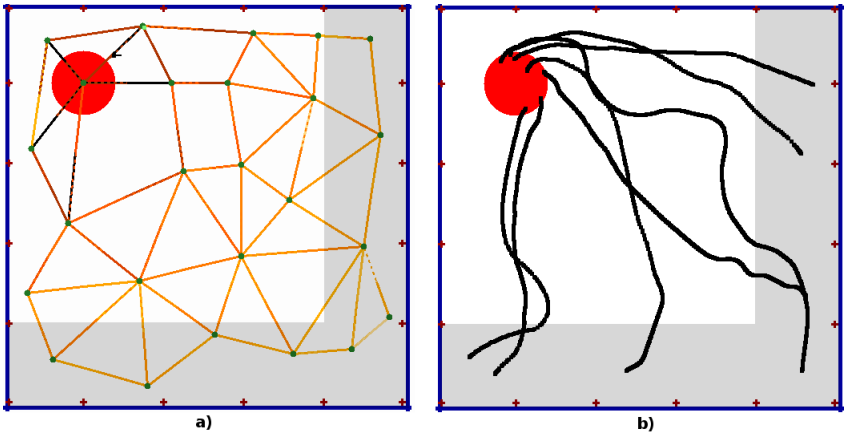

Fig. 4. a) Graph of all learned transitions in the simulated environment. Darker colors mean higher Q values for the corresponding transitions. **b)** Trajectories taken by the agent during goal-directed navigation. The goal location is represented by a disk in the upper left corner of the environment. All starting points for goal-oriented navigation trials are located in the gray area.

to the goal. Smoother trajectories could be obtained using a soft competition when selecting transitions and their associated actions. Mean escape latencies and standard deviations are given in table 4 for both the transition Q-learning and a random exploration strategy. They express the time needed by the agent to reach the goal when motivated, with starting points randomly spread in the motivation trigger area. The transition Q-learning architecture performs 3 times better than random exploration with obstacle avoidance. A soft competition for transition and action selection could be used to further increase the performances of the algorithm.

Table 1. Mean escape latency and standard deviation in seconds for transition Q-learning and random exploration. The values are computed from a set of 50 trials for each strategy. The parameters of the simulation are : learning rate $\alpha = 0.5$, discount factor $\gamma = 0.8$, reward value $r = 1$.

	Mean	Standard deviation
Transition Q-learning	36.7	14.5
Random exploration	115.1	90.1

5 Discussion

In addition to being consistent with neurobiological observations [6], the transition learning architecture could serve as a basis for several navigation strategies. The prediction of available transitions at any given time provides the system with a repertoire of possible actions. The transition-action association is learned autonomously and is dissociated from navigation strategies such as path planning or RL. As opposed to usual actor-critic models of TD learning where the motor action is the output of the RL network, a Q-learning based model can work with transitions as its sole representation of the environment and be more efficient in complex scenarios.

By using transitions as a common representation, one can easily integrate several navigation strategies in the same architecture. We previously used a cognitive map to solve navigation tasks. This strategy also provided a bias to the competition between predicted transitions. The competition leading to the selection of the next action can accept several such biases, given by different strategies working in parallel. The parallel use of the cognitive map planning and RL will bring to light the advantages/disadvantages of one system over the other and show the way for an integrated architecture with the 2 cooperating systems. More transition-based strategies, such as timed sequences of actions, could eventually be added. In this case transitions would have to be able to learn both spatial and temporal properties. Future work will involve the implementation of a system capable of modulating these concurrent strategies. The modulation could be based on a performance criterion, thus selecting the best strategy for a particular task. Internal signals could also be monitored by a meta-controller capable of detecting whether a strategy is dysfunctional or not.

We have recently built an architecture which used transitions with both spatial and temporal components to build a cognitive map and solve planning tasks involving navigation and the precise timing of particular actions. The integration of timed transitions into the present RL model would help reproduce the precise time-dependent prediction capabilities of dopaminergic neurons in the basal ganglia. This is necessary to be able to select an appropriate behavior depending on the timing of a reward. One particular case in which we are interested is the autonomous learning of a precisely timed waiting period requiring movement inhibition from the animat.

Acknowledgments This work is supported by the CNRS, as part of a PEPS project on neuroinformatics, and the DGA. We thank B. Poucet, S. Wiener and E. Save for useful discussions.

References

1. O'Keefe, J., Dostrovsky, J.: The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. Brain Res **34**(1) (Nov 1971) 171–175

2. Gaussier, P., Revel, A., Banquet, J.P., Babeau, V.: From view cells and place cells to cognitive map learning: processing stages of the hippocampal system. Biol Cybern **86**(1) (Jan 2002) 15–28
3. Cuperlier, N., Quoy, M., Gaussier, P.: Neurobiologically inspired mobile robot navigation and planning. Front Neurorobotics **1** (2007) 3
4. Giovannangeli, C., Gaussier, P., Banquet, J.P.: Robustness of visual place cells in dynamic indoor and outdoor environment. International Journal of Advanced Robotic Systems **3**(2) (jun 2006) 115–124
5. Grossberg, S., Merrill, J.W.: A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. Brain Res Cogn Brain Res **1**(1) (Jun 1992) 3–38
6. Alvernhe, A., Cauter, T.V., Save, E., Poucet, B.: Different ca1 and ca3 representations of novel routes in a shortcut situation. J Neurosci **28**(29) (Jul 2008) 7324–7333
7. Arleo, A., Gerstner, W.: Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. Biol Cybern **83**(3) (Sep 2000) 287–299
8. Foster, D.J., Morris, R.G., Dayan, P.: A model of hippocampally dependent navigation, using the temporal difference learning rule. Hippocampus **10**(1) (2000) 1–16
9. Khamassi, M., Lachèze, L., Girard, B., Berthoz, A., Guillot, A.: Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. Adaptive Behavior **13**(2) (2005) 131–148
10. Mannella, F., Baldassarre, G.: A neural-network reinforcement-learning model of domestic chicks that learn to localize the centre of closed arenas. Philos Trans R Soc Lond B Biol Sci **362**(1479) (Mar 2007) 383–401
11. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. Journal Of Artificial Intelligence Research **4** (1996) 237–285
12. Sutton, R.S.: Learning to predict by the methods of temporal differences. Machine Learning **3** (1988) 9–44
13. Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning **8**(3) (May 1992) 279–292
14. Schultz, W.: Predictive reward signal of dopamine neurons. J Neurophysiol **80**(1) (Jul 1998) 1–27
15. Houk, J.C., Adams, J.L., Barto, A.G.: A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Models of Information Processing in the Basal Ganglia, MIT Press (1995) 215–232
16. Montague, P.R., Dayan, P., Sejnowski, T.J.: A framework for mesencephalic dopamine systems based on predictive hebbian learning. J Neurosci **16**(5) (Mar 1996) 1936–1947
17. Joel, D., Niv, Y., Ruppin, E.: Actor-critic models of the basal ganglia: new anatomical and computational perspectives. Neural Netw **15**(4-6) (2002) 535–547
18. Barto, A.G.: Adaptive critics and the basal ganglia. In: Models of Information Processing in the Basal Ganglia, MIT Press (1995) 215–232
19. Amari, S.I.: Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics **27**(2) (1977) 77–87
20. Lagarde, M., Andry, P., Gaussier, P.: Distributed real time neural networks in interactive complex systems. In: CSTST '08, New York, NY, USA, ACM (2008) 95–100