HIPPOCAMPUS 10:1–16 (2000)

A Model of Hippocampally Dependent Navigation, Using the Temporal Difference Learning Rule

D.J. Foster,[1,2]* R.G.M. Morris,[1] and Peter Dayan[2]

[1]*Centre for Neuroscience, University of Edinburgh, Edinburgh, Scotland, UK*
[2]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts*

ABSTRACT: This paper presents a model of how hippocampal place cells might be used for spatial navigation in two watermaze tasks: the standard reference memory task and a delayed matching-to-place task. In the reference memory task, the escape platform occupies a single location and rats gradually learn relatively direct paths to the goal over the course of days, in each of which they perform a fixed number of trials. In the delayed matching-to-place task, the escape platform occupies a novel location on each day, and rats gradually acquire one-trial learning, i.e., direct paths on the second trial of each day. The model uses a local, incremental, and statistically efficient connectionist algorithm called temporal difference learning in two distinct components. The first is a reinforcement-based "actor-critic" network that is a general model of classical and instrumental conditioning. In this case, it is applied to navigation, using place cells to provide information about state. By itself, the actor-critic can learn the reference memory task, but this learning is inflexible to changes to the platform location. We argue that one-trial learning in the delayed matching-to-place task demands a goal-independent representation of space. This is provided by the second component of the model: a network that uses temporal difference learning and self-motion information to acquire consistent spatial coordinates in the environment. Each component of the model is necessary at a different stage of the task; the actor-critic provides a way of transferring control to the component that performs best. The model successfully captures gradual acquisition in both tasks, and, in particular, the ultimate development of one-trial learning in the delayed matching-to-place task. Place cells report a form of stable, allocentric information that is well-suited to the various kinds of learning in the model. *Hippocampus 2000;10:1–16.*

KEY WORDS: hippocampus; place cells; spatial learning; temporal difference learning; navigation

INTRODUCTION

There is an apparent discrepancy in the rodent hippocampal literature, between the putative involvement of hippocampal principal neurons in navigation, and the limited navigational correlates of neuronal activity actually observed during electrophysiological recording from these neurons. Consider a hippocampal place cell, so called because it fires when the animal occupies a restricted portion of an environment, known as its place field (O'Keefe and Dostrovsky, 1971; O'Keefe and Nadel, 1978; Wilson and McNaughton, 1993). The cell's spatial tuning suggests a role in spatial learning, in agreement with hippocampal lesion studies (Morris et al., 1982; Sutherland et al., 1983; Barnes, 1979). However, the activity of the cell, or even of a collection of such cells, simply individuates different locations—it does not directly tell the animal where it is, or where it ought to go. More complex navigational activity has been suggested, such as the replaying of long-range navigational sequences or possible navigational routes (e.g., Levy, 1996), but such activity has not been observed across more than only a few place cells at a time, and is not known to be predictive in nature (O'Keefe and Recce, 1993; Skaggs et al., 1996).

The paradox is this: how can place cells be important for navigation, but at the same time not embody all the spatial information required to navigate from one place to another? To put it another way, what might be the use for navigation of a group of cells whose firing simply divides up an environment into place fields, rather than computing specific spatial quantities like distance or direction to a goal? Two types of model of rodent navigation have attempted to make use of place cells while not assuming hitherto unobserved properties of them. However, each of these types of model has encountered a fundamental problem.

One type of model assumes that place cells provide the ideal representation for reward-based learning. Thus, when a rat encounters a goal such as hidden food in an environment, some kind of reinforcement signal enables the place cell firing near the goal to be associated with the actions which the rat took immediately prior to

Grant sponsor: McDonnell-Pew Foundation; Grant sponsor: Edinburgh University; Grant sponsor: MRC; Grant sponsor: NSF; Grant number: IBN-9634339; Grant sponsor: Surdna Foundation; Grant sponsor: Gatsby Charitable Foundation; Grant sponsor: University of Oxford McDonnell-Pew Center for Cognitive Neuroscience.
Peter Dayan's current address is: Gatsby Computational Neuroscience Unit, Alexandra House, 17 Queen Square, London WC1N 3AR, UK.
*Correspondence to: D.J. Foster, Department of Brain and Cognitive Sciences, E25-210, Massachusetts Institute of Technology, Cambridge, MA 02139.
Accepted for publication 1 September 1999

attaining the goal (Burgess et al., 1994; Brown and Sharp, 1995). The problem with this hypothesis is that there is no simple way of dealing with the fact that most locations within an environment are typically very far from the goal location, compared to the length of a place field. The rat will have no direct information as to the direction in which to move if it starts at a location far from the goal where none of the place cells associated with appropriate actions is active. We call this the *distal reward* problem. It has been dealt with in different ways, e.g., by postulating very large place fields covering the entire environment, although these are rarely observed (Burgess et al., 1994), or by making use of a memory trace, for which there is no evidence over the kinds of distances required, and which in any case leads to a rather inefficient learning algorithm (Brown and Sharp, 1995).

The second type of model assumes that place cells become associated with metric coordinates for locations within environments (Wan et al., 1994; Redish and Touretzky, 1997; Blum and Abbott, 1996; Gerstner and Abbott, 1996). A natural basis for learning the coordinates in the first place is the self-motion (or "dead reckoning") information which an animal has available. The problem with this hypothesis is that self-motion information, while suitably metric, is only *relative* in nature. Simply performing path integration on this information runs into trouble as soon as the animal loses track of its origin, as must happen during laboratory navigation tasks in which an animal is often picked up from the goal location at the end of one trial, and started again from an unpredictable starting location. If the animal path-integrates from each new starting position, it will quickly acquire inconsistent coordinates over the environment as a whole. We call this the problem of *global consistency*.

The motivation for the present work is the observation that a recently developed neural network learning rule, temporal difference (TD) learning (Sutton, 1988), can solve both the distal reward problem and the global consistency problem. Following Dayan (1991), this paper investigates a model of spatial learning in two navigational tasks, combining TD learning with a place cell representation to learn about rewards and coordinates. We pose the question: can TD learning bridge the computational gap between the observed activity of place cells and the goal-directed navigational behavior for which place cells are thought to be important?

TEMPORAL DIFFERENCE LEARNING

Temporal difference (TD) learning (Sutton, 1988; Barto et al., 1990; Bertekas and Tsitsiklis, 1996; Sutton and Barto, 1998) is a form of error-driven learning used in feed-forward neural networks in which input patterns (e.g., patterns of place cell activity) are to be associated with output values (e.g., an expectation of how close the goal is), but where additionally there is information to be had in the sequence in which input patterns and output values present themselves.

Conventional error-driven learning rules (such as backpropagation) are usually referred to as "supervised" because they use an error based on the difference between the network output and a desired, or "teaching" value. As a consequence, these learning rules require a teaching value all the time. TD learning, by contrast, uses an error based on the *difference* between successively occurring output values—a sensible strategy when a consistent relationship between these values is expected. For example, in this paper we consider the problem of a rat trying to learn an expectation of reward that increases smoothly as it follows a path towards a goal. Irrespective of where the goal is, there should be a certain temporal gradient, or "temporal difference," between values of this expectation at successive locations. TD learning uses the reward information directly available at the goal to learn where the greatest expectation of reward should be, but also uses the temporal gradient information to learn appropriate expectations everywhere else.

The use of local consistency information makes TD learning considerably more efficient than supervised learning alternatives. Consider one such alternative, the trace memory learning rule in Brown and Sharp (1995). A place cell very far from the goal can learn an expectation of reward simply by maintaining a trace memory of its activation which decays so slowly that when the animal gets to the goal, a residual trace will remain. However, this is inefficient because an animal's paths will be extremely variable during learning: early on in training, the animal sometimes gets to the goal quickly, and sometimes not, and the residual value of a particular place cell's trace will likewise be extremely variable from trial to trial. Unfortunately, it is these residual values that the learning rule of Brown and Sharp (1995) must average over. By contrast, TD learning considers a generally less variable quantity, the difference between successive estimates of the quantity being learned. Because of this, TD learning both converges faster and produces better predictions than supervised learning (Sutton, 1988).

TD learning has provided a simple but powerful model of associative learning in classical conditioning, effectively extending the Rescorla-Wagner rule to the temporal domain (Sutton and Barto, 1987). In particular, TD provides an explanation for second-order conditioning, whereby a conditioned stimulus (CS) that has acquired predictive value can itself condition another preceding CS. This process is similar to the dissemination of reward information through an environment in a TD model of navigation: place cells, acting as CSs, can become predictive of reward, even when they are not *directly* followed by reward, but instead followed by other CSs, which have themselves become predictive of reward. Furthermore, a neural basis for the involvement of TD learning in classical and instrumental conditioning was recently proposed, following the discovery in the primate ventral tegmental area of neurons whose firing during conditioning tasks is consistent with the main error term which forms the basis for TD learning (Schultz et al., 1997; Montague et al., 1996).

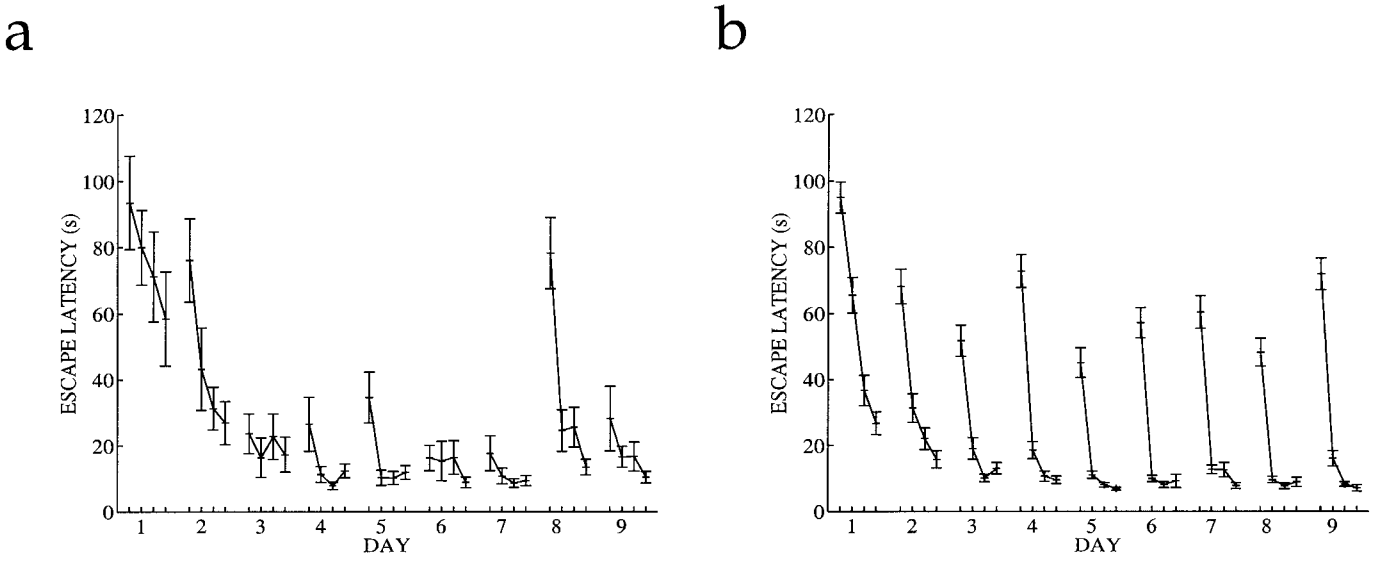

FIGURE 1. Performance of rats on (a) reference memory (RMW), N = 12, and (b) delayed matching-to-place (DMP), N = 62. For both, escape latency (time taken to reach platform) is plotted across days (RMW task: 4 trials/day, fixed platform location, days 1–7; reversal to new platform location, days 8–9; DMP task: 4 trials/day, new platform location each day). Note 1) asymptotic performance in RMW task, 2) one-trial learning in DMP task, and 3) difference in escape latency on second trial of day 8, between the two tasks. Trial 1 performance differs from day to day, due to platform position. It was observed that platforms nearer the center of the pool, or near to a starting position, were easier to find under random search than others. (b) is from Steele and Morris (1999); data for (a) were obtained in the same apparatus and using the same methods as those described for the DMP task by Steele and Morris (1999), with permission, except that: 1) the platform remained in the same location across days, until moved to the opposite quadrant on day 8; and 2) the intertrial interval was always 15 s.

THE BEHAVIOR TO BE MODELED

We have chosen to model two behavioral tasks that are highly sensitive to hippocampal lesions and represent the kind of navigational problems for which TD learning, used in conjunction with place cells, might provide a solution.

Reference memory in the watermaze (RMW) involves placing rats into a circular tank of water in which there is a hidden escape platform towards which they are highly motivated to swim (Morris, 1981). The tank itself affords no local cues as to the position of the platform, and the use of a different starting location on each trial renders useless the strategy of replaying a series of motor commands that worked previously. The rats must learn to navigate to the platform location from any possible starting position. Normal rats show more or less direct paths to the platform after 20 trials, as implied by their short escape latencies (Fig. 1a, days 5–7). If the platform is then moved to a new location, performance is disrupted and animals take several trials before they show direct paths to the new platform location (Fig. 1a, days 8 and 9).

RMW has been modelled as an instance of conventional reward-based learning using place cells (Brown and Sharp, 1995). However, the task presents a distal reward problem. We examine a simple TD-learning based "actor-critic" model of learning (Barto et al., 1983, 1990), in which a set of place cells is associated with a representation of reward expectation, and also with a representation of action choice. Critically, the TD learning rule is used to predict rewards.

Delayed matching-to-place (DMP) is a new protocol for the watermaze (Steele and Morris, 1999), though similar tasks have been explored (Morris, 1983; Panakhova et al., 1984; Whishaw, 1985, 1991). As in RMW, rats are given several trials per day with a platform that stays in the same location throughout the day. The critical difference is that the platform is at a new and different location on each day. Within each of the first few days, normal rats show a gradual decrease in the time taken to reach the platform (see Fig. 1b, days 1–5). A different pattern of escape latencies emerges by about day 6. Rats by then show "one-trial learning," i.e., near-asymptotic navigational performance on the second trial of the day to a novel platform position.

DMP is computationally more demanding than RMW. Unlike RMW, this task involves altering actions after only one trial of experience. It does not, however, only involve rapid learning, as is demanded in a standard delayed match-to-sample task. DMP in the watermaze is a complex navigation task in which a whole sequence of navigational actions has to be inferred from the single learning experience. This suggests that rats learn a representation of space that is goal-independent, which we model as a metric coordinate system, learned from self-motion information. However, previous attempts at modeling coordinate learning using self-motion information encountered a global consistency problem (Wan et al., 1994; Redish and Touretzky, 1997). In our

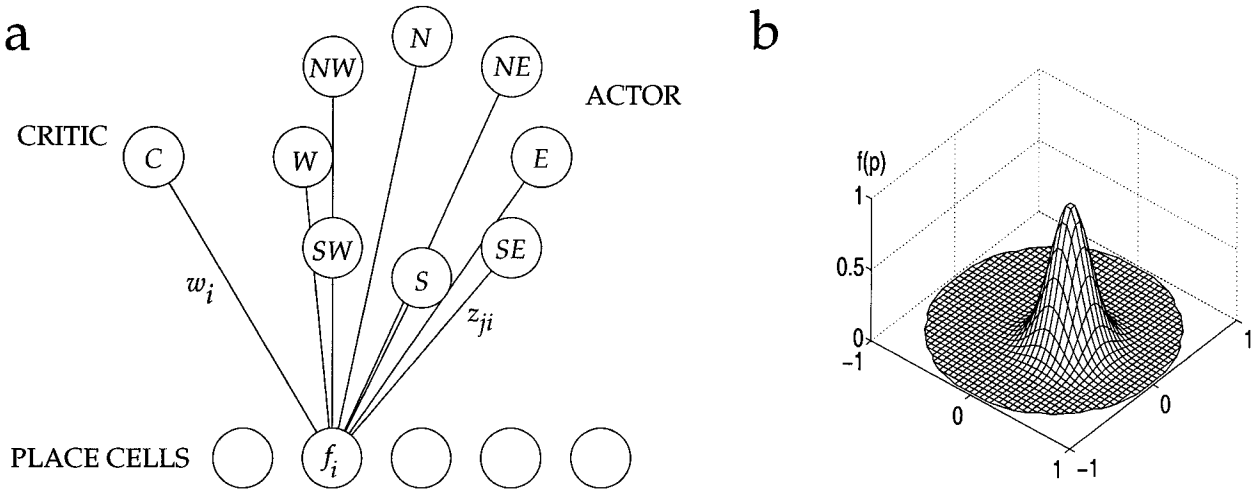

FIGURE 2. The actor-critic system. a: An input layer of place cells projects to the critic cell, C, whose output is used to evaluate behavior. Place cells also project to eight action cells, which the actor uses to select between eight possible directions of movement from any given location. b: An example of a Gaussian place field (x and y axes represent location, z axis represents firing rate).

model, TD learning is used, in association with a stable place cell representation, to develop consistent coordinates directly.

The paper begins by presenting the reward-based component of the model, demonstrating that this component alone captures some aspects of spatial learning, but not all. In particular, it does not capture the flexible way in which rats can learn about novel goal locations. The second component of the model, the learned coordinate system, is then described, along with a simple way in which the components can be made to work together. Simulation results are presented which capture performance in both RMW and DMP tasks. The discussion addresses the role of place cells within the model, what can be inferred from the model about the nature of the two tasks, and the relationship of the model to experimental data and to other models of hippocampal function. Finally, a set of novel experimental predictions is presented.

REWARD-BASED NAVIGATION

Consider a simulated animal in an environment with control of its own actions. At any given time *t*, the animal is able to choose an action. Also at any given time *t*, the environment provides the animal with a reward R_t. If the animal moves onto the platform (a certain region of the environment) at time *t*, $R_t = 1$; otherwise $R_t = 0$. The difficult problem is to learn correct actions given such a sparse reward signal.

To solve this problem we use an "actor-critic" architecture. A computational unit called the *actor* continually produces actions, taking a simulated animal around an environment. While it does so, a second computational unit called the *critic* continually criticizes the actions taken. The actor adapts its action choices using the critic's information. The critic also adapts in the light of the changing actor. The critic's role is as a go-between, between the actions on one hand, and the reward information on the other, the latter being too sparse and uninformative to criticize the actor directly.

Our implementation of the actor-critic has three parts (Fig. 2a): 1) an input layer of *place cells*, 2) a *critic* network that learns appropriate weights from the place cells to enable it to output information about the value of particular locations, and 3) an *actor* network that learns appropriate weights from the place cells which enable it to represent the direction in which the rat should swim at particular locations.

Hippocampal Place Cells

Following experimental data (O'Keefe and Burgess, 1996), the activities of place cells are modelled as Gaussian functions of location in the maze (Fig. 2b). If the rat is at position p, then the activity of place cell $i = 1 \ldots N$ is given by:

$$f_i(\mathrm{p}) = \exp\left(-\frac{\|\mathrm{p} - \mathrm{s}_i\|^2}{2\sigma^2}\right) \quad (1)$$

where s_i is the location in space of the center of cell *i*'s place field, and σ is the breadth of the field, equivalent to the radius of the circular contour where firing is 61% of the maximal firing rate. We consider an ensemble of place cells ($N = 493$) with place fields distributed in an overlapping manner throughout the maze, each with width $\sigma = 0.16$ m.

Although clearly idealized, these place cells illustrate the limitations pointed out in the Introduction: they are not intrinsically informative about spatial or navigational quantities such as distance or direction from a distant goal. However, such units form a basis function representation (e.g., Poggio and Girosi, 1990) of location. As such, they would support the representation and learning of functions which vary (usually smoothly) with location. This paper explores this hypothesis, that hippocampal place cells play the limited but nonetheless critical role of providing a particular representational substrate.

The Critic

The critic has a single output cell, whose firing rate at a location p is given by a weighted sum of the firing rates of place cell inputs $f_i(\mathrm{p})$:

$$C(\mathrm{p}) = \sum_i w_i f_i(\mathrm{p}) \tag{2}$$

where w_i is the weight from place cell i.

The standard approach is for the critic to attempt to learn what is called a value function over location, $V(\mathrm{p})$, which is really an evaluation of the actions currently specified by the actor. The value function is usually defined as, for any location p, the discounted total future reward that is expected, on average, to accrue after occupying location p and then following the actions currently specified by the actor. If p_t is the location at time t, we may define the value as:

$$V(\mathrm{p}_t) = \langle R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots \rangle \tag{3}$$

where γ is a constant discounting factor, set such that $0 < \gamma < 1$, and $\langle \cdot \rangle$ denotes the mean over all trials. Three features can be noted about this quantity. First, if we call the time at which each watermaze trial ends T (noting that the value of T will vary from trial to trial), then because this is the only time at which there is any reward, the value simplifies to:

$$V(\mathrm{p}_t) = \langle \gamma^{T-t} \rangle. \tag{4}$$

Second, because the constant discounting factor γ is set such that $0 < \gamma < 1$, $V(\mathrm{p}_t)$ is a monotonic measure of the average time it takes to get to the platform from p. Third, $V(\mathrm{p}_t)$ can actually suggest improvements to the actions of the actor, since an action which leads to a large increase in value is guaranteed to take the animal closer to the platform. Therefore, a good strategy for the actor is to try several actions at each location, with the aim of choosing the action which involves the largest increase in value.

However, the value function is not given; the critic must learn it using TD learning, i.e., the weights w_i must be adapted so that $C(\mathrm{p}) = V(\mathrm{p})$. TD works by enforcing *consistency* between successive critic outputs. Specifically, from Equation 3, the following relationship holds between successively occurring values, $V(\mathrm{p}_t)$ and $V(\mathrm{p}_{t+1})$:

$$V(\mathrm{p}_t) = \langle R_t \rangle + \gamma V(\mathrm{p}_{t+1}). \tag{5}$$

If it were true that $C(\mathrm{p}) = V(\mathrm{p})$, then a similar relationship *should* hold between successively occurring critic outputs, $C(\mathrm{p}_t)$ and $C(\mathrm{p}_{t+1})$:

$$C(\mathrm{p}_t) = \langle R_t \rangle + \gamma C(\mathrm{p}_{t+1}). \tag{6}$$

TD uses the *actual* difference between the two sides of equation 6 as a *prediction error*, δ_t, which drives learning:

$$\delta_t = R_t + \gamma C(\mathrm{p}_{t+1}) - C(\mathrm{p}_t) \tag{7}$$

using the instantaneous sample R_t in place of the desired average value $\langle R_t \rangle$ which is, of course, unavailable. Note that the above equation is more complex than it need be; in fact, R_t and $C(\mathrm{p}_{t+1})$ ought never to be both nonzero, since $R_t = 1$ only on the platform, and at this point a trial ends, so $V(\mathrm{p}_{t+1}) = 0$. We therefore enforce this condition by making δ_t include either one term or the other, but never both. TD reduces the error by changing the weights w_i from those place cells that are active:

$$\Delta w_i \propto \delta_t f_i(\mathrm{p}_t). \tag{8}$$

Under various conditions on the learning rate and on the representation provided by the place cells, this rule is bound to make $C(\mathrm{p})$ converge to the value function $V(\mathrm{p})$ as required. Following standard reinforcement learning practice, we use a fixed learning rate to avoid slow learning. The price to be paid is residual error; however, the results show that this error is insignificant.

The Actor

The actor is shown in Figure 2a. For convenience, the rat is allowed to move in one of eight possible directions at each time step (e.g., north, northeast, east), and so the actor makes use of eight action cells a_j, $j = 1 \ldots 8$. Just as in Equation 2, at position p the activity of each action cell is

$$a_j(\mathrm{p}) = \sum_i z_{ji} f_i(\mathrm{p})$$

where z_{ij} is the weight from place cell i to action cell j. This activity is interpreted as the relative preference for swimming in the jth direction at location p: the actual swimming direction is chosen stochastically, with probabilities P_j related to these activities by:

$$P_j = \frac{\exp(2a_j)}{\sum_k \exp(2a_k)} \tag{9}$$

Following the logic described above, the actor should try various actions at each location, with the aim of choosing an action which produces the greatest increase in value. The stochastic action choice ensures that many different actions are tried at similar locations. To choose the best action, a signal is required from the critic about the change in value that results from taking an action. It turns out that an appropriate signal is the same prediction error δ_t used in the learning of the value function. For example, consider what happens when appropriate values have been learned that are consistent with the actions specified by the actor throughout the environment, i.e., when $C(\mathrm{p}) = V(\mathrm{p})$. At this point, the currently specified actions should produce, on average, zero prediction error, i.e., $\delta_t = 0$. However, other actions will produce, on average, nonzero δ_t. In particular, if $\delta_t > 0$, i.e., $V(\mathrm{p}_t) < \gamma V(\mathrm{p}_{t+1})$, then the new action is a better one. If $\delta_t < 0$, the new action is a worse one.

The actor weights z_{ji} are adapted according to:

$$\Delta z_{ji} \propto \delta_t f_i(\mathrm{p}_t) g_j(t) \tag{10}$$

where $g_j(t) = 1$ if action j was chosen at time t, and $g_j(t) = 0$ otherwise. This is a form of Hebbian learning modified by δ_t: the connection between a place cell and an action cell is strengthened if (1) they fire together, and (2) if what resulted from taking that action at that place was an improvement in value. Likewise, the

connection between a place cell and an action cell is weakened if (1) they fire together, and (2) what resulted from taking that action at that place was that the value got worse.

Learning Actor and Critic Simultaneously

So far, two separate mechanisms have been described. First, a critic can develop a value function, which serves as an evaluation of the current actions of the animal. However, the method was presented as if the actor was constant, i.e., as if the specified actions did not change. Second, an actor can use the critic's value function to improve the actions it specifies. However, this was presented as if the value function was correct for the current actions of the actor. Given that both mechanisms must work together, it has been suggested that learning in the actor should proceed much more slowly than in the critic (Witten, 1977).

In fact, the scheme is robust enough for learning to proceed quickly in both actor and critic; thus, the actor is being criticized by a critic which has not necessarily completely learned the appropriate value function. The reason this "bootstrapping" can work is because learning in the critic is characterized by what might be called "graceful improvement:" even when poorly learned, the critic's value function can lead to improvements in the actor, e.g., near the platform.

Theoretical guarantees are not available for this joint learning of the actor and the critic (though they are, for closely related algorithms). However, there is quite extensive empirical evidence, in addition to the results we present here, showing that it works well (Barto et al., 1990).

PERFORMANCE OF REWARD-BASED NAVIGATION

Simulation Procedures

We simulated the swimming behavior of a rat in a 2-m-diameter circular watermaze, which contained a 0.1-m-diameter escape platform. These parameters are the same as those in Steele and Morris (1999). The swimming speed of the rat was constant, at 0.3 ms^{-1}. The walls were treated as reflecting boundaries: the rat "bounced" off. Any move into the platform area was counted as a move onto the platform. Space was treated as a continuous variable; however, time was discretized into steps of 0.1 s. Simulations with 0.01-s bins produced similar results to those with the coarser discretization, and so show that this discretization does not produce artifacts.

In reality, a rat cannot choose a different direction at the fine-grained time steps of the temporally discrete simulation. To model momentum, the direction the rat heads was given by a mixture of control as specified by the actor, and the previous heading, in the ratio 1:3. This restricts the turning curve of the rat, and is particularly important early on, when the whole pool must be searched fairly quickly. One technical concern about momentum is that it means that the path to the goal from a location is partly determined by the direction in which it was swimming when it arrived at that location. This disturbs the formal theory, although simulations demonstrate that it does not prevent good performance by the simulated rats.

Following the experimental protocols, each trial began at one of four starting locations located at the north, south, east, and west edges of the pool, and ended when either the rat reached the platform, or a time-out of 120 s was reached. For RMW, the platform remained in the same location throughout the simulation. In DMP, the platform was moved to a novel location after every four trials.

The learning rate parameters, which determine the constants of proportionality in Equations 8 and 10, were optimized.

Simulation Results

Figure 3 shows the gradual development of the value function. For the first few trials, it is informative about only a small area close the platform location. Later in learning, however, values have spread out to all parts of the environment. This enables appropriate actions to be learned, as reflected in ever shorter paths to the platform.

The actor-critic model of Figure 2 was first applied to the reference memory (RMW) task. Figure 4a shows that the actor-critic captures learning in this task; path lengths reach asymptotically low values as quickly as the latencies of rats shown in Figure 1a. However, when the platform is moved during the reversal phase of days 8 and 9, this model diverges from the performance of rats.

Likewise, when applied to the delayed matching-to-place (DMP) task, the results are strikingly different. Figure 4b demonstrates that the actor-critic component of the model fails by itself to capture the performance of rats in DMP, because the value function that is learned confounds spatial and reward information, and so neither the value function nor the policy are flexible to changes in reward location. The model incorrectly predicts that learning a new platform position is much slower because of interference from previous days.

COORDINATE-BASED NAVIGATION

Learning Globally Consistent Coordinates From Self-Motion Information

The actor-critic is a general solution to the problem of navigating to a fixed goal location. Nothing is assumed about the shape or topology of the environment, and short paths to the goal would ultimately be learned even in the presence of complicated barriers. However, the actor-critic model fails by itself to capture the performance of rats in DMP for two reasons. First, it incorrectly predicts that learning a new platform position is much slower because of interference from previous days. Second, it

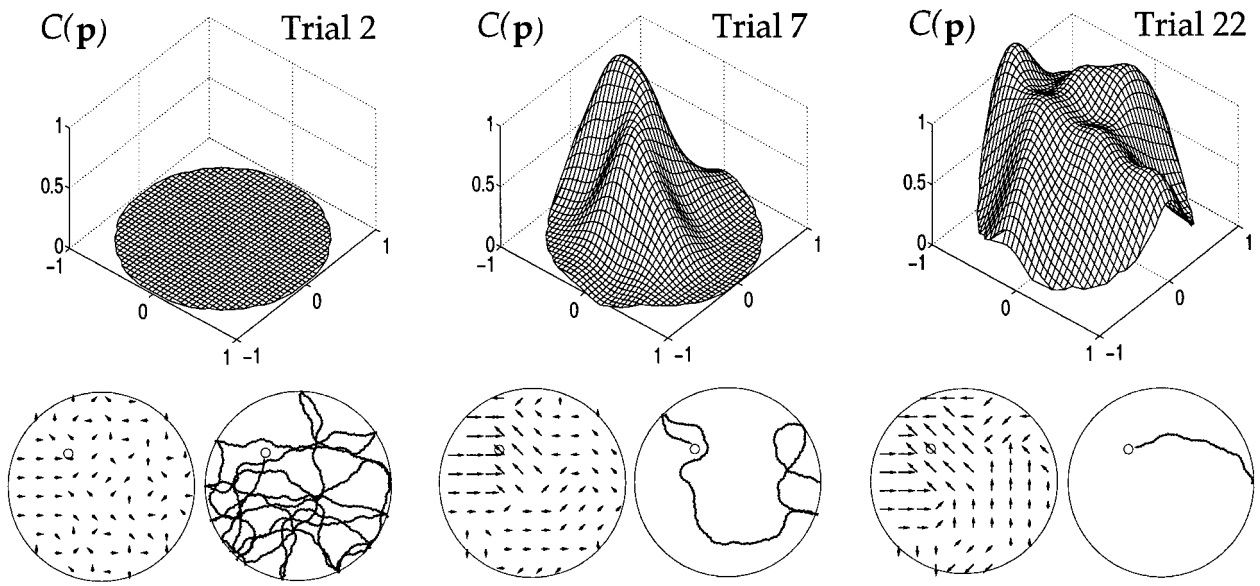

FIGURE 3. Learning in the actor-critic system in RMW. For each trial, the critic's value function $C(p)$ is shown in the upper, three-dimensional plot; at lower left, the preferred actions at various locations are shown (the length of each arrow is related to the probability that the particular action shown is taken by a logarithmic scale); at lower right is a sample path. Trial 2: After a timed-out first trial, the critic's value function remains zero everywhere, the actions point randomly in different directions, and a long and tortuous path is taken to the platform. Trial 7: The critic's value function having peaked in the northeast quadrant of the pool, the preferred actions are correct for locations close to the platform, but not for locations further away. Trial 22: The critic's value function has spread across the whole pool and the preferred actions are close to correct in most locations, and so the actor takes a direct route to the platform.

provides no mechanism by which the experience of previous days can provide any help with learning a new platform position.

One-trial learning by rats on DMP reveals that rats suffer neither of these limitations. Under appropriate training conditions, rats can not only avoid interference between training on successive days, but can also generalize from experience on early days to help performance on later days. To make this clear in computational terms, consider trial 2 on day 6 of training (Fig. 1b). The starting position may be in an area of the environment not explored on trial 1 of that day; nevertheless, the rat swims immediately to the platform. Clearly, knowledge from previous days is being used.

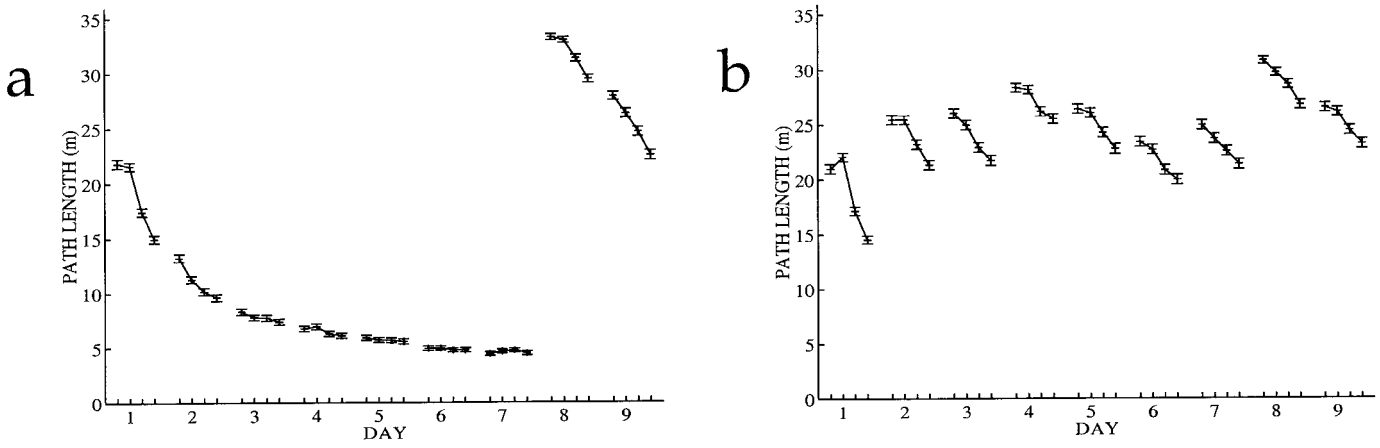

FIGURE 4. Performance of the actor-critic model. For each data point, the mean and standard error in the mean are obtained from 1,000 simulation runs. (a) RMW task, in which the platform occupies the same location. The actor-critic captures acquisition, producing direct paths after around 10 trials. For the last eight trials, however (days 8 and 9), the platform is moved to a different position (reversal), and the model fails to adapt rapidly enough. These simulation results can be compared to Figure 1a. (b) DMP task, in which the platform remains in the same position within a day, but occupies a novel position on each new day. The actor-critic model captures acquisition for the four trials of day 1, for which the task is indistinguishable from RMW. However, as a soon as the platform is moved, the actor-critic not only fails to generalize to the new goal location, but suffers from interference from the previous days' goal locations. Rats suffer neither of these limitations (Fig. 1b).

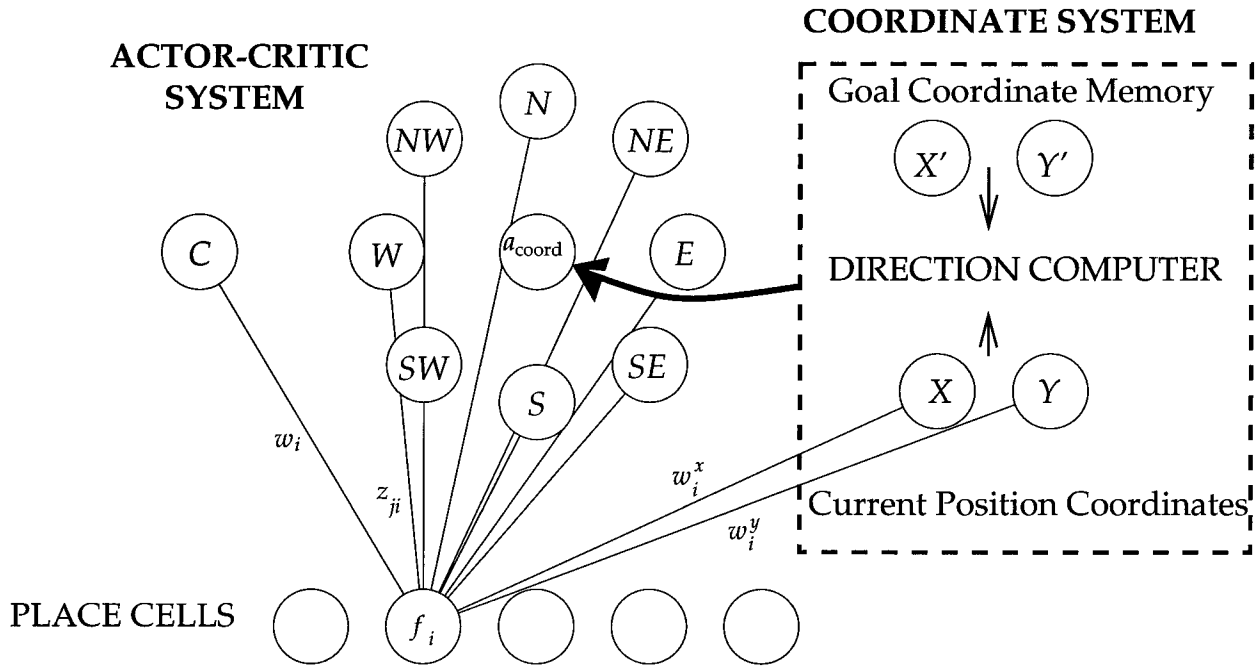

FIGURE 5. The combined coordinate and actor-critic model incorporates both the actor-critic system and a coordinate system. The coordinate system consists of three components: 1) a coordinate representation of current position made up of two cells *X* and *Y*, the firing of which is a function of place cell input; 2) a goal coordinate memory consisting of two cells, *X′* and *Y′*, whose firing reflects the coordinate location of the last place at which the platform was found; and 3) a mechanism which computes the direction in which to swim to get from the current position to the goal. The output direction from the coordinate system is integrated with that from the actor-critic through the "abstract action," marked a_{coord}, which receives reinforcement depending on its performance.

Our model of coordinate learning is based on the observation that the computations involved in the *dead-reckoning* abilities of animals could subserve an all-to-all navigation system for open spaces like a watermaze, if only the dead-reckoning coordinates could be made to be consistent across separate trials, i.e., tied to an allocentric representation of the environment. In effect, we consider making a dead-reckoning system hippocampal-dependent, i.e., dependent on input from the place cell system, and show how such a system can be used to account for one-trial learning in the DMP task.

Dead-reckoning abilities have been documented in (at least) ants, bees, wasps, geese, gerbils, pigeons, rats, and humans (Gallistel, 1990). These abilities are based on the availability of instantaneous estimates of the animal's self-motion, which can be integrated in order to calculate the direction back to a starting point. We will make use of this self-motion information, and the simple geometrical processing required to calculate a heading from the current position; however, we will not make use of path integration. Instead, we will use place cell responses and a predictive TD-based learning rule to acquire a coordinate system in the maze which is defined allocentrically, i.e., independent of the animal's point of origin.

It is hard to acquire an appropriate coordinate system using path integration information alone because of the problem of consistency. When the rat is put in the maze in a new place, there is no way of ensuring that the dead reckoning coordinates it assigns are automatically consistent with those it has assigned elsewhere in previous traversals of the maze. The essential task for the model is learning this consistency (see also Wan et al., 1994). Note that the problem of having a consistent report of head direction (implicitly required in the model) is quite similar. However, head direction generalizes over a much greater spatial extent than does dead reckoning, and, in the experiments being modeled, vestibular disorientation or other manipulations of the head direction system were not used.

The problem for the rat is therefore to learn *globally* consistent coordinates based only on local *relative* self-motion. The key observation is that for every move that the rat makes, the difference between its estimates of coordinates at the ending and starting locations should be exactly the relative self-motion during the move. This consistency condition can be used as the basis for a TD learning rule for learning coordinates.

Figure 5 shows a simple model of learning and using coordinates. The coordinate system consists of two networks, one which learns *X* coordinates (as $X(\mathrm{p}) = \Sigma_i w_i^X f_i(\mathrm{p})$), and one which learns *Y* coordinates (as $Y(\mathrm{p}) = \Sigma_i w_i^Y f_i(\mathrm{p})$), both using inputs from place cells which act in exactly the same way as in the actor-critic model, each producing a firing rate $f_i(\mathrm{p})$ as a function of location p. The choice of *X* and *Y* coordinates, or even just two orthogonal directions, is of course arbitrary, but the basic problem of making coordinates consistent will exist whatever particular coordinate system is used. The *X* and *Y* coordinates have been chosen for simplicity, and to illustrate clearly the learning problem.

As the rat moves around, the weights $\{w_i^X\}$ and $\{w_i^Y\}$, $i = 1, \ldots, N$ that define the coordinates are adjusted according to:

$$\Delta w_i^X \propto (\Delta x_t + X(p_{t+1}) - X(p_t)) \sum_{k=1}^{t} \lambda^{t-k} f_i(p_k) \quad (11)$$

$$\Delta w_i^Y \propto (\Delta y_t + Y(p_{t+1}) - Y(p_t)) \sum_{k=1}^{t} \lambda^{t-k} f_i(p_k) \quad (12)$$

where Δx_t and Δy_t are the self-motion estimates in the x direction and y direction, respectively. An interesting technical issue is the use of a more general form of the TD algorithm, which leads to the sums on the right of Equations 11 and 12. This form works by enforcing consistency between coordinates not only across one timestep, but across many. The parameter λ determines to what extent more distant timesteps are also considered. Theoretical arguments suggest that since the terms Δx_t and Δy_t are likely to be quite accurate, distant timesteps are useful, and therefore a high value of λ should make learning fastest (Watkins, 1989). Simulations confirmed this, and so we set λ to 0.9.

Using Coordinates to Control Actions

In dead reckoning, an animal computes, from its current coordinate, a bearing back to a point of origin. In the model, a coordinate controller computes, given its current allocentrically defined coordinate, a bearing to whatever other coordinate is of interest. This requires performing a simple vector subtraction, which is just the same computation that dead reckoning also requires (although we do not explicitly model the computation in neural or connectionist terms). The additional, nontrivial requirement for the general coordinate system is some form of goal coordinate memory, a point we will return to in the Discussion. At certain times, however, there will be no remembered goal coordinate: during the first trial, and, on DMP, every time the rat reaches the position where it thinks the goal is, and finds it to be moved. When there is no goal coordinate in memory, we make the coordinate controller specify random, exploratory actions.

When coordinates have been learned, a coordinate controller such as that described above is potentially extremely useful; however, if coordinates are poorly learned, there are no guarantees that the controller is at all useful. Early on, the controller will produce paths which are not only indirect, but are even prone to catastrophic loops (see results). The ability of the controller to switch to random exploration can sometimes alleviate this problem, but even then is guaranteed to produce highly suboptimal paths.

The solution adopted in this paper is to combine coordinate control with the actor-critic architecture. One way to do this is shown in Figure 5. Here, there is an additional action cell, a_{coord}, representing the rat's preference for the swimming direction offered by the coordinate system. This coordinate action can be chosen stochastically, in competition with normal actions, rather like the "abstract actions" of Singh (1992). The coordinate action is reinforced by the critic in a similar manner to the other actions: when the coordinate action is chosen, the weighting of the coordinate action cell is changed by an amount proportional to the prediction error provided by the critic. Unlike the normal actions, preference for the coordinate action is independent of location in the watermaze. A second difference is that when there is no remembered goal coordinate—and the controller is specifying random exploratory actions instead of actions based on its coordinates—then the controller does not participate in learning, i.e., a_{coord} is not updated. The effect is that coordinate control comes to be relied upon gradually, as it gives increasingly accurate information about where both the animal and the goal are located. Note that the coordinate system suggests appropriate actions without suggesting values associated with these actions.

PERFORMANCE OF THE COMBINED COORDINATE AND ACTOR-CRITIC MODEL

Simulation Methods

The combined model was tested in simulated versions of the RMW and DMP tasks, using the same simulation environment as described for the actor-critic model. Learning rate parameters (including those governing the constants of proportionality in Equations 11 and 12 and Equation 10 for the abstract action) were again optimized.

Simulation Results

Figure 6a shows the development of the X and Y coordinates over days. Early on, e.g., day 2, trial 2, the coordinate surface is uneven. By day 6, it is relatively smooth. Note that the coordinate learning system receives no direct information about how the coordinates should be centered. Three factors control the centering: the boundary of the arena, the prior setting of the coordinate weights (in this case, all were zero), and the position and prior value of any absorbing area (in this case the platform). These factors are arbitrary, and one might worry that the coordinates could drift over time and thereby invalidate coordinates that have been remembered over long periods. Consider, for example, a rat that had learned coordinates throughout a maze but was then confined for a period of time to a particular region of the maze. If the rat was later released, but coordinates had drifted in the meantime, navigation within the maze as a whole would be affected. However, since the expected value of the prediction error at timesteps should be zero for any self-consistent coordinate mapping, such a mapping should remain stable. This is demonstrated for a single run: Figure 6c,d shows the mean value of coordinates $\bar{X}$ evolving over trials, indicating that there is little drift after the first few trials.

The difficulty in using the coordinates by themselves to specify actions is clear from the nature of the gradient of these functions (Figure 7). Early on in learning, the coordinate functions are highly irregular, and a direction specified on the basis of these functions is worse than simply suboptimal, since catastrophic

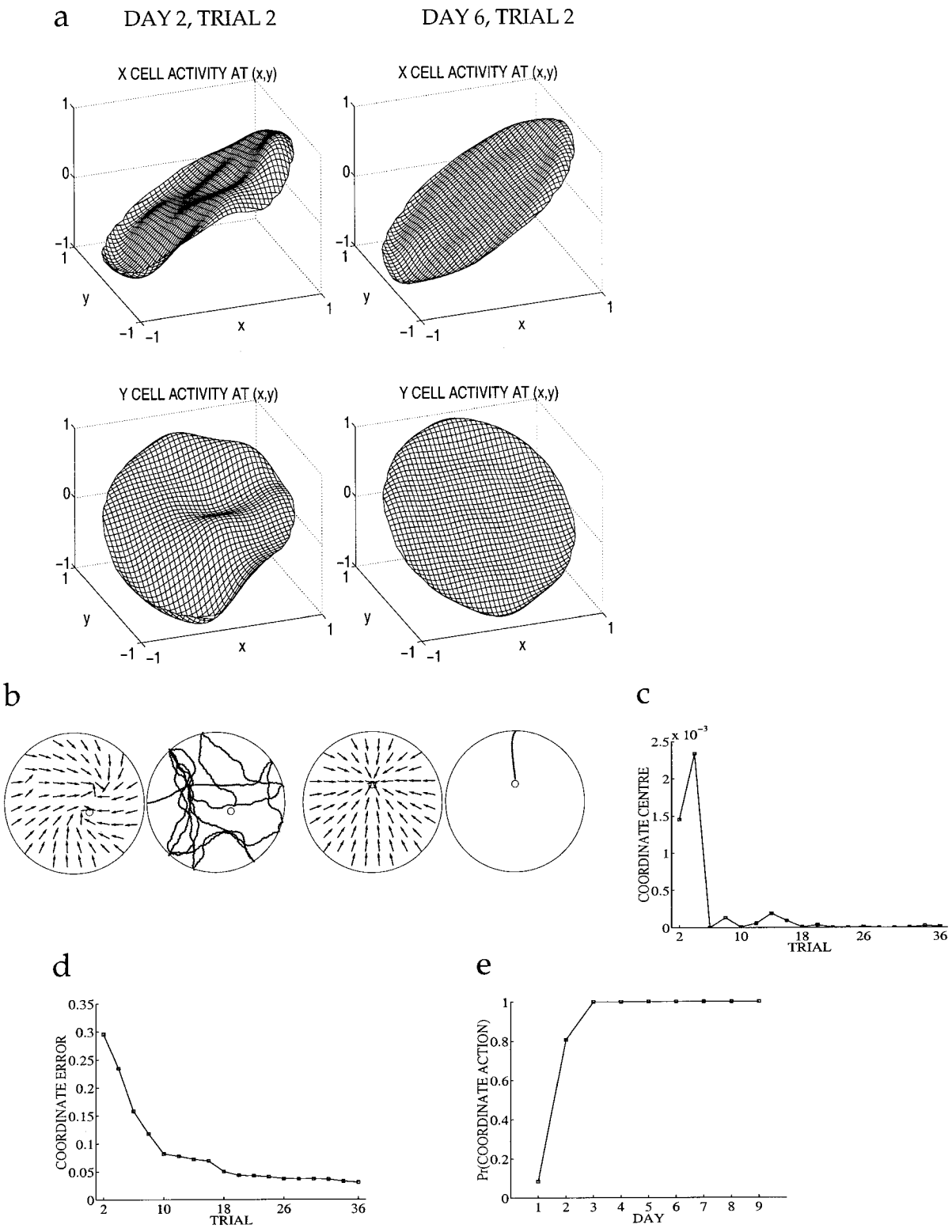

FIGURE 6. **(a) The *X* and *Y* coordinate functions develop gradually over days, at first being quite uneven (e.g., day 2), but becoming quite smooth by day 6. (b) Below each coordinate are examples of preferred actions, and paths, for trial 2 of a simulated run of DMP using the full model. On the second trial of day 2, performance is quite poor. By day 6, one-trial learning is evident. (c) The centering of the *X* coordinates, as measured by the mean, does not drift by the time coordinates are smooth. This is expected, since as the coordinates become consistent, all weight changes tend to zero. (d) Error in the *X* coordinates for the same simulation, measured as the variance for each coordinate about its desired value relative to the mean. The error stabilizes after a few trials. (e) As coordinates improve, the weighting of the coordinate-based action increases. Thus, the probability of taking the coordinate action, averaged over all time points within a trial, and over all the trials of a day, is shown to increase.**

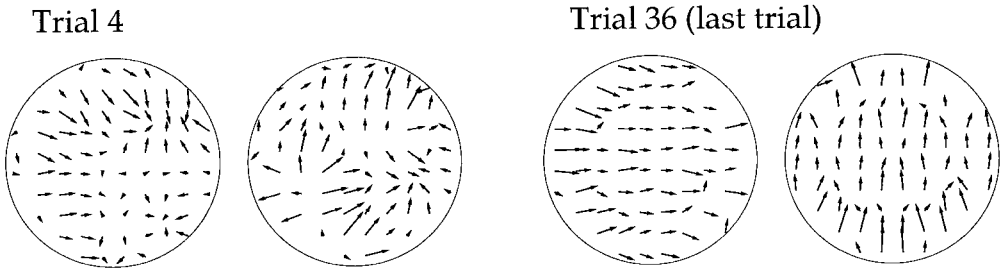

FIGURE 7. Gradient of the coordinate functions. The gradient is a very sensitive measure of smoothness. On trial 4, coordinates are still not at all smooth; navigation based on these functions alone would be prone to catastrophic loops, i.e., would never reach the platform. By comparison, the actor-critic scheme develops effective values and actions for control by trial 4 (Fig. 3), and it is this control that allows the rat to move through the environment, and so improve its coordinate functions. By trial 36, coordinates are smoother and the gradients reflect the *X* and *Y* directions.

loops are possible. This difficulty motivates the combination of the coordinate control with the actor-critic, allowing the conventional actions of the actor-critic to dominate early on, but enabling coordinate control to come to dominate as its actions prove more reliable than the conventional ones. This transfer of control happens rapidly during the DMP task (Fig. 6e).

Figure 8a shows the performance of the combined model in the RMW task. Like the actor-critic model discussed above, the combined coordinate and actor-critic model successfully captures the acquisition of this task. Moreover, this model can also account for the rapid learning to the novel platform during the reversal phase, as seen in Figure 1a. Figure 8b shows the performance of the combined model in the DMP task. Just as in Figure 1b, acquisition during early days is gradual, while by day 6, one-trial learning is evident in the difference in performance between trials 1 and 2.

DISCUSSION

A model of hippocampally dependent navigation has been presented that uses place cells as a representational substrate for learning three different functions of position in an environment. The actor-critic component of the model learns the temporal proximity of locations to a single escape platform and also appropriate actions that get there quickly. By itself, the actor-critic model captures initial acquisition performance in RMW. However, its performance diverges from that of rats the moment the platform is moved, failing to account for the good reversal performance shown by rats, or for the even more striking one-trial learning in DMP. A further component of the model learns *X* and *Y* coordinates, a goal-independent representation of the environment, and this provides the flexibility necessary for DMP by

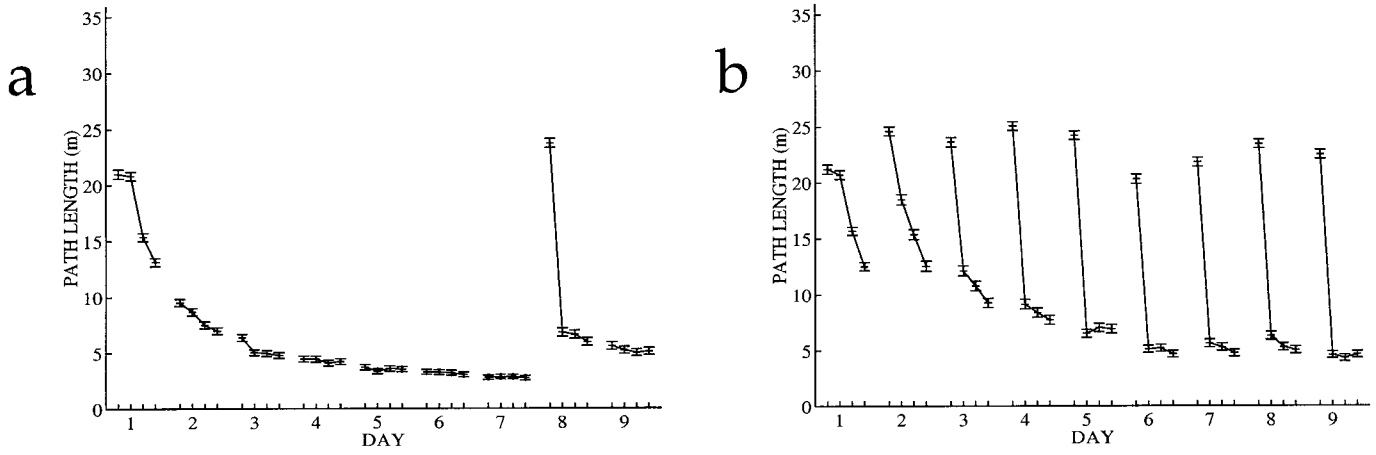

FIGURE 8. Performance of the combined coordinate and actor-critic model. For each data point, the mean and standard error in the mean are obtained from 1,000 simulation runs. (a) RMW task, in which the platform occupies the same location. The combined model captures both acquisition, producing direct paths after around 10 trials, and reversal, producing rapid adaptation to the change in platform position on day 8 (see Fig. 1a). (b) DMP task, in which the platform remains in the same position within a day, but occupies a novel position on each new day. The combined model captures the acquisition of one-trial learning: the improvement within each day is gradual early in training, but becomes a one-trial improvement by day 6. The model provides a good match to the data (Fig. 1b).

allowing navigation to arbitrary goals. The complete model combines coordinates with the actor-critic architecture and accounts for the performance of rats in the RMW task, including the reversal, and in the DMP task.

The Contribution of the Model

The question posed at the beginning of the paper was: how might place cell activity be useful for navigation, without containing all the spatial information necessary for navigation? We have shown that place cells provide an excellent representation for learning values, actions, and coordinates. In the watermaze, the overlap between adjacent place fields in the model supports generalization because nearby places have similar optimal values, coordinates, and correct actions. The point is particularly well-illustrated by the performance of the actor-critic in the RMW task. Reinforcement learning methods such as the actor-critic are infamous for the large numbers of training trials required for learning, which in most applications run to the thousands. With place cells as an input representation, the actor-critic learns the RMW task in about 10 trials.

As well as considering a somewhat standard application of TD learning, i.e., the actor-critic, we have also presented a novel application of TD learning in the form of a network that learns consistent coordinates in an environment. This learning is found to be extremely fast, with smooth coordinates acquired after about 16 trials. Moreover, the coordinates learned are stable, despite being learned from relative information. The problem of global consistency is a general one that affects all navigating systems which use self-motion information to build map-like representations. The solution presented here partners a statistically efficient learning algorithm, TD learning, with the stable, allocentrically defined representation of the environment that hippocampal place cells provide.

What does the model tell us about the spatial tasks themselves? First, since the actor-critic component can capture acquisition performance of rats in RMW, this acquisition does not provide evidence for a "cognitive map" (Morris et al., 1982). The actor-critic is not the first model to provide a nonmapping account of the task (Zipser, 1986; Wilkie and Palfrey, 1987; Burgess et al., 1994; Brown and Sharp, 1995; Blum and Abbott, 1996). It is, however, the first to incorporate a principled solution to the distal reward problem, the critical component of which is the temporal difference (TD) learning rule. This solution is quite general, since nothing is assumed about the topology of the environment (beyond the structure implicit in the place cell representation), and so the actor-critic has the potential to learn in more complex environments, such as environments with barriers.

Second, the model demonstrates that it may be dangerous to conclude, as in a recent review of models of navigation by Trullier et al. (1997), that *metric* navigation methods subsume *topological* navigation methods. The DMP task can be solved using metric information supplied by the learned coordinates, but the model knows very little about the topological structure of the environment, and this is its principal weakness. Likewise, other demonstrations of navigational ability, such as execution of paths in the dark (Collett et al., 1986) or shortcuts (Menzel, 1973; Gallistel, 1990), provide evidence for the use of metric information, but not necessarily for the learning or use of topological information about environments. Few spatial tasks demand even coordinates, and a challenge for the future is to explore whether rats use still more sophisticated (i.e., topologically richer) representations of space.

Our work contains several simplifications. A way by which the coordinate controller might suggest a value to the critic was not included, and so the critic itself becomes inaccurate as one-trial learning is established. This may be justified on the grounds of parsimony: there is little evidence to constrain the choice of mechanism either for this, or for the closely related issue of learning "set," i.e., the information about the task that a rat acquires as it finds the platform changing position each day. Furthermore, "set" learning is clearly incomplete, since on the first trial of each new day, normal rats continue to revisit the position of the platform on the previous day, even though this is always incorrect (Steele and Morris, 1999). We have also not addressed the problem of learning to navigate to goals in many different environments, or the formation of place fields themselves, nor the possibility that place fields change during either task. An important related issue is the possibility of transfer between environments, such as enhanced learning in one watermaze after pretraining in a different one of a similar size and shape, and to a similar task (Whishaw, 1991; Bannerman et al., 1995; Saucier and Cain, 1995). The key representational role played by place cells in our model implies that following a complete redistribution of place fields, coordinates in particular would have to be relearned. It is hard to make a strong prediction about transfer, however, because it is not known exactly how place fields come to form in a novel environment, and to what extent they might show regularities with previously experienced environments that have relevance to the current behavioral situation. Additionally, the search strategy of the simulated rats was based on a random walk, commencing at the starting point and ending at the platform. A better strategy would be to search all areas of the pool more uniformly, and this may indeed be what the experimental rats did. The key difference between the two strategies lies in not returning to previously searched areas. In fact, the actor-critic has the potential to learn such a strategy, if punishments are associated with moves which do not take the rat onto the platform. In this case, novel areas will appear more attractive than previously searched areas.

Relationship to Experimental Data

The hippocampus

In the model, place cells in the hippocampus provide a representation of the current position of the rat, i.e., its current state in the control task. This representation has almost ideal properties for learning the actor and critic because nearby places, which will have similar values, and require similar actions, are represented similarly by the place cell system. Interestingly, where this condition is violated, as for example in the case of nearby places on opposite sides of a barrier, the place cell representation

adapts accordingly: place fields which cross barriers do not generally occur, and if a barrier is added to the environment so as to split an extant place field, the field extinguishes, i.e., the cell associated with the field ceases to fire (Muller et al., 1987). The implication is that proximity is defined by the task, not purely by space. Activity-dependent changes in place fields' shape and centering have been observed (Mehta et al., 1996), and these can be interpreted in terms of improving the state representation for use by learning systems such as the actor-critic (Dayan, 1993).

A straightforward generalization from this role in navigation is that hippocampal neurons provide a state space representation appropriate to whatever task is at hand, so that nonspatial tasks might lead to nonlocation-based representations of the environment. This is supported by physiological evidence for the nonspatial tuning of rodent hippocampal cells (Wiener, 1996; Eichenbaum, 1996; Wood et al., 1999) and of cells in the primate hippocampus (O'Mara et al., 1994), and is in line with a number of current theories of hippocampal function (McClelland and Goddard, 1996; Rudy and Sutherland, 1995). This hypothesis is not at odds with the possibility that animals learn specialized representations of the environment, like the coordinates of the model, in order to solve particular problems. As suggested by various theories, it is also quite possible that the representations are present in the hippocampus for only a limited time, and are ultimately consolidated into the cortex.

In the model, weight changes due to navigational learning occur downstream of hippocampal place cells. Consonant with this, Steele and Morris (1999) found that, after 9 days of pretraining on DMP, animals can, at short memory delays, continue to perform one-trial learning to novel platform positions during pharmacological blockade of N-methyl-D-aspatate (NMDA) receptors in the hippocampus. However, 9 days is long enough to learn a coordinate system, and so the experiment of Steele and Morris (1999) does not distinguish between models in which coordinate-like information is stored inside the hippocampus, and models in which it is stored outside.

An interesting issue raised by the model concerns the explicit memory for current goal location, demanded by the coordinate model. Evidence suggests that goal memory may be a dissociable computational factor in navigation. Steele and Morris (1999) found that, after 9 days of pretraining, animals in which hippocampal synaptic plasticity has been blocked by an NMDA antagonist show a delay-dependent impairment during DMP, i.e., trial 2 performance in DMP is impaired if, and only if, the delay between trials 1 and 2 is long (20 min or 2 h; short delay was 15 s). Within the framework of the model, this impairment corresponds to a selective disruption of the goal coordinate memory. Moreover, the data suggest that the normal operation of this goal memory is dependent on hippocampal NMDA receptors. The important question of what happens to place cells themselves if hippocampal synaptic plasticity is blocked has only recently begun to be addressed. Pioneering studies using NMDA antagonists suggest that hippocampal synaptic plasticity is necessary for the long-term stability of place fields within an unfamiliar environment, but not necessary if the environment has become familiar during a period of preoperative training (Kentros et al., 1998). By this account, place fields ought to have been stable throughout all stages of the DMP task conducted by Steele and Morris (1999).

The actor-critic

The actor-critic is a general learning scheme that has been used to model phenomena in classical and instrumental conditioning that are likely to be largely independent of the hippocampal formation. For example, Montague et al. (1996) built an actor-critic model of rewarded conditioning behavior based on electrophysiological evidence on the activity of cells in the dopamine system (Schultz et al., 1997). In this model, neurons in the ventral tegmental area and the substantia nigra pars compacta report the prediction error term δ_t in Equation 7, and the dorsal striatum plays the role of the actor. Both the ventral and dorsal striatum of the rat receive outputs from the CA1 hippocampal subfield, an area where place cells are found (Wiener, 1996). However, little is currently known about the activity of these systems during navigation, or how or where the values may be stored.

The coordinates

There is no evidence as yet for the neural implementation of the coordinate representation. However, the phenomenon of dead reckoning is well-documented in many animals, and strongly suggests both that a coordinate representation of some sort exists, and that neural mechanisms exist to perform simple vector subtraction. The particular X and Y coordinate representation we have used is extremely simple; we have used it to demonstrate clearly the problem of building globally consistent coordinates from relative self-motion information, which will be present for any coordinate system. The key feature of the model is to make the coordinate representation hippocampally dependent, in the sense of relying upon information from place cells, and the model demonstrates both that place cells provide an appropriate representation from which to learn coordinates, and that, with the TD learning rule, coordinate learning can be extremely fast.

Relationship to Other Models

The two key issues separating models of navigation are, from a neural perspective, the extent to which the hippocampus itself solves the navigation problem, and, from a computational perspective, the generality of the suggested control scheme. Both actor-critic and coordinate components use the hippocampus only for a representation of state (i.e., place). The actor-critic is a completely general control mechanism, working in environments with arbitrarily complicated shapes and reward contingencies, but is fairly inflexible. The coordinate model is flexible, but specialized to navigation in a restricted class of environments.

The model of Blum and Abbott (1996) (see also Abbott and Blum, 1995; Gerstner and Abbott, 1996) is very closely related to dynamic programming, the control mechanism underlying TD rules. They proposed that place cells express a decodable population code for position, and that subtle changes in the population code, due to the operation of temporally asymmetric Hebbian synaptic plasticity between place cells in field CA3 while the rat is

swimming, can be interpreted as reporting at each location the average swimming direction that takes the rat to the goal. This essentially performs one step of the dynamic programming technique of policy improvement, starting from a random policy. However, for general control problems, just one step of policy improvement is inadequate; even in the RMW task which they modeled, it was necessary to include a reinforcement process which modulated the Hebbian plasticity, in a manner similar to that of Brown and Sharp (1995).

Gerstner and Abbott (1996) extended the model of Blum and Abbott (1996) to the case of navigation to multiple goal locations. In their model, the (remembered) position of the goal modulates the activities of place cells, allowing the connections between the single set of place cells that are active in an environment to store the swimming direction appropriate to the multiple goals. Having learned synaptic weights appropriate for a few goals, navigation to novel goals is possible by interpolation. The model might use this feature to solve DMP, even in the face of the pharmacological blockade. However, there are various counts against the model. First, the modulation of place cell activity by goal position is not observed; indeed, there is evidence against it (Speakman and O'Keefe, 1991). Second, both versions of this model embed the whole problem for navigation in the hippocampus proper, in the connections between CA3 cells. This is hard to reconcile with the results of Bannerman et al. (1995) and Saucier and Cain (1995), which suggest that plasticity in this region may not be necessary to learn a watermaze task in a novel environment. Third, one of the key computational operations in the models is population decoding of the position of the rat that is encoded in the activities of the place cells. Calculating this requires knowledge of something equivalent to coordinates in the environment, i.e., a priori knowledge of the location (in some coordinate system) of the center of each place field. Some additional, unspecified scheme for learning these coordinates consistently across the environment is essential.

Like the actor-critic system, Burgess et al. (1994) and Brown and Sharp (1995) also suggested schemes in which place cells play the more limited role of providing a reliable code for space. Both papers considered an RMW-like task which presents a distal reward problem. Burgess et al. (1994) used the output of place cells to construct subicular cells with extended place fields, which in turn were used to learn postulated goal cells, which fired across the extent of an entire environment, performing a job like the actor. Learning of the goal cells only happens when the animal actually reaches the goal, but this is sufficient because the extended range of the goal cells means, in effect, that there is no longer a distal reward problem. If by some means the firing of goal cells for different goals could be distinguished, it is possible that the model could also address the DMP task, by having a subicular cell for every possible goal. However, the use of large firing field representations in this manner raises a number of issues. First, if the subicular cells that fire when the animal is at the goal do not cover the whole environment, there will be places for which the animal will not learn appropriate actions. Second, the mechanism which generates large subicular fields can be expected to learn more slowly than TD learning methods, since it attempts to produce a smooth, monotonic function of distance in the subicular cells by essentially averaging over place cell activity traces for each subicular cell (i.e., for each potential goal). Third, the model does not use a general learning scheme for control, and so can only accomplish tasks such as avoiding obstacles by making detours that are significantly larger than necessary and which, for inconveniently located barriers, may not work at all. Brown and Sharp (1995) presented a simpler model in which place cells are associated with responses, and in which learning is gated by reward. However, as noted in the Introduction, the model relies on a trace-like learning rule which is likely to be a very inefficient way of learning predictions compared to the TD learning rule used in the actor-critic model. The model does, however, suffer the same limitations as the actor-critic with respect to the learning of a DMP task.

The problems involved in learning a coordinate system have been addressed by Wan et al. (1994). In their model, coordinates are represented by an extrahippocampal path integration module that operates more conventionally, representing coordinates with respect to some current point of origin. Their model demonstrates how place cell firing might come through learning to be independent of sensory information, at least for a short while, relying instead on input from the path integrator. It also addresses the inverse problem of what happens when the path integrator becomes invalid, as for example on each new trial of a watermaze task, because the path integrator learns to set itself by the output of place cells. In a completely novel region, a new origin is selected and new coordinates laid down. However, if previous experience is of value to the animal, it must return to areas of the environment where place cells can correctly set the path integrator; hence, for example, trial 2 of a watermaze task could not produce any learning until a familiar area was traversed, thus throwing away potentially valuable experience, as well as constraining the animal's search. The TD-based model of this paper avoids both shortcomings by directly tackling the problem of inconsistent coordinates.

Finally, a quite different view of hippocampal function from that taken by the models discussed so far is that the hippocampus is directly involved in some forms of flexible processing, e.g., manipulating sequences of mnemonic or spatial information (Levy, 1996) or performing complicated computations, as in the demonstration of transitive inference (Bunsey and Eichenbaum, 1996). Although direct experimental support for this view is lacking, it is not possible, on the basis of current evidence, to rule it out. However, transitive inference may be a case in point, because working out a global order from local relationships is a similar task to calculating globally consistent coordinates from local dead-reckoning information. It is possible that the hippocampus computes the inference directly; it is also possible that downstream systems make the computation, but rely on the hippocampal representation to do so. With regard to navigation tasks, we have demonstrated that although the observed activity of place cells appears limited, it makes sense if used in the right system with the right learning rule. Indeed, according to the model presented here, the very characteristics that make place cell activity seem so redundant (namely, localization, directional

independence, and stability) contribute most to their suitability within a navigational learning context.

Predictions of the Model

On the basis of the model, the following three predictions can be made.

1. Placement trials should support DMP, once rats have acquired one-trial learning. After a certain amount of training, rats should have a system that specifies the coordinates of any location they occupy. This implies that, by this stage of learning, mere placement on a platform in a novel position might be sufficient to allow asymptotic performance of the next trial. This prediction would, however, depend on the learning set behavior of the rats in terms of knowing the appropriate response, having just been placed on a platform.
2. Rats for which hippocampal synaptic plasticity is blocked, but only after place fields have been established in an environment, should be unimpaired in learning a RMW task. The model suggests that the actor-critic is located outside of the hippocampal formation, and just uses information from place cells as a representation of state. Therefore, provided the place cells have been established (e.g., during a latent learning period of some sort), actor-critic learning should progress normally. The complication again is learning set behavior: if blocking hippocampal plasticity prevented the animals from learning the nature of the task, this too would have to be ensured during a pretraining period.
3. Rats for which hippocampal synaptic plasticity is blocked, but only after place fields have been established in an environment, might also be unimpaired in learning a DMP task. If this was found to be true, it would suggest that the coordinate system (in particular, cells X and Y in the model) is located outside the hippocampus. An impairment, on the other hand, would suggest that coordinates are located within the hippocampus. The experiment of Steele and Morris (1999) did not distinguish between the two alternatives, because synaptic plasticity was blocked only after extensive pretraining (which provided the one-trial learning data which we have modeled). The same considerations apply for this prediction as for the previous one, in terms of establishing place fields, and acquiring the learning set.

Acknowledgments

Funding for this work came from the McDonnell-Pew Foundation, an Edinburgh University Holdsworth Scholarship (D.J.E.), an MRC Programme Grant (R.G.M.M.), NSF grant IBN-9634339 (P.D.), the Surdna Foundation (P.D.), and the Gatsby Charitable Foundation (P.D.). Support also came from the University of Oxford McDonnell-Pew Centre for Cognitive Neuroscience (D.J.F. and R.G.M.M.). We thank Robert Steele for detailed discussions, and Richard Sutton, Matt Wilson, and three anonymous reviewers for comments. Opinions expressed are those of the authors.

REFERENCES

Abbott LF, Blum KI. 1995. Functional significance of long-term potentiation for sequence learning and prediction. Cereb Cortex 6:406–416.

Bannerman DM, Good MA, Butcher SP, Ramsay M, Morris RGM. 1995. Distinct components of spatial learning revealed by prior training and NMDA receptor blockade. Nature 378:182–186.

Barnes CA. 1979. Memory deficits associated with senescence: a neurophysiological and behavioral study in the rat. J Comp Phys Psych 93:74–104.

Barto AG, Sutton RS, Anderson CW. 1983. Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans Syst Man Cybernet 13:834–846.

Barto AG, Sutton RS, Watkins CJCH. 1990. Learning and sequential decision making. In: Gabriel M, Moore J, editors. Learning and computational neuroscience: foundations of adaptive networks. Cambridge, MA: MIT Press, Bradford Books. p 539–602.

Bertsekas DP, Tsitsiklis JN. 1996. Neuro-dynamic programming. Belmont, MA: Athena Scientific.

Blum KI, Abbott LF. 1996. A model of spatial map formation in the hippocampus of the rat. Neural Comput 8:85–93.

Brown MA, Sharp PE. 1995. Simulation of spatial learning in the Morris water maze by a neural network model of the hippocampal formation and nucleus accumbens. Hippocampus 5:171–188.

Bunsey M, Eichenbaum H. 1996. Conservation of hippocampal memory function in rats and humans. Nature 379:255–257.

Burgess N, Recce M, O'Keefe J. 1994. A model of hippocampal function. Neural Networks 7:1065–1081.

Collett TS, Cartwright A, Smith BA. 1986. Landmark learning and visuo-spatial memories in gerbils. J Comp Physiol [A] 158:835–851.

Dayan P. 1991. Navigating through temporal difference. In: Lippmann RP, et al., editors. In: Lippman RP, Moody JE, Touretzky DS, editors. Advances in Neural Information Processing System 3. p 464–470.

Dayan P. 1993. Improving generalisation for temporal difference learning: the successor representation. Neural Comput 5:613–624.

Eichenbaum H. 1996. Is the rat hippocampus just for "place"? Curr Opin Neurobiol 6:187–195.

Gallistel CR. 1990. The organization of learning. Cambridge, MA: MIT Press.

Gerstner W, Abbott LF. 1996. Learning navigational maps through potentiation and modulation of hippocampal place cells. J Comput Neurosci 4:79–94.

Kentros C, Hargreaves E, Hawkins RD, Kandel ER, Shapiro M, Muller RV. 1998. Abolition of long-term stability of new hippocampal place cell maps by NMDA receptor blockade. Science 280:2121–2126.

Levy WB. 1996. A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. Hippocampus 6:579–590.

McClelland JL, Goddard NH. 1996. Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. Hippocampus 6:654–665.

Mehta MR, McNaughton BL, Barnes CA, Suster MS, Weaver KL, Gerrard JL. 1996. Rapid changes in the hippocampal population code during behavior: a case for Hebbian learning in vivo. Soc Neurosci Abstr 22:724.

Menzel EW. 1973. Chimpanzee spatial memory organization. Science 182:943–945.

Montague PR, Dayan P, Sejnowski TJ. 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.

Morris RGM. 1981. Spatial localisation does not require the presence of local cues. Learn Motiv 12:239–260.

Morris RGM. 1983. An attempt to dissociate spatial-mapping and working-memory theories of hippocampal function. In: Seifert W,

editor. The neurobiology of the hippocampus. London: Academic Press.
Morris RGM, Garrud P, Rawlins JNP, O'Keefe J. 1982. Place navigation impaired in rats with hippocampal lesions. Nature 297:681–683.
Muller RU, Kubie JL, Ranck JB. 1987. Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. J Neurosci 7:1935–1950.
O'Keefe J, Burgess N. 1996. Geometrical determinants of the place fields of hippocampal neurons. Nature 381:425–428.
O'Keefe J, Dostrovsky J. 1971. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely moving rat. Brain Res 34:171–175.
O'Keefe J, Nadel L. 1978. The hippocampus as a cognitive map. London: Clarendon.
O'Keefe J, Recce ML. 1993. Phase relationship between hippocampal place cells and the EEG theta rhythm. Hippocampus 3:317–330.
O'Mara SM, Rolls ET, Berthoz A, Kesner RP. 1994. Neurons responding to whole-body motion in the primate hippocampus. J Neurosci 14:6511–6523.
Panakhova E, Buresova O, Bures J. 1984. Persistence of spatial memory in the Morris water maze tank task. Int J Psychophysiol 2:5–10.
Poggio T, Girosi F. 1990. Networks for approximation and learning. Proc IEEE 78:1481–1497.
Redish AD, Touretzky DS. 1997. Navigating with landmarks: computing goal locations from place codes. In Ikeuchi K, Veloso M, editors. Symbolic Visual Learning. Oxford University Press. Chapter 12, pp. 325–351.
Rudy JW, Sutherland RJ. 1995. Configural association theory and the hippocampal formation: an appraisal and reconfiguration. Hippocampus 5:375–389.
Saucier D, Cain DP. 1995. Spatial learning without NMDA receptor dependent long-term potentiation. Nature 378:186–189.
Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. Science 275:1593–1599.
Singh SP. 1992. Transfer of learning by composing solutions of elemental sequential tasks. Machine Learning, 8:323–339.
Skaggs WE, McNaughton BL, Wilson MA, Barnes CA. 1996. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. Hippocampus 6:149–172.
Speakman A, O'Keefe J. 1991. Hippocampal complex spike cells do not change their place fields if the goal is moved within a cue controlled environment. Eur J Neurosci 2:544–555.
Steele RJ, Morris RGMM. 1999. Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the NMDA-antagonist D-AP5. Hippocampus 9:118–136.
Sutherland RJ, Whishaw IQ, Kolb B. 1983. A behavioural analysis of spatial localisation following electrolytic, kainate or colchicine induced damage to the hippocampal formation in the rat. Behav Brain Res 7:133–153.
Sutton RS. 1988. Learning to predict by the methods of temporal difference learning. Machine Learn 3:9–44.
Sutton RS, Barto AG. 1988. Reinforcement learning: an introduction. MIT Press. A temporal-difference model of classical conditioning. Tech Report, GTE Labs, TR87-509.2, 1987.
Taube JS. 1995. Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. J Neurosci 15:70–86.
Trullier O, Wiener SI, Berthoz A, Meyer J. 1997. Biologically-based artificial navigation systems—review and prospects. Prog Neurobiol 51:483–544.
Wan HS, Touretzky DS, Redish AD. 1994. Towards a computational theory of rat navigation. In: Mozer M, Smolensky P, Toutetzky D, Elman J, Weigend A, editors. Proceedings of the 1993 Connectionist Models Summer School. Hillsdale, NJ: Lawrence Erlbaum. p 11–19.
Watkins CJCH. 1989. Learning from delayed rewards. Ph.D. thesis, University of Cambridge.
Whishaw IQ. 1985. Formation of a place learning set in the rat: a new paradigm for neurobehavioral studies. Physiol Behav 35:139–145.
Whishaw IQ. 1991. Latent learning in a swimming pool place task by rats: evidence for the use of associative and not cognitive mapping processes. Q J Exp Psychol [B] 43:83–103.
Wiener SI. 1996. Spatial, behavioral and sensory correlates of hippocampal CA1 complex spike cell activity: implications for information processing functions. Prog Neurobiol 49:335–361.
Wilkie DM, Palfrey R. 1987. A computer simulation model of rats' place navigation in the Morris water maze. Behav Res Methods Instr Comput 19:400–403.
Wilson MA, McNaughton BL. 1993. Dynamics of the hippocampal ensemble code for space. Science 261:1055–1058.
Witten IH. 1977. An adaptive optimal controller for discrete-time Markov environments. Information Control 34:286–295.
Wood ER, Dudchenko PA, Eichenbaum H. 1999. The global record of memory in hippocampal neuronal activity. Nature 397:613–616.
Zipser D. 1986. Biologically plausible models of place recognition and goal location. In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing, volume 1. Cambridge, MA: MIT Press.