Modelling hippocampal and dorsolateral striatal contributions to learning across domains

Running Title: Hippocampal and striatal learning model

Jesse P. Geerts[1,2,+], **Fabian Chersi**[2,3,+], **Kimberly L. Stachenfeld**[4], **and Neil Burgess**[2,*]

[1]Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London, 25 Howland Street, London W1T 4JG, UK
[2]Institute for Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AZ, UK
[3]GrAI Matter Labs, 74 Rue du Faubourg Saint-Antoine, 75012 Paris, France
[4]DeepMind, 6 Pancras Square, London N1C 4AG, UK
[*]Corresponding author: n.burgess@ucl.ac.uk
[+]These authors contributed equally to this work

Acknowledgements

We thank Kevin Miller and Andrea Banino for helpful comments on the manuscript. This work was supported by the Gatsby Charitable Foundation, the Wellcome Trust and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 720270 Human Brain Project SGA1 and grant agreement No. 785907 Human Brain Project SGA2.

ABSTRACT

Behavioural and neural data suggest that multiple strategies and brain systems support learning about reward. Reinforcement learning (RL) considers these strategies in terms of model-based (MB) learning, which involves learning an explicit transition model of the available state-space, or model-free (MF) learning, which involves directly estimating the cumulative future reward associated with each state. The MB-MF distinction parallels that between "place learning" and "response learning" within spatial navigation. Lesion and inactivation studies suggests that dorsal hippocampus (dHC) underlies both MB and place learning, whereas dorsolateral striatum (DLS) underlies MF and response learning. Here, we present a computational model of the dHC and DLS contributions to reward learning that applies to spatial and nonspatial domains. In the model, hippocampal 'place cell' firing reflects the geodesic distance of other states from its 'preferred' state along the graph of task structure. Accordingly, this population can support a value function or 'goal cell' firing rate, via one-shot learning, on which gradient ascent corresponds to taking the shortest path on the graph, behaviour associated with MB planning. In contrast, the DLS learns through MF stimulus-response associations with egocentric environmental cues. We show that this model reproduces animal behaviour on spatial navigation tasks using the Morris Water Maze and the Plus Maze, and human behaviour on non-spatial two-step decision tasks. We discuss how the geodesic 'place cell' fields could be learnt, and how this type of representation helps to span the gap between MB and MF learning. The generality of our model, originally shaped by detailed constraints in the spatial literature, suggests that the hippocampal-striatal system is a general-purpose learning device that adaptively combines MB and MF mechanisms.

Keywords: Reinforcement Learning, Navigation, Hippocampus, Striatum

Introduction

Behavioural and neural studies suggest that animals can apply multiple strategies to the problem of maximising future reward, referred to as the reinforcement learning (RL) problem (Sutton and Barto, 1998). One solution is to build a model or map of the environment, that can be used to simulate the future to plan optimal actions (Tolman, 1948; Behrens et al., 2018; Muller et al., 1996). This model-based (MB) approach is powerful and flexible but computationally expensive (Russek et al., 2017). An alternative, model-free (MF) approach uses trial and error to estimate a mapping from the animal's state to its expected future reward, or "value" (Rescorla and Wagner, 1972; Sutton, 1988). This value computation is thought to be carried out in the brain through prediction errors signalled by phasic dopamine responses (Montague et al., 1996). While MF methods enable rapid action selection, these methods are considerably less flexible than MB and adapt poorly to environments with changing demands. While these two approaches are often contrasted, there are solutions that lie in between MF and MB (Dayan, 1993; Gustafson and Daw, 2011; Bengio et al., 2012), some of which have attracted attention in neuroscience recently (Russek et al., 2017; Stachenfeld et al., 2017; Momennejad et al., 2016; Gustafson and Daw, 2011). These algorithms learn to aggregate statistics over the environmental structure instead of performing expensive forward simulations, settling for intermediate flexibility at a lower computational cost.

In the spatial memory literature, the dominant dichotomy over learning strategies is the distinction between "response learning" and "place learning" navigation strategies (Chersi and Burgess, 2015). When navigating to a previously visited location, a response learning strategy involves learning a sequence of actions, each of which depends on the preceding action or a sensory cue (expressed in egocentric terms). For example, one might remember a sequence of left and right turns starting from a specific landmark. An alternative place learning strategy involves learning a flexible internal representation of the spatial layout of the environment (expressed in allocentric terms). This "cognitive map" can be found in the hippocampal formation, where there are neurons tuned to place and heading direction (O'Keefe and Nadel, 1978; Taube et al., 1990; Hafting et al., 2005). Spatial navigation using this map is considered more flexible because it can be used with arbitrary starting locations and destinations which need not be marked by immediate sensory cues.

We posit that the distinction between place and response learning is analogous to that between model-based and model-free RL (Poldrack and Packard, 2003). Indeed, there is evidence from both rodents (Packard and McGaugh, 1996; Packard, 1999; Mcdonald and White, 1994) and humans (Doeller et al., 2008; Doeller and Burgess, 2008) that spatial response learning relies on the same basal ganglia structures that support model-free RL. The equivalence between model-based reasoning and hippocampus-based navigation has been less clear – in rodents, multiple hippocampal lesion and inactivation studies failed to elicit an effect on action-outcome learning, a hallmark of model-based planning (Kimble and BreMiller, 1981; Kimble et al., 1982; Corbit and Balleine, 2000; Corbit et al., 2002; Ward-Robinson et al., 2001; Gaskin et al., 2005). However, there are indications that hippocampus might contribute to a different aspect of model-based RL, namely the representation of structure. Tasks that require memory of the relationships between stimuli do show dependence on hippocampus (Bunsey and Eichenbaum, 1996; Dusek and Eichenbaum, 1997; DeVito and Eichenbaum, 2011) and evidence from humans suggests hippocampus encodes these relationships (Schapiro et al., 2016; Garvert et al., 2017). Furthermore, the non-local "replay" of previously visited locations resembles the forward sweeps required for MB-style planning (Johnson and Redish, 2007; Pezzulo et al., 2014; Mattar and Daw, 2017; Foster and Knierim, 2012), suggesting hippocampus has a role in navigating these representations of structure.

Two recent studies have provided more direct evidence for hippocampal involvement in model-based planning as well as supporting the cognitive map. Firstly, Miller et al. (2017) trained rats on a two-step

decision task specifically designed to test model-based behaviour (Daw et al., 2011), and found that lesioning hippocampus caused deficits in model-based planning. In a similar vein, Vikbladh et al. (2018) trained healthy participants and patients on both the two-step decision task (Daw et al., 2011) and a spatial memory task (Doeller et al., 2008) and found that hippocampal damage impaired both model-based planning and place memory. The authors further noted that in control subjects, model-based planning and place memory covaried with one another.

This work formalizes computationally the perspective that hippocampal contributions to model-based learning and place learning are the same contribution, as are the dorsolateral striatal contributions to model-free and response learning. In our model, hippocampus supports flexible behaviour by encoding transition statistics in the state-space structure separately from reward information Gustafson and Daw (2011), while dorsolateral striatum (DLS) supports in associative reinforcement (Yin et al., 2005, 2004). We show that hippocampus and dorsolateral striatum can maintain these roles across multiple task domains, including a range of spatial and nonspatial tasks. Our model can explain quantitatively a range of key findings including place and response strategies in spatial navigation (Packard and McGaugh, 1996; Pearce et al., 1998; Doeller et al., 2008; Doeller and Burgess, 2008) and choices made by humans on non-spatial multi-step decision tasks (Doll et al., 2015; Daw et al., 2011).

Methods

Hippocampal and striatal systems for decision making

We implemented a model of hippocampal and dorsolateral striatal contributions to learning, shown in Figure 1. In our model, each system independently proposes an action and estimates its "value." Drawing from the standard RL framework, we define the value $Q_{s,a}$ of taking action a while being in state s to be the expected discounted cumulative return:

$$Q_{s,a} = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t) | s_0 = s, a_0 = a \right], \tag{1}$$

where s_0 and a_0 are the starting state and action at time $t = 0$, r is a reward function specifying the instantaneous reward found in each state, $\gamma \in [0,1]$ is a discount factor that upweights immediate rewards relative to distal rewards and $\pi(a|s)$ is the policy specifying a distribution over available actions given the current state. The objective of an RL agent is to discover an optimal policy π^* that will maximise value over all states.

Following earlier work in spatial RL (Chersi and Burgess, 2015, 2016; Dollé et al., 2010, 2018) two systems in our model estimate value using qualitatively different strategies, which can cause them to generate divergent predictions about which action is best. The dorsal striatal component uses a model-free temporal difference (TD) method (Sutton, 1988) to learn stimulus-response (S-R) associations directly from sensory inputs. The hippocampus has access to allocentric place information provided by place cells that fire at a specific location, together providing a cognitive map of the environment (O'Keefe and Nadel, 1978). This cognitive map provides an implicit model of the transition structure of the environment. In our model, it is used as a basis for value learning (Gustafson and Daw, 2011) through Hebbian association with "goal cells" that fire specifically at goal locations (Gauthier and Tank, 2018). Arbitration between the two systems was done by comparing the action values associated with the systems' respective outputs and selecting the higher-valued action. Although not modelled in detail here, we suggest this arbitration takes place in the medial prefrontal cortex (mPFC), following previous theoretical and experimental studies (Killcross and Coutureau, 2003; Daw et al., 2005).

In the next sections, we describe the hippocampal and striatal models and their interaction in detail, and test them by simulating several key spatial and non-spatial learning paradigms.

Striatal system

The dorsolateral striatum module was implemented as a model-free RL system that learned direct associations between sensory stimuli and actions. Striatal neurons coded for the value of each action, where actions were expressed as egocentric heading directions in the spatial navigation tasks and left or right button presses in the non-spatial tasks. Sensory input given to the model was coded by a set of transformed sensory input (or landmark) cells coding for the presence or absence of a landmark in a particular egocentric direction in the visual field of the agent, as well as for the distance of the landmark to the agent, in a manner analogous to the model presented in (Dollé et al., 2010, 2018). These neurons have Gaussian receptive fields centred at an angle in the agent's visual space, where the width of the Gaussian indicates the distance to the landmark:

$$u_i^{LC} = \exp\left(-\frac{\Delta\psi_i}{2(\sigma_{LC}/d(\boldsymbol{x}_a,\boldsymbol{x}_L))}\right), \tag{2}$$

where $\Delta\psi_i = \psi^L - \psi^{LC}$ is the angular distance between the direction of the landmark ψ^L and the preferred direction of the landmark cell ψ^{LC}, $d(\boldsymbol{x}_a,\boldsymbol{x}_L)$ is the distance from the agent to the landmark and σ_{LC} parameterises the width of the landmark cell's receptive field.

Neurons in the sensory layer project to neurons in the dorsal striatum in a fully connected one-layer network (see Figure 2):

$$u_a^{\mathrm{DLS}} = Q^{\mathrm{DLS}}(s,a) = \phi\left[\sum_{i=1}^{N} w_{i,a} u_i^{\mathrm{sensory}}(s)\right], \tag{3}$$

where u_a^{DLS} is the firing rate of the dorsolateral striatal neuron corresponding to striatal estimated value Q^{DLS} of action a given state s, ϕ the neuron's transfer function, which was chosen to be a sigmoid function $\phi(x) = \frac{e^x}{e^x+1}$, N the total number of sensory neurons, u_i^{sensory} the firing rate of sensory neuron i and $w_{i,a}$ the weight from sensory neuron i to striatal neuron a. The firing rate of the striatal neurons estimates the value of the possible actions associated with sensory states.

Learning in the striatal network happens through updating the weights in this network following a Q-learning rule (Watkins and Dayan, 1992). This allows the model to compute a temporal difference (TD) prediction error δ_t:

$$\delta_t = r_{t+1} + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t), \tag{4}$$

where r_{t+1} is the reward received at time $t+1$. This prediction error is then used to update the weights:

$$\Delta w_{i,a} = \alpha \delta_t e_{i,a}, \tag{5}$$

with learning rate α and eligibility trace $e_{i,a}$, which tracks which weights are eligible for updating. Every time step, the eligibility trace is updated according to the following rule:

$$e_{i,a}(t+1) = u_i^{sensory} u_j^{DLS} + \lambda e_{i,a}(t), \tag{6}$$

where λ is the trace decay parameter, controlling for how long synapses stay eligible for updating. Eligibility traces enable faster learning by making it possible to update weights that were active in the recent past instead of only the very last time step (Sutton and Barto, 1998).

Hippocampal system

The hippocampal component of our model was implemented with a single layer of rate-based place cells encoding the position of the agent in an allocentric reference frame. Applying earlier work in spatial RL from Gustafson and Daw (2011), we modelled the decline of place cell activity from the centre of the place field as a function of geodesic distance (along a path, around obstacles) rather than Euclidean distance (as the crow flies), and used these place cells as a basis for learning a function indicating the distance to the goal (Figure 2A). Since these geodesic place cells contain information about possible transitions among states in the task, associative learning over these state representations constitutes a middle ground between model-based and model-free strategies (Gustafson and Daw, 2011). It is worth noting that geodesic and Euclidean simulated cells yield qualitatively the same results in the spatial tasks we modelled later, since both capture a great deal of structural information in such tasks. However, geodesic distance permits us to apply our model to non-spatial domains, whereas Euclidean distance is not necessarily defined in these domains (see Discussion for further consideration on the relative considerations of different place cell models).

The response of hippocampal neurons is a bell-shaped Gaussian function of the geodesic (path) distance to the centre of their receptive field:

$$u_j^{\mathrm{PC}} = \exp\left(-\frac{d(\boldsymbol{x}, s_j)^2}{2\sigma_{\mathrm{PC}}^2}\right) \tag{7}$$

where $d(\boldsymbol{x}, s_j)$ is the approximate geodesic distance between the agent's position $\boldsymbol{x}$ and the place field centre $\boldsymbol{s}_j$ (see below for a detailed description of how this was computed) and σ_{PC} parameterizes the place field width. Place cell centres were randomly, uniformly distributed over the agent's environment.

Learning about goal locations in hippocampus was achieved using a qualitatively different mechanism from the striatum. Next to place cells, our model proposes the existence of "goal cells" that represent sensory attributes such as rewards regardless of location, and which can be associated to a location by synaptic connections with place cells to remember its location. This type of cell was predicted by early theoretical models (Burgess and O'Keefe, 1996; Burgess et al., 1994, 2001), and evidence for the existence of similar hippocampal reward-coding cells was recently reported (Gauthier and Tank, 2018). When the agent reached a goal location, the firing rate u^G of the goal cell was set to maximal, and connections between place cells and goal cell were updated using a Hebbian learning rule:

$$\Delta z_i = \eta u_i^{PC} u^G, \tag{8}$$

with learning rate η. When these connections are learnt, the firing rate map of the goal cell constitutes a goal function (Chersi and Burgess, 2015, 2016), the value of which indicates how far away the goal location is. When the goal is moved, the top of this goal function is reached without the goal cell being set to maximal and Hebbian unlearning happens, with a negative learning rate ρ.

$$\Delta z_i = -\rho u_i^{PC} u^G. \tag{9}$$

Whenever place cells drove a goal cell above a maximum firing rate, divisive normalisation was applied to ensure stability in the weights. Since this normalisation rule sets a maximum to the sum of weights, it ensures that unused associations are weakened when new associations are learnt (Oja, 1982). Figure 3 shows how the firing rate of a goal cell can represent this goal distance function for a continuous spatial environment (3A) and a discrete abstract state space (Doll et al., 2015) (3B). In both cases, the goal cell firing rate encodes the shortest path distance to the goal. Note that we implemented learning this way for biologically plausible reasons, but this amounts to MF learning over allocentric features.

The hippocampal module selects actions that maximise the increase in goal cell firing rate at the resulting state. The value of an action proposed by this subsystem was defined as the increase on this slope resulting from that action:

$$Q^{\mathrm{HC}}(s,a) = u^{G}(s) \tag{10}$$

The place cells encode path distance which corresponds to distance on the underlying task graph, independent of reward. Using these place cells as features for MF learning of a goal function results in a representation that contains the relevant information to navigate to a goal.

Although the learning rule described above is essentially MF, this system exhibits much of the flexibility that is usually associated with MB planning because MF learning is applied to features that capture the transition statistics of the environment, which constitute a sort of partial model or cognitive map. This approach therefore belongs to a class of semi-MB methods (see Russek et al., 2017, for discussion). Moreover, since the representation of task structure by place cells is decoupled from the representation of reward, when a new reward location is observed the agent can then quickly recompute the goal function. In contrast, a purely MF learner such as the striatal system described above would have to start over, because the only representation it has of the environment is the value function. Thus, the fact that this geodesic representation captures information about the graph transition structure means that some of the weaknesses associated to MF learning over egocentric features can be overcome. This flexibility is highlighted by the role of the hippocampal module in our simulations.

For spatial navigation tasks, using geodesic features for value learning prevents falsely generalising value over boundaries, which makes it a more efficient representation for learning (Gustafson and Daw, 2011; Mahadevan, 2005). For non-spatial multi-step decision tasks such as the one depicted in Figure 3, the geodesic features capture the structure of the graph. When a goal location is learnt, this means that the best action amounts to ascending this slope regardless of what state the agent started in when it first found the goal. In contrast, a MF learner will only learn to repeat actions from states that have led it to rewards before. Therefore, as we will show below, in our model an allocentric map-based representation of the environment can produce behaviour that is usually associated to MB planning, without the need for the computations normally associated to planning, such as tree search or dynamic programming. Note that when state transitions are non-deterministic, as in one of the tasks we consider, this does not affect the graph distance between two nodes.

We have not simulated how geodesic place representations are learnt, instead focussing on the downstream learning process. However, there are multiple ways in which a biological system could plausibly learn such representations. One way this could be achieved during exploration of the environment is by observing actual transitions, strengthening the connections between place fields that are nearby on the task graph (Muller et al., 1996). Note that such a mechanism would give rise to behaviour-dependent place fields, which are observed for example on a linear track (Sharp et al., 1996; Mehta et al., 2000). A particularly appealing feature of such type of learning is that it captures state transitions in general, whether those states are spatial or not, such as in the non-spatial two-step tasks discussed later. Similarly, the boundary vector cell model (Barry et al., 2006) produces place fields that respect the environment's boundaries in spatial environments. Another mechanism that could give rise to these place fields is the "arc length" cell proposed by Hasselmo (2007), which computes path distance using an oscillatory interference mechanism. Lastly, one can learn a geodesic distance matrix G, such that $G(s,s')$ is the is the geodesic distance between state s and state s', online with a simple off-policy update rule:

$$G_{t+1}(s,s') = \min[G_t(s_{t+1},s')+1, G_t(s_t,s')] \tag{11}$$

where G_t is the estimate of G and s_t is the state at time step t. For a detailed discussion of how geodesic firing fields can be learnt we refer the reader to Gustafson and Daw (2011).

Combining outputs from hippocampus and striatum

The striatal and hippocampal systems support qualitatively different algorithms for navigation, as outlined above. The hippocampal system provides an allocentric, map based representation whereas the striatum allows for stimulus-response associations. Our full model (Figure 1) incorporates both these systems. At every time step, both systems propose an action according to their own learning system, with a corresponding action value. Selection between these two actions was done by comparing the value of both proposed actions, weighted by a parameter ζ governing the trade-off between the two systems.

$$a^* = \begin{cases} a^{\text{DLS}}, & \text{if } Q^{\text{DLS}} > \zeta Q^{\text{HPC}} \\ a^{\text{HPC}}, & \text{otherwise} \end{cases} \tag{12}$$

The agent followed a ε-greedy policy with regard to either value function, meaning at every time step it took a random action with probability ε, or else took the action associated with the greatest increase in value. An additional free parameter governed the model's preference for either system.

Task simulation parameters

Hippocampal (η) and striatal (α) learning rate parameters, eligibility trace parameter λ and discount factor γ were hand-tuned for each task separately. A scalar parameter ζ governed the preference for striatal over hippocampal outputs in final action selection. This parameter was hand-tuned for each task separately. All other parameters were kept constant and described below. All parameter values used in simulations shown in the figures can be found in Table 1. Note that the learning rates were set much lower in the spatial environments because the environment's scale is much larger in these cases.

Spatial tasks

Agents moved at a constant swimming speed of $0.3m/s$ in the Morris Water Maze (Foster et al., 2000) and running speed of $0.7m/s$ in the Plus Maze. The Water Maze had a radius of $1m$ (Steele and Morris, 1999) and the Plus Maze was 1m long in both directions, with a width of 30 cm. Space was treated as continuous, but time was discretised into steps of $0.1s$. At every time step, the agent chose an action that corresponded to a new heading direction. Since rats cannot choose a new direction at the time resolution of this simulation, the change in heading direction was restricted to 60 degrees. The width of the sensory cell receptive field was set to 20 degrees.

Non-spatial tasks

In the non-spatial tasks, "visual input" to the striatum simply indicated state identity, and actions corresponded to choosing one of the two pictures shown on the screen as in Figure 9. This is a simple implementation of the fact that visual input uniquely identifies state in this task. Hippocampal place cells also coded for state identity, with their activity falling off with geodesic distance along the graph, without direction selectivity.

Geodesic transformation

To compute the place cell responses, the Euclidean coordinates in the spatial tasks discussed in this paper were transformed such that the distance between two points reflected the geodesic distance (i.e. the distance along the shortest walkable path, rather than how the crow flies). Following the approach of Gustafson and Daw (2011), we computed this transformation using the ISOMAP algorithm (Tenenbaum

et al., 2000) as implemented in scikit-learn (Pedregosa et al., 2011). In brief, the spatial environment was discretised into N points and an N-by-N matrix was generated containing the shortest path from each point to each other point, computed using the Floyd-Warshall algorithm (Floyd, 1962). Next, multidimensional scaling (MDS) was used to find a new set of 2D coordinates whose Euclidean distances reflect the geodesic distances in Euclidean space. Place cell responses where then computed using equation 7 with coordinates in geodesic embedding as input. Note that for the non-spatial tasks, geodesic place cell responses simply reflected distance on the task graph.

Results

In this section, we will describe a set of simulations and experimental results that are captured by the model in both spatial and non-spatial contexts.

Spatial decision making

Striatum but not hippocampus is sensitive to spatial blocking

A classical test of spatial memory for rodents is the Morris Water Maze (Morris, 1984). In this task, animals swim in a circular tank and are motivated to find a hidden platform that is submerged under opaque water where they can rest. Early experimental work has shown that this task is initially dependent on hippocampus, but becomes hippocampus independent after extensive training when animals start from the same location repeatedly (Morris et al., 1990). Figure 4 shows example trajectories of our model performing this task. Early in training, actions are mostly selected from hippocampus. However, since the agent always starts from the same place in this experiment, the striatal component is also sufficient to solve the task. Over training, the agent switches from a predominantly hippocampal strategy to a predominantly striatal strategy.

To tease apart the behavioural signatures of these different strategies, we next simulated a blocking experiment. Blocking is characteristic of learning rules that utilize a single prediction error variable to learn from multiple input signals (Rescorla and Wagner, 1972). Learning of one stimulus-response association hinders the learning of subsequent stimulus-responses because the prediction error becomes small, thereby reducing further weight updates (see Figure 5. In humans, spatial blocking has been shown to occur when learning locations relative to discrete landmarks, but not to boundaries (Doeller and Burgess, 2008). Furthermore, learning with respect to landmarks elicited activations in the dorsal striatum, whereas learning with respect to boundaries activated the posterior hippocampus (Doeller et al., 2008).

We aimed to capture these effects by looking at the behaviour of the hippocampal and striatal model systems separately, following a paradigm similar to Doeller and Burgess (2008) (see Figure 5): the agent navigated through a Morris Water Maze to find an unmarked platform submerged under opaque water, using a landmark to guide navigation. After 10 trials, a second landmark was added, and after 20 the first landmark was removed. As predicted by the Rescorla-Wagner rule, learning about landmark 2 was blocked by the prior learning about landmark 1, as evidenced by the drop in performance after the removal of landmark 1. The agent using its hippocampus to navigate was unaffected by the removal of landmark 2, showing incidental learning of location relative to both landmarks.

The effect of hippocampal lesions on adapted Water Maze navigation

An adaptation to the Morris Water Maze task involves putting an intra-maze landmark into the pool at a fixed offset from the platform, and moving both to a different location within the tank at the start of each block of four trials (Pearce et al., 1998, Figure 6A). In this version of the task, hippocampally lesioned animals perform *better* than healthy animals on the first trial of each session, because healthy animals initially linger at the previously rewarded goal location, see Figure 6C. In addition, learning over

sessions is relatively unimpaired. However, these animals show little intra-session learning. Note that the unimpaired learning across sessions in lesioned animals indicates that they are largely representing the goal location relative to the landmark, and this relationship does not change from session to session.

We simulated this task by comparing the performance of the full model with a model where we silenced the hippocampal component. As shown in figure 6, our model lesion accurately captures the effects of the original lesion experiments: within-session learning is impaired, between-session learning is unimpaired and healthy animals perform worse than hippocampally lesioned animals on the first trial after the platform has been moved. Looking at the trajectories (Figure 6C, right panel) reveals that the kind of mistake made by the agent is the same as was observed by the rat. The hippocampal spatial memory system guides the agent to the previously rewarded location. Only when it reaches that location and the reward is not there does it start unlearning the hippocampal goal representation. Since the striatal system learns about the location with respect to landmarks, the hippocampally lesioned agent can use the landmark to navigate directly to the correct location. Over the course of multiple sessions, the striatal agent starts dominating behaviour, driving lower first-trial escape times.

Animals switch to a response strategy on the Plus Maze

The distinct roles of the hippocampus and dorsal striatum have also been investigated using the place/response learning task (Packard and McGaugh, 1996; Packard, 1999). In this task, rats were trained to find a food reward from one arm of a Plus Maze, starting in the same arm every time, while the opposite arm is blocked (Figure 7). After training, a probe trial is performed in which the animal starts at the opposite end of the maze. If animals take the same egocentric turning direction as before, thus ending up at the opposite goal arm, their strategy is interpreted as response learning (relying on a remembered sequence of egocentric turns). If, on the other hand, they take the opposite turn to end up in the same goal arm, their strategy is interpreted as flexible place learning (relying on an allocentric representation of space).

Figure 8 shows the results of the original experiment and our simulations. Early in training most control rats (injected with saline) use a place strategy, but they switch to a response strategy after extensive training. Inactivation of the dorsal striatum with lidocaine prevented this switch. Inactivation of the hippocampus, by contrast, caused the response strategy to be used more often even early in training. These results indicate that the dorsal striatum underlies response learning, while the hippocampus underlies place learning. Indeed, when we simulated this task, we recapitulated the broad pattern of results found by Packard and McGaugh. Early in training, our full model shows a preference for actions proposed by the hippocampus, leading the agent to follow a place strategy. This is because the hippocampal goal representation is learnt fast in a one-shot update. Over the course of training, the slower incremental learning of the dorsal striatum takes over. Inactivation of the dorsal striatal and hippocampal components of the model can bias the agent to follow a place or response strategy, respectively.

One deviation from the original results is that lidocaine inactivation of the hippocampus caused only a half-way switch to a response strategy in rats early in training, whereas our model shows a bigger effect. A possible explanation is that that the hippocampus might not have been fully silenced by the amount of lidocaine injected. Our model could capture this by we only weakening the hippocampal output compared to that of the striatum.

Non-spatial decision making

Outside of the spatial domain, the distinction between model-free and model-based reinforcement learning has been heavily investigated using multi-step decision tasks. In this section we describe how the proposed model can be used to solve cognitive decision tasks. In particular we focus on two reference papers: the work of Doll et al. (2015) (see Figure 9A), and the work of Daw et al. (2011) (see Figure 9B).

Deterministic task

In the first experiment, human subjects where shown one of two sets of two pictures (faces or tools) and where asked to choose one picture. This first situation, of faces or tools on the screen, was defined as the start state. The subjects' initial choice deterministically controlled which of two second-stage states they would transition to. These second stage states corresponded to pictures of scenes or body parts (see Figure 9A)). Each second-stage option was either rewarded with money or not rewarded. The reward probability for each outcome drifted slowly and randomly such that subjects continuously learned by trial and error which sequence of choices was most likely to be rewarded. The total expected value of both scene and body part states was made equal to avoid inducing a bias. The first-stage options in both start states deterministically led to different outcomes: selecting one of the tools or one of the faces always led to the scenes, while the other tool or face always led to the body parts. This task structure allowed the authors to dissociate behaviour consistent with model-based and model-free learning approaches because model-free and model-based learners show different behaviours on this task. A model-based learner represents the states in terms of their transition probabilities, and uses this transition model to compute the best action. For this reason, when a model-based learner encounters a reward, this should affect its behaviour in the first-stage states regardless of whether the next trial starts in the same state as the previous trial (for example, faces followed by faces) or in a different one (for example, faces followed by tools). In contrast, a canonical model-free learner evaluates options in terms of the outcomes they have previously produced. Therefore, a model-free learner, upon receiving a reward, will only increase the probability of taking the same action in the next trial if that next trial starts in the same state as the previous one. Consistently with humans making use of both strategies, Doll and colleagues showed that human performance on this task lies somewhere in between these strategies (Figure 9B, rightmost panel).

Our model recapitulates the effects found by Doll and colleagues. When trained on this decision task, the hippocampal model mimics model-based behaviour by separating out reward information, as coded by the goal cell, from information about the transition structure. When a goal is reached, a value function is learnt in the weights between place and goal cells, as illustrated in Figure 3B. This value function falls off with (geodesic) distance along the graph. Thus, regardless of whether the next trial starts in the same or a different state, the hippocampal model will learn to take the same action again (Figure 9B). In contrast, the striatal learner learns separate action values for each state. Therefore, rewards obtained following one start state will not affect action values in the other start state (Figure 9B). Combining these two models gives a pattern of behaviour in between model-based and model-free, akin to human performance.

Non-deterministic task

A different version of the two-step decision task involves a single start state (Figure 9D). In this version of the task, left or right choices lead to different corresponding second-stage states with high probability (common transitions), but there is a small probability (rare transitions) that the agent transitions to the opposite state. For example, in Figure 9C, the left icon in the first (green) state usually leads to the choice in the pink state (common transition), but occasionally leads to the choice in the blue state (rare transition). In this task, model-free learners increase the likelihood of repeating their first-stage action following a reward, regardless of whether a common or rare transition was made. In contrast, a model-based learner uses its knowledge of the task's transition structure, such that rewards obtained after a rare transition have the opposite effect. The key finding in the original study was that human choices reflect both model-based and model-free influences (Figure 9E).

Our model recapitulates these findings. The model striatum, implementing a model-free RL system, increases stay probability after rewards regardless of whether a rare or common transition was made (Figure 9F). In contrast, the hippocampus uses geodesic distance to compute a value function over the

whole graph. When a goal state is reached and a reward is obtained, value is generalised over the graph according to graph distance. Therefore, on the next trial, the action is taken that will most likely lead to the state closest to the recent goal state. This encoding of the graph structure thus recapitulates true model-based behaviour. Combining the two systems results in behaviour that is similar to humans.

Discussion

We presented a model of hippocampal and dorsolateral striatal contributions to learning across domains that recapitulates experimental findings from both spatial navigation and non-spatial decision making. Our simulations support the view that the hippocampus serves both allocentric place learning and model-based decision making by supplying a map of the underlying structure of the environment environment, whereas the dorsolateral striatum underlies egocentric response learning and model-free learning. These results suggest that the hippocampal-striatal system is a general purpose learning device that adaptively combines model-based and model-free mechanisms.

The involvement of the hippocampus in abstract non-spatial tasks raises interesting questions about its role throughout evolution. It may be that the system evolved initially in the spatial domain but became recruited more generally (O'Keefe and Nadel, 1978), or that spatial decision making was always a sub-set of a more general ability (Eichenbaum et al., 1992). In the simulations presented here, we showed that an incidental Hebbian learning mechanism, proposed as a system for spatial memory (Burgess et al., 2001), can extend to "model-based" behaviour on a two-step decision task. These results might explain recent findings that model-based behaviour and spatial memory performance covary, and that hippocampal lesions impair both model-based behaviour and spatial memory (Vikbladh et al., 2018; Miller et al., 2017).

Our work follows several proposed models of spatial decision making by hippocampal and striatal systems (Chersi and Burgess, 2015; Dollé et al., 2010, 2018; Foster et al., 2000; Gustafson and Daw, 2011). Here we use this type of model, specified in the well-characterised spatial tasks and spatial neuronal responses, to address non-spatial decision making, to explore whether the same network could also solve non-spatial tasks. In the spatial domain, Dollé et al. (2010, 2018) used a similar hippocampo-striatal model and used it to explain behaviour on the adapted Water Maze task (Pearce et al., 1998) presented in Figure 6. Our model differs in a couple of important ways. Firstly, in their model place cells connected to "graph cells" that were used to construct an explicit topological graph of the spatial environment. A tree search algorithm was then used to explicitly plan a path through the environment in a model-based way. In the present model, by contrast, the topological structure of the environment is implied in the geodesic place cell fields. This allowed us to mimic true model-based behaviour by using MF learning over features of an allocentric map. Secondly, their model used another expert network that learned whether to take striatal or hippocampal outputs using TD learning.

It is important to note that the striatal model presented here reflects the function of the dorsolateral striatum specifically. Lesion and inactivation studies have shown that the dorsal striatum is functionally very heterogeneous (Yin and Knowlton, 2006). Lesions of the dorsomedial striatum (DMS), like those of the dorsal hippocampus shown in Figure 8, result in a switch to response strategies on the Cross Maze (Yin and Knowlton, 2004), and to cue based responding in the Water Maze, while the DLS underlies response learning (Devan et al., 1999). Furthermore, the DMS has been implicated in learning action-outcome contingencies outside the spatial domain (Yin et al., 2005; Yin and Knowlton, 2006). Anatomical connectivity supports this functional dissociation in the dorsal striatum (Yin and Knowlton, 2006; Devan and White, 1999). Whereas the DLS receives inputs mostly from sensorimotor cortex and dopaminergic input from the substantia nigra, the DMS receives input from several meso and allocortical areas including the hippocampus. Indeed, cells encoding route and heading direction have been found in the DMS (Mulder

et al., 2004; Ragozzino et al., 2001). It is therefore likely that the dorsal hippocampus and the DMS are part of a single circuit involved in flexible goal-directed decision making, whereby the hippocampus provides map-based information, and the DMS is involved in action selection.

The "goal cells" used in this model have been proposed by earlier theoretical models (Burgess and O'Keefe, 1996; Burgess et al., 2001). As noted in this earlier work, these are expected to be found downstream of the place cells. For example, in the subiculum or in the ventral striatum, which receives strong inputs from hippocampus and is involved in action selection (Yin and Knowlton, 2006). Finally, they might be located in the hippocampus itself, where "reward cells" with similar properties have recently been found (Gauthier and Tank, 2018).

In the present work, we have chosen for a very simple arbitration mechanism between the striatal and hippocampal outputs based on the slope of a goal function that was controlled by a parameter weighting the hippocampal and striatal outputs, but this is by no means the only way such an arbitration could be implemented. For instance, Dollé et al. (2010, 2018) used a TD learning algorithm learning to select the appropriate outputs of a response learning expert network and a place learning expert network. In contrast, Daw et al. (2005) proposed a Bayesian mechanism for arbitrating between MB outputs originating in the PFC and MF outputs originating in the DLS, based on the uncertainty associated with these outputs. These mechanisms all predicted a gradual switch to habitual learning (i.e. selecting striatal outputs) over time. Future work should find out whether animals arbitrate between systems on the basis of uncertainty or estimated value by designing experiment that specifically decorrelates the two. For example, by considering a task with rapidly changing reward contingencies. A MB strategy will have lower uncertainty about its value estimate (Daw et al., 2005). If the MB strategy is still chosen even if the associated value estimate is lower, that is evidence for the uncertainty based competition mechanism.

As shown in table 1, the values of the striatal learning rate α chosen in different tasks differ substantially, especially between the spatial and non-spatial experiments. The reason for this is that the nature of the tasks is qualitatively different. In the spatial navigation tasks, agents explored a continuous state space, where an action is chosen at every time step, resulting in many actions per trial. The striatal value prediction is computed after every action, which means a low learning rate causes more stable learning. By contrast, agents took only two actions per trial on the non-spatial tasks, requiring a high learning rate to learn efficiently within a few trials. We have defined all actions at the same level of abstraction for simplicity but in the brain, it is likely that these actions are defined at a higher level of abstraction than the moment-by-moment choice of heading direction. Hierarchical reinforcement learning provides a principled way of dealing with actions defined at different levels of abstraction (Botvinick et al., 2009; Solway et al., 2014).

There are a few reasons for modelling place cells with geodesic fields. Firstly, unlike real place cells, Euclidean place cells do not respect the boundaries of an environment, whereas geodesic place cells do (Gustafson and Daw, 2011). Because of this, geodesic place cells as basis functions allow for more efficient downstream reinforcement learning. Another reason is that hippocampal place-like coding has been found representing non-spatial variables such as sound frequency (Aronov et al., 2017), time (Manns et al., 2007) and abstract states (Miller et al., 2017), in line with the view of the hippocampal cognitive map as a general representation of relationships between cognitive entities (Constantinescu et al., 2016; Buzsáki and Moser, 2013; Behrens et al., 2018). Unlike Euclidean distance, geodesic distance is a meaningful measure in all these state spaces because it describes the distance along a path in a graph regardless of the variables represented by that graph. This makes gradient ascent on the goal cell rate equivalent to following the shortest path in the graph of the task's state space.

While in our implementation fully formed idealised geodesic place cells were provided, geodesic distance can be estimated in a biologically plausible manner through the weight strength in a connected

network of place cells via a Hebbian learning rule (Muller et al., 1996). Alternatively, geodesic distance might be estimated by "arc-length" cells (Hasselmo, 2007) that compute path length using an oscillatory interference mechanism. A Hebbian mechanism such as proposed by Muller et al. (1996) would naturally extend to non-spatial tasks because it is based on cells firing closely together in time. In this sense, it is very similar to a recently proposed model with very similar properties to ours proposes that place cells encode the discounted expected future occupancy of surrounding states (Stachenfeld et al., 2017). This successor representation (SR; Dayan, 1993) differs from the idealised geodesic model in that it is policy-dependent, meaning that the shape of the place fields will depend on the behaviour of the agent. The SR can be estimated using a temporal difference rule. In the spatial domain, Euclidean and geodesic distance are equivalent for open field environments. They differ for environments with obstacles or walls such as the Plus Maze (see Figure 10), although not so substantially that they lead to different behavioural predictions in the tasks considered here. Under a random policy, the SR place field is nearly equivalent to the idealised geodesic representation and indeed the results for the tasks simulated here would be no different. Therefore, the results in this paper do not differentiate between these models (but see Stachenfeld et al., 2017). One potential problem with experience dependent place representations is that models might be incapable of generating novel shortcuts if those shortcuts were not present in the previous policy. A potential role for grid cells is to alleviate this problem by providing a more Euclidean representation of space, allowing shortcuts across unvisited states (Bush et al., 2015; Banino et al., 2018; Bicanski and Burgess, 2018), although it has also been proposed that grid cells form a low-dimensional basis set for the place cell representation (Dordek et al., 2016; Stachenfeld et al., 2017).

The hippocampal model presented here encodes the topological structure of the environment, by making firing rate depend on the length of the shortest path to the field centre (Gustafson and Daw, 2011). By encoding this structure of the environment and separating it from a reward representation, it occupies a space in between model-free and model-based methods, and like the SR (Russek et al., 2017; Stachenfeld et al., 2017), it can mimic truly model-based behaviour in cases where the distance to the goal is close enough (Figure 9). However, full model-based planning classically refers to the way that a model is used to plan using tree search or dynamic programming (Daw et al., 2005; Dollé et al., 2018). An interesting body of work focuses on the replay of hippocampal sequences during sharp wave ripples as a potential neural substrate of this process (Johnson and Redish, 2007; Pezzulo et al., 2014; Mattar and Daw, 2017; Foster and Knierim, 2012).

The fact that, as shown in this paper, MF learning over a map-based representation can reproduce many behavioural effects generally attributed to MB planning, highlights the importance of finding a good representation of a task: an appropriate representation can alleviate the need for expensive computation (Dayan, 1993; Bengio et al., 2012). In addition, it might provide an explanation for results found in lesion studies. If hippocampus provides a structural model of the environment, and DMS supports computations over this model, this would explain why lesions to both areas impair MB planning (Miller et al., 2017; Yin et al., 2005; Yin and Knowlton, 2006).

In conclusion, dorsal hippocampus and DLS support qualitatively different strategies for learning about reward. This becomes apparent in spatial as well as non-spatial contexts, as illustrated by the model presented here. The fact that the same model explains behaviour in both types of task implies that the hippocampal-striatal system is a general purpose learning device that adaptively combines model-based and model-free mechanisms.

References

Aronov, D., Nevers, R., and Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, 543(7647):719–722.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M. J., Degris, T., Modayil, J., Wayne, G., Soyer, H., Viola, F., Zhang, B., Goroshin, R., Rabinowitz, N., Pascanu, R., Beattie, C., Petersen, S., Sadik, A., Gaffney, S., and King, H. (2018). Vector-based navigation using grid-like Representations in Artificial Agents. *Nature*, 26.

Barry, C., Lever, C., Hayman, R., Hartley, T., Burton, S., O'Keefe, J., Jeffery, K., and Burgess, N. (2006). The boundary vector cell model of place cell firing and spatial memory. *Reviews in the Neurosciences*, 17(1-2):71–98.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A., Stacenhfeld, K. L., and Kurth-Nelson, Z. (2018). What is a cognitive map? Organising knowledge for flexible behaviour. *BioRxiv*.

Bengio, Y., Courville, A., and Vincent, P. (2012). Representation Learning: A Review and New Perspectives. *arXiv*, (1993):1–30.

Bicanski, A. and Burgess, N. (2018). A Neural Level Model of Spatial Memory and Imagery. *eLife*, 436(7052):1–3.

Botvinick, M. M., Niv, Y., and Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280.

Bunsey, M. and Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature*, 379(6562):255.

Burgess, N., Becker, S., King, J. A., and O'Keefe, J. (2001). Memory for events and their spatial context: Models and experiments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356(1413):1493–1503.

Burgess, N. and O'Keefe, J. (1996). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus*, 6(6):749–762.

Burgess, N., Recce, M., and O'Keefe, J. (1994). 1994 Special Issue: A model of hippocampal function. *Neural Netw.*, 7(6-7):1065–1081.

Bush, D., Barry, C., Manson, D., and Burgess, N. (2015). Using Grid Cells for Navigation. *Neuron*, 87(3):507–520.

Buzsáki, G. and Moser, E. I. (2013). Memory, navigation and theta rhythm in the hippocampal-entorhinal system. *Nature Neuroscience*, 16(2):130–138.

Chersi, F. and Burgess, N. (2015). The Cognitive Architecture of Spatial Navigation: Hippocampal and Striatal Contributions. *Neuron*, 88(1):64–77.

Chersi, F. and Burgess, N. (2016). Hippocampal and striatal involvement in cognitive tasks : a computational model. In *Proceedings of the 6th International Conference on Memory, ICOM16*, pages 24–28.

Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.

Corbit, L. H. and Balleine, B. W. (2000). The role of the hippocampus in instrumental conditioning. *Journal of Neuroscience*, 20(11):4233–4239.

Corbit, L. H., Ostlund, S. B., and Balleine, B. W. (2002). Sensitivity to instrumental contingency degradation is mediated by the entorhinal cortex and its efferents via the dorsal hippocampus. *Journal of Neuroscience*, 22(24):10976–10984.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.

Dayan, P. (1993). Improving Generalisation for Temporal Difference Learning: The Successor Representation. *Neural Computation*, 5(4):613–624.

Dayan, P. and Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective and Behavioral Neuroscience*, 8(4):429–453.

Devan, B. D., McDonald, R. J., and White, N. M. (1999). Effects of medial and lateral caudate-putamen lesions on place- and cue- guided behaviors in the water maze: Relation to thigmotaxis. *Behavioural Brain Research*, 100(1-2):5–14.

Devan, B. D. and White, N. M. (1999). Parallel information processing in the dorsal striatum: relation to hippocampal function. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 19(7):2789–2798.

DeVito, L. M. and Eichenbaum, H. (2011). Memory for the order of events in specific sequences: contributions of the hippocampus and medial prefrontal cortex. *Journal of Neuroscience*, 31(9):3169–3175.

Doeller, C. F. and Burgess, N. (2008). Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proceedings of the National Academy of Sciences*, 105(15):5909–5914.

Doeller, C. F., King, J. A., and Burgess, N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proceedings of the National Academy of Sciences*, 105(15):5915–5920.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., and Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, 18(5):767–772.

Dollé, L., Chavarriaga, R., Guillot, A., and Khamassi, M. (2018). Interactions of spatial strategies producing generalization gradient and blocking: A computational approach. *PLoS Computational Biology*, 14(4):1–35.

Dollé, L., Sheynikhovich, D., Girard, B., Chavarriaga, R., and Guillot, A. (2010). Path planning versus cue responding: A bio-inspired model of switching between navigation strategies. *Biological Cybernetics*, 103(4):299–317.

Dordek, Y., Soudry, D., Meir, R., and Derdikman, D. (2016). Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife*, 5(MARCH2016):1–36.

Dusek, J. A. and Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences*, 94(13):7109–7114.

Eichenbaum, H., Otto, T., and Cohen, N. J. (1992). The hippocampus: What does it do? *Behavioral & Neural Biology*, 57(1):2–36.

Floyd, R. W. (1962). Algorithm 97: Shortest path. *Communications of the ACM*, 5(6):345.

Foster, D. J. and Knierim, J. J. (2012). Sequence Learning and the Role of the Hippocampus in Rodent Navigation. *Current opinion in neurobiology*, 22(2):294–300.

Foster, D. J., Morris, R. G., and Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, 10(1):1–16.

Garvert, M. M., Dolan, R. J., and Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife*, 6:1–20.

Gaskin, S., Chai, S., and White, N. M. (2005). Inactivation of the dorsal hippocampus does not affect learning during exploration of a novel environment. *Hippocampus*, 15(8):1085–1093.

Gauthier, J. L. and Tank, D. W. (2018). A Dedicated Population for Reward Coding in the Hippocampus. *Neuron*, 0(0):179–193.

Gustafson, N. J. and Daw, N. D. (2011). Grid cells, place cells, and geodesic generalization for spatial

reinforcement learning. *PLoS Computational Biology*, 7(10).

Hafting, T., Fyhn, M., Molden, S., Moser, M., and Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806.

Hasselmo, M. E. (2007). Arc length coding by interference of theta frequency oscillations may underlie context-dependent hippocampal unit data and episodic memory function. *Learning and Memory*, 14(11):782–794.

Johnson, A. and Redish, A. D. (2007). Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point. *Journal of Neuroscience*, 27(45):12176–12189.

Killcross, S. and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, 13(4):400–408.

Kimble, D. P. and BreMiller, R. (1981). Latent learning in hippocampal-lesioned rats. *Physiology & behavior*, 26(6):1055–1059.

Kimble, D. P., Jordan, W. P., and BreMiller, R. (1982). Further evidence for latent learning in hippocampal-lesioned rats. *Physiology & behavior*, 29(3):401–407.

Mahadevan, S. (2005). Proto-value functions.

Manns, J. R., Howard, M. W., and Eichenbaum, H. (2007). Gradual Changes in Hippocampal Activity Support Remembering the Order of Events. *Neuron*, 56(3):530–540.

Mattar, M. G. and Daw, N. D. (2017). Prioritized memory access explains planning and hippocampal replay. *bioRxiv*, pages 0–17.

Mcdonald, R. J. and White, N. M. (1994). Parallel Information Processing in the Water Maze : Evidence for Independent Memory Systems Involving Dorsal Striatum and Hippocampus. *Behavioral and Neural Biology*, 270:260–270.

Mehta, M. R., Quirk, M. C., and Wilson, M. A. (2000). Experience-Dependent Asymmetric Shape of Hippocampal Receptive Fields. 25:707–715.

Miller, K. J., Botvinick, M. M., and Brody, C. D. (2017). Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, 20(9):1269–1276.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N., and Gershman, S. J. (2016). The successor representation in human reinforcement learning. *bioRxiv*, pages 1–27.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of neuroscience*, 16(5):1936–1947.

Morris, R. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods*, 11(1):47–60.

Morris, R. G., Schenk, F., Tweedie, F., and Jarrard, L. E. (1990). Ibotenate Lesions of Hippocampus and/or Subiculum: Dissociating Components of Allocentric Spatial Learning. *European Journal of Neuroscience*, 2(12):1016–1028.

Mulder, A. B., Tabuchi, E., and Wiener, S. I. (2004). Neurons in hippocampal afferent zones of rat striatum parse routes into multi-pace segments during maze navigation. *European Journal of Neuroscience*, 19(7):1923–1932.

Muller, R. U., Stead, M., and Pach, J. (1996). The Hippocampus as a Cognitive Graph. *Journal of General Physiology*, 107(6):663–694.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273.

O'Keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

Packard, M. G. (1999). Glutamate infused posttraining into the hippocampus or caudate-putamen differentially strengthens place and response learning. *Proceedings of the National Academy of Sciences of the United States of America*, 96(22):12881–12886.

Packard, M. G. and McGaugh, J. L. (1996). Inactivation of Hippocampus or Caudate Nucleus with Lidocaine Differentially Affects Expression of Place and Response Learning. *Neurobiology of Learning and Memory*, 72(0007):65–72.

Pearce, J. M., Roberts, A. D. L., and Good, M. (1998). Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors. *Nature*, 62(1989):1997–1999.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

Pezzulo, G., van der Meer, M. A., Lansink, C. S., and Pennartz, C. M. (2014). Internally generated sequences in learning and executing goal-directed behavior. *Trends in Cognitive Sciences*, 18(12):647–657.

Poldrack, R. and Packard, M. (2003). Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia*, 41:245–251.

Ragozzino, K., Leutgeb, S., and Mizumori, S. (2001). Dorsal striatal head direction and hippocampal place representations during spatial navigation. *Experimental Brain Research*, 139(3):372–376.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, 2:64–99.

Russek, E. M., Momennejad, I., Botvinick, M. M., and Gershman, S. J. (2017). Predictive representations can link model - based reinforcement learning to model - free mechanisms. *PLoS Computational Biology*, (September):1–42.

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., and Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, 26(1):3–8.

Sharp, P. E., Blair, H. T., and Brown, M. (1996). Neural network modeling of the hippocampal formation spatial signals and their possible role in navigation: A modular approach. *Hippocampus*, 6(6):720–734.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., and Botvinick, M. M. (2014). Optimal Behavioral Hierarchy. *PLoS Computational Biology*, 10(8).

Stachenfeld, K. L., Botvinick, M. M., and Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653.

Steele, R. J. and Morris, R. G. M. (1999). Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the NMDA-antagonist D-AP5. *Hippocampus*, 9(2):118–136.

Sutton, R. S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, pages 9–44.

Sutton, R. S. and Barto, a. G. (1998). Reinforcement learning: an introduction. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 9(5):1054.

Taube, J. S., Muller, R. U., and Ranck, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience*, 10(2):420–435.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for non linear dimensionality reduction. *Science*, 290(December):2319–2323.

Tolman, E. C. (1948). Cognitive Maps in Rats and Man. *Psychological Review*, 55(4):189–208.

Vikbladh, O., Michael, M., John, K., Blackmon, K., Orrin, D., Daphna, S., Burgess, N., and Daw, N. D. (2018). Two Sides of the Same Coin: The Hippocampus as a Common Neural Substrate for Model-Based Planning and Spatial Memory. *bioRxiv*.

Ward-Robinson, J., Coutureau, E., Good, M., Honey, R. C., Killcross, A. S., and Oswald, C. J. P. (2001). Excitotoxic lesions of the hippocampus leave sensory preconditioning intact: Implications for models of hippocampal functioning. *Behavioral neuroscience*, 115(6):1357.

Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3-4):279–292.

Yin, H. H. and Knowlton, B. J. (2004). Contributions of striatal subregions to place and response learning. *Learning and Memory*, 11(4):459–463.

Yin, H. H. and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6):464–476.

Yin, H. H., Knowlton, B. J., and Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, 19(October 2003):181–189.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., and Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, 22(2):513–523.

	ε	η	ρ	α	λ	γ	ζ
Water Maze	0.1	0.5	.05	0.001	0.76	0.98	5
Plus Maze	0.1	0.5	.05	0.0008	0.76	0.98	0.5
DawTask	0.1	0.5	.05	0.3	0.9	0.98	1
Dolltask	0.2	0.5	.05	0.8	0.8	0.98	1

Table 1. Parameters used in the different simulations environments: randomness parameter ε, hippocampal learning rate η, hippocampal forgetting rate ρ, striatal learning rate α, trace decay parameter λ, discount parameter γ and trade-off parameter ζ.

Supplementary materials

Reinforcement learning

We consider the Reinforcement Learning (RL) framework: an agent interacts with an environment in a Markov Decision Process consisting of a set of states $\mathbb{S}$, a set of actions $\mathbb{A}$, a next-state distribution $p(s'|s,a)$ giving the probability of transitioning from s to s' when taking action a, a reward function $r(s)$ specifying the reward found in state s and a discount factor $\gamma \in [0,1]$ down-weighting distal rewards. The agent chooses actions according to a policy $\pi(a|s)$. The agent's goal is to maximise over value, defined as the discounted expected total future reward:

$$V(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)|s_0 = s\right] \tag{13}$$

where s_t is the state visited at time t. As described in the Results section, the striatal component of the model learned value through Q-learning (Watkins and Dayan, 1992) and the hippocampal model uses Hebbian learning with a goal cell to learn geodesic distance to the goal. To construct this place cell matrix the model implicitly uses a state transition model, describing the transition probabilities from state to state:

$$T(s,s') = \sum_{a} \pi(a|s)p(s'|s,a) \tag{14}$$

The hippocampal model therefore occupies a space in between pure model-based and pure model-free (see also Gustafson and Daw, 2011; Dayan, 1993; Stachenfeld et al., 2017; Russek et al., 2017; Dayan and Daw, 2008). Agents followed an ε-greedy policy with respect to the value function.

Figures

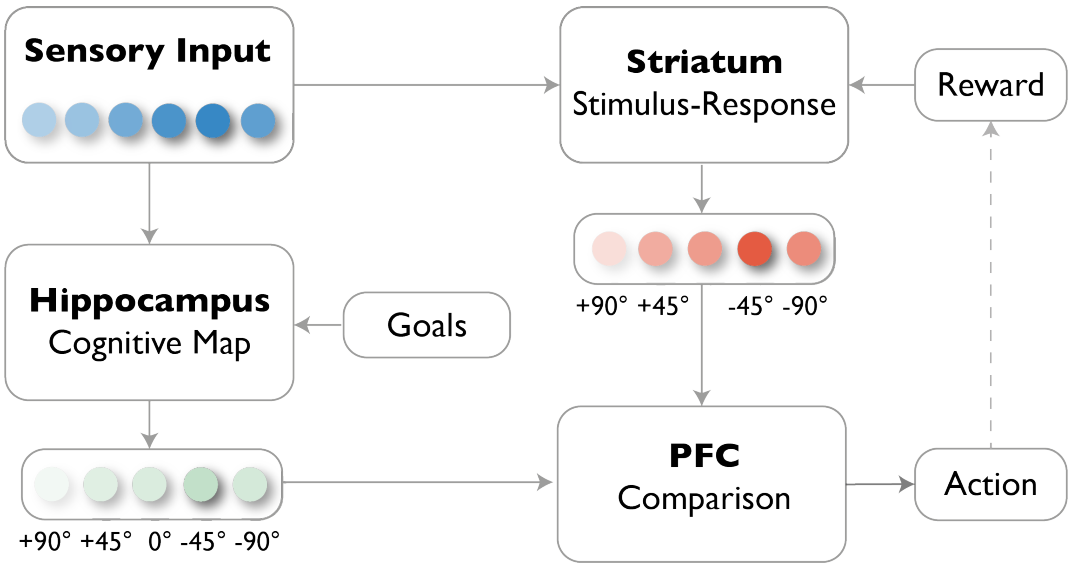

Figure 1. Schematic representation of the hippocampal-striatal circuit that guides spatial navigation (Chersi and Burgess, 2015). The hippocampus provides a "cognitive map," with information about locations and their adjacency for goal-directed decision making, and the dorsolateral striatum learns stimulus-response associations. Each area outputs its estimated optimal action (e.g. the turning angle for spatial tasks), which is then compared and selected by the prefrontal cortex.

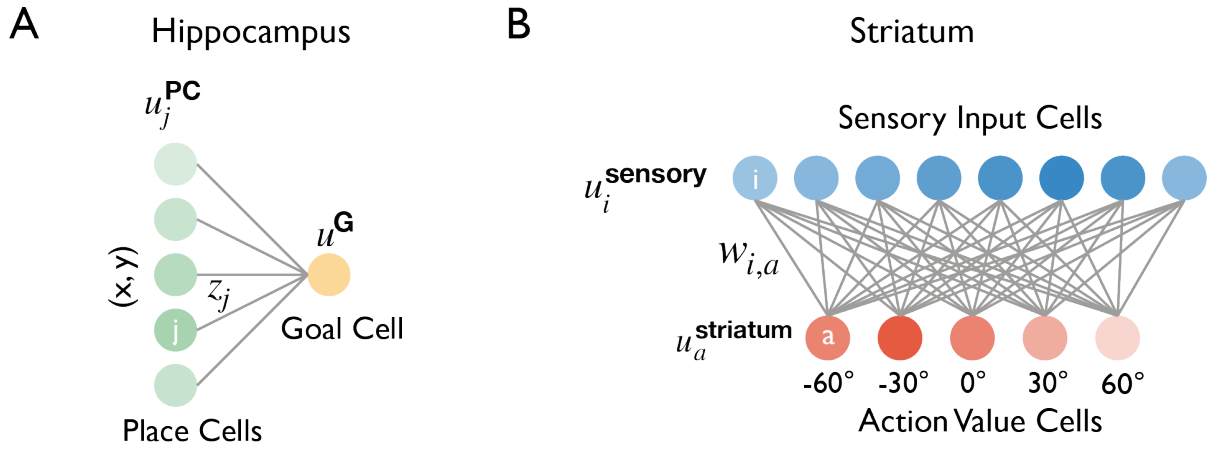

Figure 2. Illustration of hippocampal and striatal model components. (A) The hippocampal component of the model consists of place cells coding for the agent's position in space, and a goal cell encodes vicinity to the goal. Actions that maximise the gradient of the goal cell rate are selected. (B) In the striatal model, sensory input cells code for the presence or absence of a landmark in a specific turning direction. These project to neurons in the dorsolateral striatum coding for the cumulative discounted future reward if that action is taken in the current state.

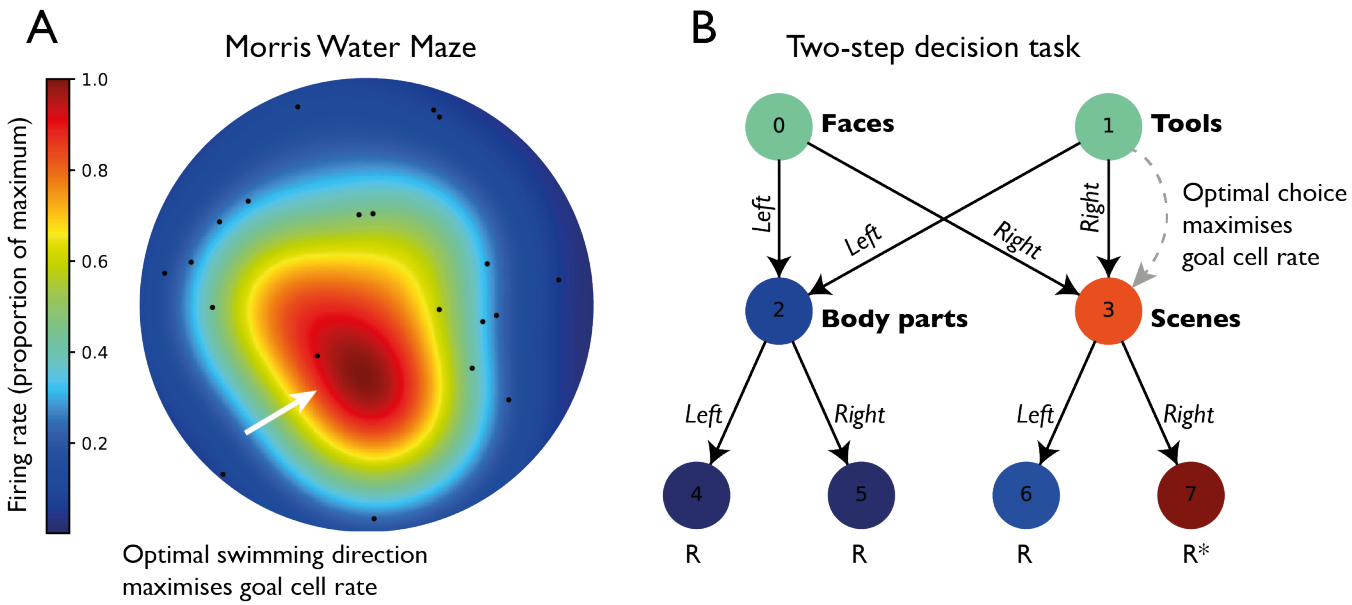

Figure 3. Goal cell firing maps encode a value function that can be used for navigating to a goal. When a goal is reached, connections between place and goal cells are strengthened. Navigating back to this goal is then achieved by ascending the gradient of this firing field. (A) Goal cell firing rate across a continuous spatial environment (the Morris Water Maze). Black dots show place field centres that were uniformly distributed over the environment. (B) Goal cell firing rate over the discrete non-spatial state space of the multi-step decision task shown in Figure 9B.

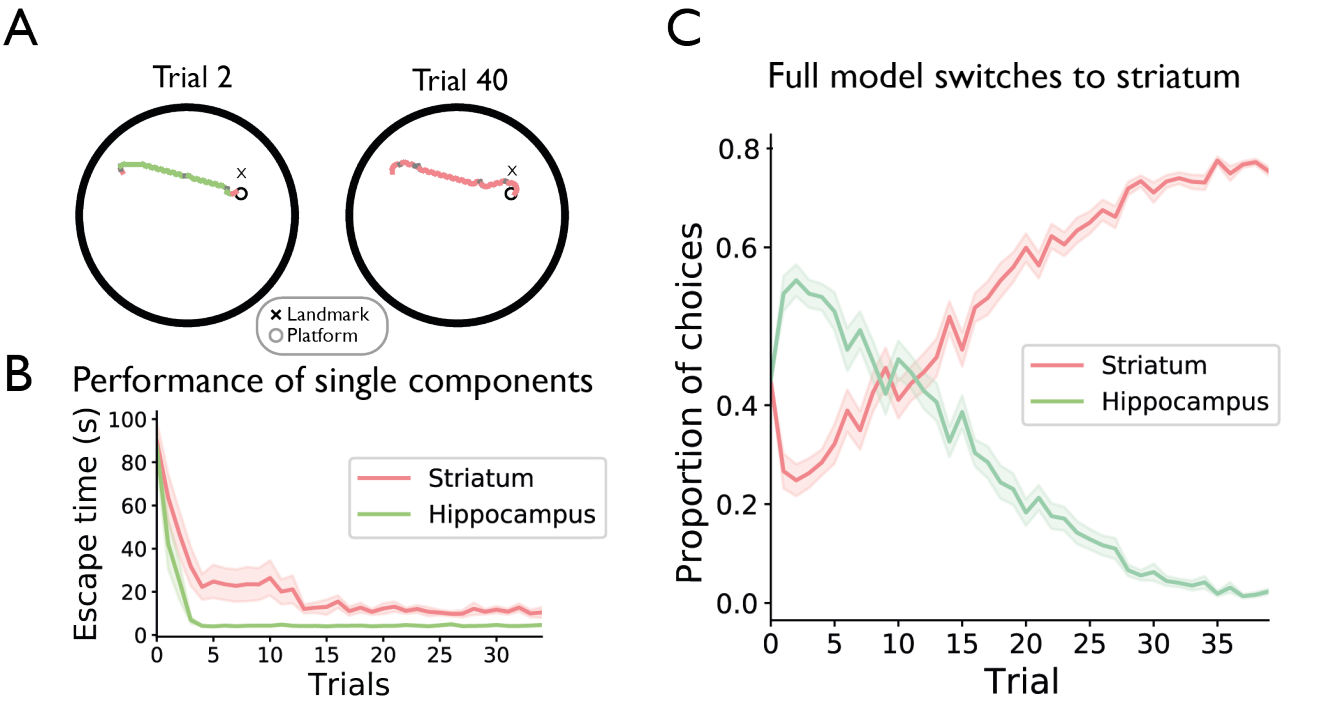

Figure 4. Simulated navigation in the Morris Water Maze. In this experiment, the agent started from the same place in every trial, and a landmark was placed at a fixed distance from it. Both model systems can learn this task individually, but once striatum reliably predicts actions with values comparable to hippocampus, it overtakes hippocampus. (A) Example trajectories, where each step is colour-coded by whether the action originated in the hippocampus (green), striatum (red) or whether it was a random action (grey). (B) Escape times (time until the agent finds the hidden platform) decrease for both the DLS and dHC components, indicating that both can learn the task. (C) Mean proportion of choices originating from both systems as a function of the number of trials trained in the same starting position.

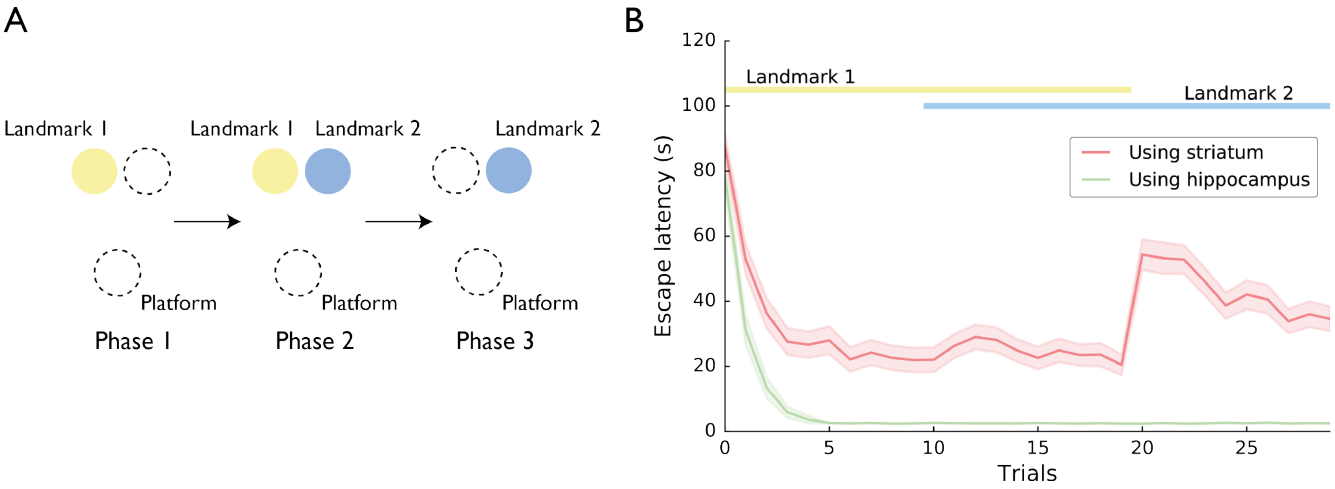

Figure 5. Illustration of the blocking experiment. (A) Schematic of experimental phases. In phase 1, a single landmark is present. In phase 2, a second landmark is added, and in phase 3 the first landmark is removed. Associative reinforcement would predict an increase in escape latency (time it takes to find the platform) upon removal of the first landmark. (B) Escape latency as a function of trial number. Yellow and blue horizontal lines indicate the presence of landmarks 1 and 2, respectively.

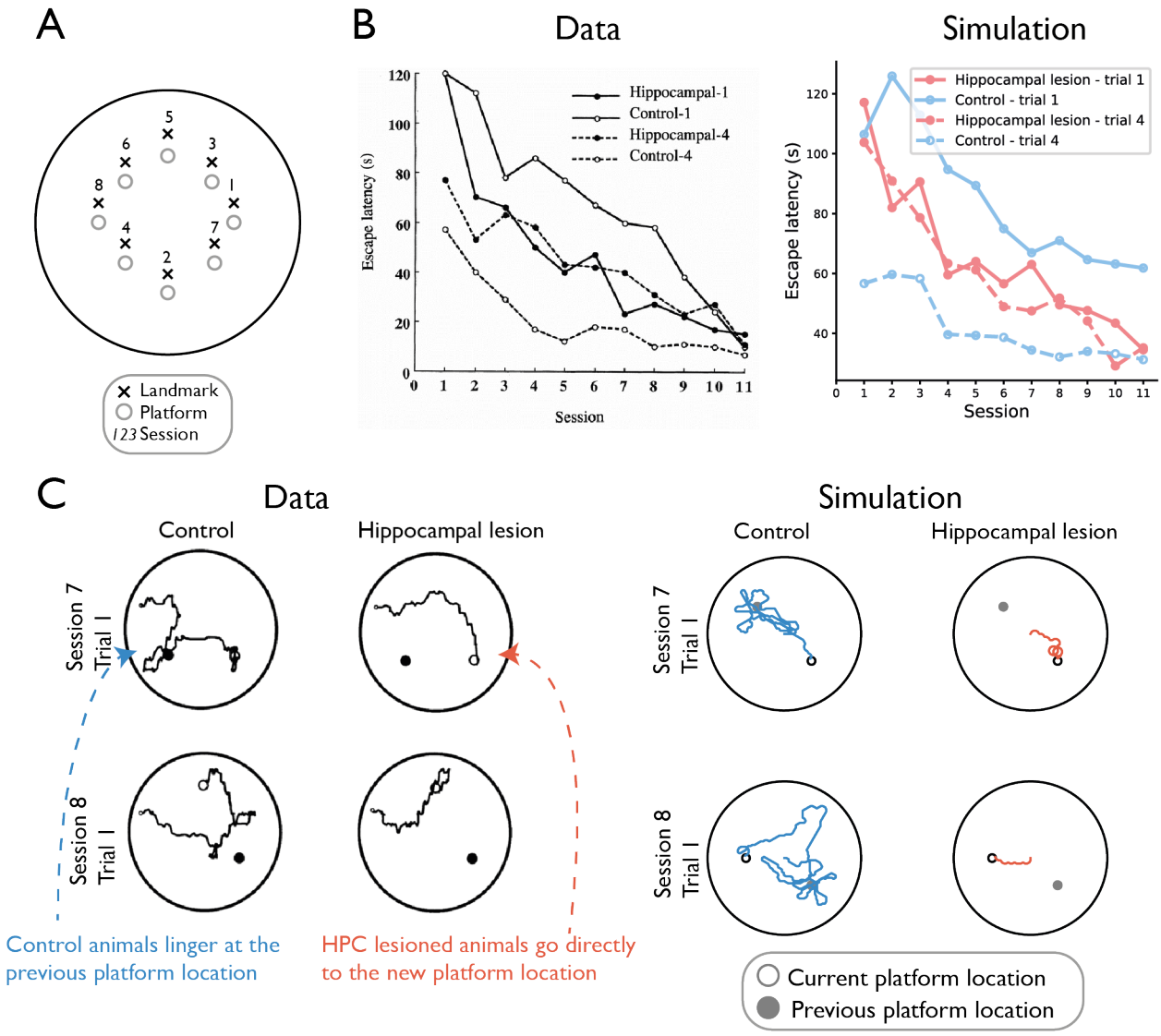

Figure 6. Results and simulations of the experiment described in Pearce et al. (1998). Sessions lasted 4 trials, and platform and landmark were moved at the beginning of each session. (A) : Possible locations of the hidden platform (white circle) and the corresponding landmark (crosses). Numbers indicate in which session this was the location. (B) Escape latency in the water maze for hippocampal lesioned and control animals (left) and agents (right) on trial 1 (solid lines) and 4 (dashed line) of each session. Hippocampal damage impairs intra-trial learning but preserves learning over trials. Because animals with hippocampal damage follow a response strategy based on egocentric visual input, their performance on the first trial of each session is better than that of healthy controls. (C) Example trajectories from the first trials of sessions 7 and 8. Animals (left panel, adapted from Pearce et al., 1998) and agents (right panel) using a hippocampal place strategy tend to wander around the previous platform location.

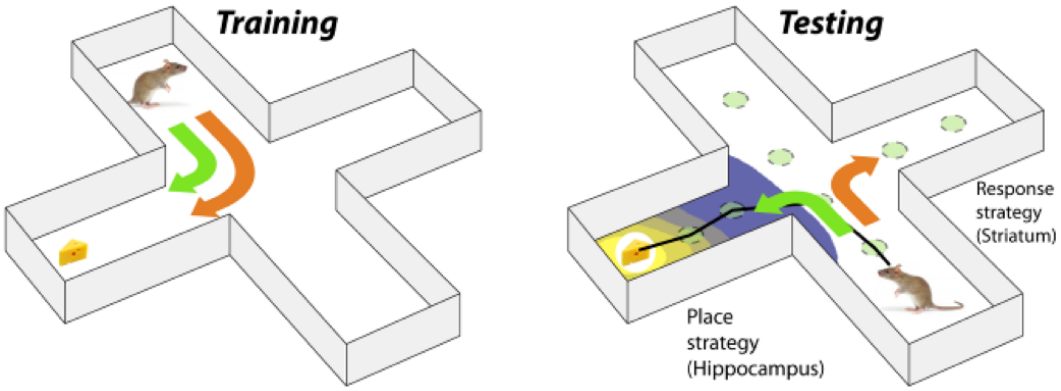

Figure 7. Representation of the Plus Maze used by Packard and McGaugh (1996). Left panel: rats learn to find the food placed at one end of an arm from the start location. Right panel: in unrewarded probe trials, the starting position is moved to the opposite side of the maze and the control and treated rats are tested for their ability in finding the correct location

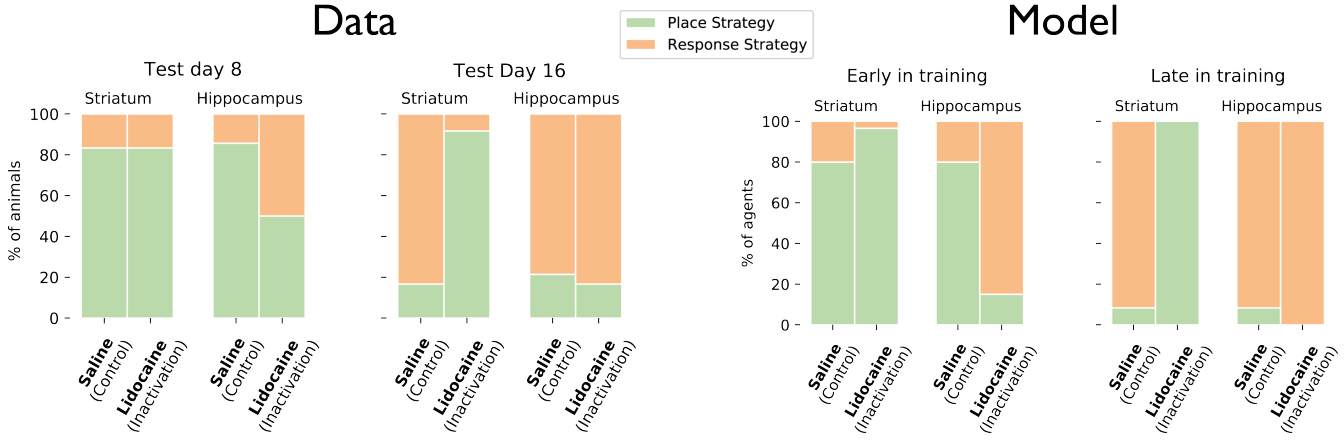

Figure 8. Left panel: distribution of strategy choices (response or place strategy) during the 8th and 16th test days of the experiment conducted by Packard and McGaugh (1996). Control animals (injected with saline) exhibited place learning early in training (test day 8), but switched to response learning later in training (test day 16). Inactivation of the hippocampus using lidocaine caused an earlier switch to response learning. Conversely, inactivating the dorsal striatum prevented the switch to response learning. The graph in the right panel reports the results of the simulated experiment.

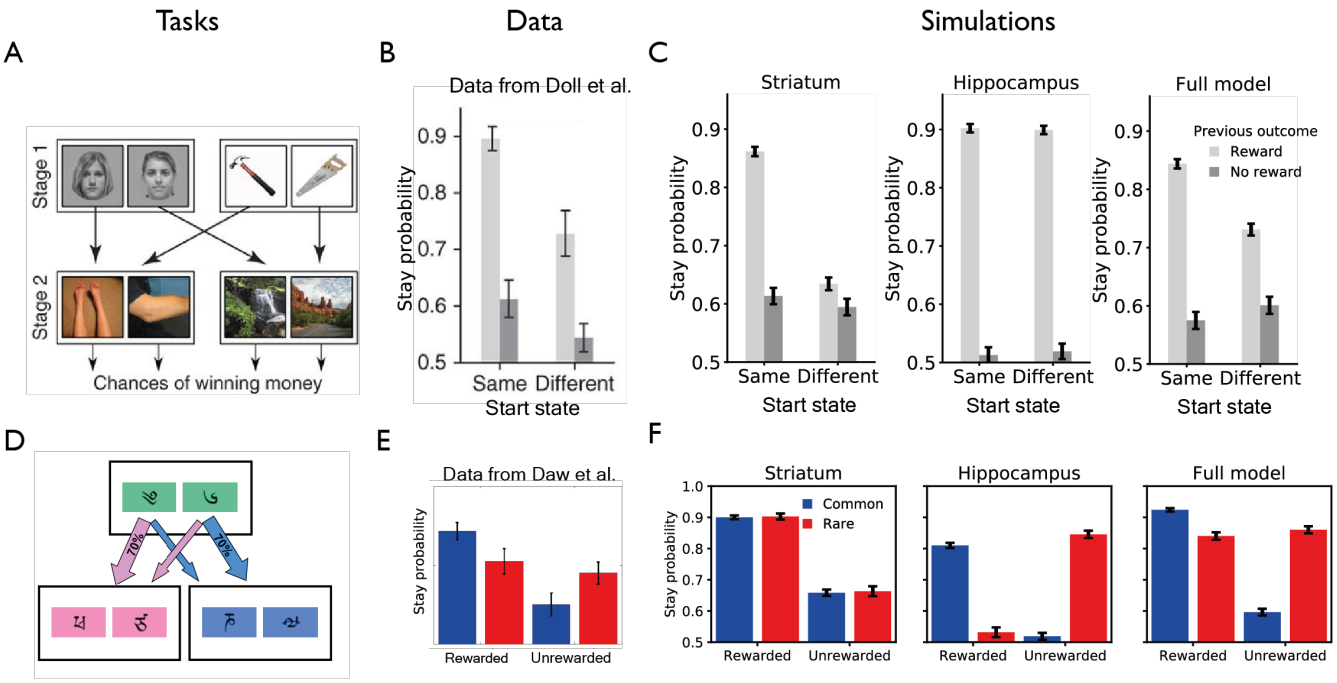

Figure 9. The non-spatial two step tasks. (A) In the task employed by Doll et al. (2015), subjects started from one of two start states (faces or tools). Picking the one of the tools or faces initiated a transition that leaded deterministically to one of the two second-stage states (body parts or scenes), associated with a slowly drifting reward probability. (B) The probability that human participants in Doll et al. (2015) chose the same first-stage action (stay) as on the previous trial binned by whether the previous choice was rewarded, and whether they started in the same state. While true MB agents would perfectly transfer knowledge between start states, and MF agents would only repeat rewarding actions in the same state, human performance resulted in a pattern in between the two. (C) Simulation results. The hippocampal model separates information about reward (in the goal cell firing) and transition structure (in the place cell firing) that can be combined to construct a value function (Figure 3B). This update will immediately affect value in both start states, and will thus affect choices regardless of the subsequent start state, mimicking the true model-based agent presented in the original paper. The striatal model learns separate action values for each start state, so rewards only affect choices on subsequent trials with the same start state. Combining the two models results in a behavioural pattern that lies halfway between model-based and model-free, akin to the human performance observed by Doll et al. (D) A similar task was employed by Daw et al. (2011). Here, a single start state leaded probabilistically to one of either two second state, depending on the action chosen and whether by chance a rare (70%) or common (30%) transition was made. (E) Data from Daw et al. (2011) showing that human performance lies in between MF, which would always repeat rewarding actions, and MB, which would use the task's transition structure to repeat rewarded actions after a common transition, and unrewarded actions after a rare transition. (F) Simulation results. For the striatal model, a first-stage choice resulting in reward is more likely to be repeated on the subsequent trial, regardless of transition probability. The separate reward and transition structure representations based on graph distance make that the hippocampal model's rare transitions affect the value of the other first-stage option. Performance of the model combining both systems results in a pattern half-way between the two, akin to the pattern that human behaviour follows (right-most plot).

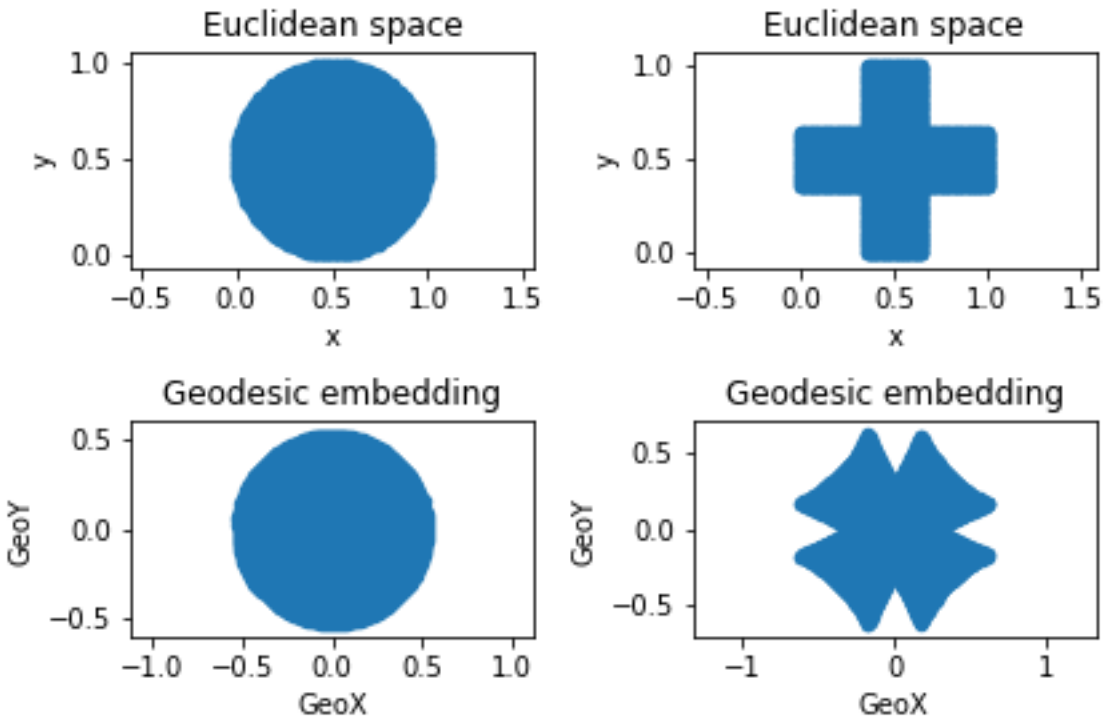

Figure 10. In open field environments, Euclidean and Geodesic distance are equivalent, which is demonstrated in the left panels. When there are obstacles or walls in the environment, Geodesic distance is not equal to Euclidean.