A general model of hippocampal and dorsal striatal learning and decision making

Jesse P. Geerts[a,b,1], Fabian Chersi[b,c,1], Kimberly L. Stachenfeld[d], and Neil Burgess[b,a,2]

[a]Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London, London W1T 4JG, United Kingdom; [b]Institute of Cognitive Neuroscience, University College London, London WC1N 3AZ, United Kingdom; [c]GrAI Matter Labs, 75012 Paris, France; and [d]DeepMind, London N1C 4AG, United Kingdom

Edited by György Buzsáki, New York University Langone Medical Center, New York, NY, and approved October 20, 2020 (received for review April 24, 2020)

Humans and other animals use multiple strategies for making decisions. Reinforcement-learning theory distinguishes between stimulus–response (model-free; MF) learning and deliberative (model-based; MB) planning. The spatial-navigation literature presents a parallel dichotomy between navigation strategies. In "response learning," associated with the dorsolateral striatum (DLS), decisions are anchored to an egocentric reference frame. In "place learning," associated with the hippocampus, decisions are anchored to an allocentric reference frame. Emerging evidence suggests that the contribution of hippocampus to place learning may also underlie its contribution to MB learning by representing relational structure in a cognitive map. Here, we introduce a computational model in which hippocampus subserves place and MB learning by learning a "successor representation" of relational structure between states; DLS implements model-free response learning by learning associations between actions and egocentric representations of landmarks; and action values from either system are weighted by the reliability of its predictions. We show that this model reproduces a range of seemingly disparate behavioral findings in spatial and nonspatial decision tasks and explains the effects of lesions to DLS and hippocampus on these tasks. Furthermore, modeling place cells as driven by boundaries explains the observation that, unlike navigation guided by landmarks, navigation guided by boundaries is robust to "blocking" by prior state–reward associations due to learned associations between place cells. Our model, originally shaped by detailed constraints in the spatial literature, successfully characterizes the hippocampal–striatal system as a general system for decision making via adaptive combination of stimulus–response learning and the use of a cognitive map.

reinforcement learning | spatial navigation | hippocampus | striatum

Behavioral and neuroscientific studies suggest that animals can apply multiple strategies to the problem of maximizing future reward, referred to as the reinforcement-learning (RL) problem (1, 2). One strategy is to build a model of the environment that can be used to simulate the future to plan optimal actions (3) and the past for episodic memory (4–6). An alternative, model-free (MF) approach uses trial and error to estimate a direct mapping from the animal's state to its expected future reward, which the agent caches and looks up at decision time (7, 8), potentially supporting procedural memory (9). This computation is thought to be carried out in the brain through prediction errors signaled by phasic dopamine responses (10). These strategies are associated with different tradeoffs (2). The model-based (MB) approach is powerful and flexible, but computationally expensive and, therefore, slow at decision time. MF methods, in contrast, enable rapid action selection, but these methods learn slowly and adapt poorly to changing environments. In addition to MF and MB methods, there are intermediate solutions that rely on learning useful representations that reduce burdens on the downstream RL process (11–13).

In the spatial-memory literature, a distinction has been observed between "response learning" and "place learning" (14–16). When navigating to a previously visited location, response learning involves learning a sequence of actions, each of which depends on the preceding action or 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" is thought to be supported by the hippocampal formation, where there are neurons tuned to place and heading direction (17–19). Spatial navigation using this map is 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 MB and MF RL (20). Under this view, associative reinforcement is supported by the DLS (21, 22). Indeed, there is evidence from both rodents (23–25) and humans (26, 27) that spatial-response learning relies on the same basal ganglia structures that support MF RL. Evidence also suggests an analogy between MB reasoning and hippocampus (HPC)-based place learning (28, 29). However, this equivalence is not completely straightforward. For example, in rodents, multiple hippocampal lesion and inactivation studies failed to elicit an effect on action-outcome learning, a hallmark of MB planning (30–35). Nevertheless, there are indications that HPC might contribute to a different aspect of MB RL: namely, the representation of relational structure. Tasks that require

Significance

A central question in neuroscience concerns how humans and animals trade off multiple decision-making strategies. Another question pertains to the use of egocentric and allocentric strategies during navigation. We introduce reinforcement-learning models based on learning to predict future reward directly from states and actions or via learning to predict future "successor" states, choosing actions from either system based on the reliability of its predictions. We show that this model explains behavior on both spatial and nonspatial decision tasks, and we map the two model components onto the function of the dorsal hippocampus and the dorsolateral striatum, thereby unifying findings from the spatial-navigation and decision-making fields.

Author contributions: J.P.G., F.C., K.L.S., and N.B. designed research; J.P.G. performed research; J.P.G. analyzed data; and J.P.G., F.C., K.L.S., and N.B. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]J.P.G. and F.C. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: n.burgess@ucl.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2007981117/-/DCSupplemental.

memory of the relationships between stimuli do show dependence on HPC (36–42).

Here, we formalize the perspective that hippocampal contributions to MB learning and place learning are the same, as are the dorsolateral striatal contributions to MF and response learning. In our model, HPC supports flexible behavior by representing the relational structure among different allocentric states, while dorsolateral striatum (DLS) supports associative reinforcement over egocentric sensory features. The model arbitrates between the use of these systems by weighting each system's action values by the reliability of the system, as measured by a recent average of prediction errors, following Wan Lee et al. (43). We show that HPC and DLS maintain these roles across multiple task domains, including a range of spatial and nonspatial tasks. Our model can quantitatively explain a range of seemingly disparate findings, including the choice between place and response strategies in spatial navigation (23, 44) and choices on nonspatial multistep decision tasks (45, 46). Furthermore, it explains the puzzling finding that landmark-guided navigation is sensitive to the blocking effect, whereas boundary-guided navigation is not (27), and that these are supported by the DLS and HPC, respectively (26). Thus, different RL strategies that manage competing tradeoffs can explain a longstanding body of spatial navigation and decision-making literature under a unified model.

Results

We implemented a model of hippocampal and dorsolateral striatal contributions to learning, shown in Fig. 1. Each system independently proposes an action and estimates its value. The value $Q(s,a)$ of taking action a while being in state s is the expected discounted cumulative return:

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

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 gives smaller weight to distal rewards, and $\pi(a|s)$ is the policy specifying a distribution over available actions given the current state. The objective of the RL agent is to discover an optimal policy π^{*} that will maximize value over all states.

Similarly to earlier work in spatial RL (15, 47–49), the two systems in our model estimate value using qualitatively different strategies, which can cause them to generate divergent predictions for the optimal policy. The dorsal striatal component uses an MF temporal difference (TD) method (50) to learn stimulus–response associations directly from egocentric sensory inputs given by landmark cells (LCs) tuned to landmarks at given distances and egocentric directions from the agent (Fig. 1*A* and *Materials and Methods*).

The hippocampal component, in contrast, has access to state information provided by place cells that, in spatial tasks, fire when the agent occupies specific locations. We draw on previous work by Stachenfeld et al. (51) and model hippocampal place cells as encoding the successor representation (SR; ref. 11). The SR is a predictive representation, containing the discounted future occupancy of each state s' from current state s:

$$M^{\pi}(s,s')=\mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}\gamma^{t}\mathbb{I}(s_t=s')|s_0=s\right], \qquad [2]$$

where $\mathbb{I}(s_t=s')=1$ if $s_t=s'$ and 0 otherwise. Each entry $M^{\pi}(s,s')$ of the SR estimates the exponentially discounted count of the number of times state s' is visited in the future, given that the current state is s, conditioned on the current policy $\pi(a|s)$. In addition to the SR, the hippocampal system learns a vector of rewards R associated to each state, which is multiplied with the SR to compute state values (Eq. **8**). Crucially, the hippocampal SR algorithm learns aggregate statistics over the relational structure between states, which allows for some of the flexibility of fully MB systems at lower computational cost. Specifically, SR-based systems decouple learning about transition dynamics from learning about reward, which

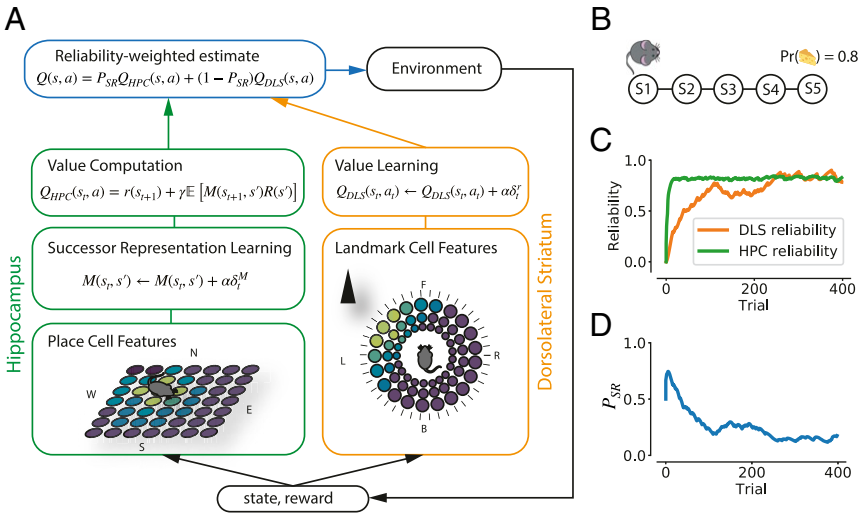

Fig. 1. (*A*) Model architecture. DLS (orange) learns value directly from landmark features in egocentric directions with respect to the agent: L (left), R (right), F (front), or B (back). HPC (green) learns an SR *M* over allocentric input features (north, N; east, E; south, S; or west, W), which is subsequently used for value computation. An arbitrator (blue) computes an average of these values, weighted by each system's reliability (*Materials and Methods*). Lighter colors mean higher firing rates. α, learning rate; δ^{M}, SPE; δ^{r}, reward-prediction error; P_{HPC}, proportion of influence of HPC component. (*B*) A linear track environment with five states. Terminal state S5 gives a reward with probability 0.8. (*C*) Reliability of the hippocampal SR system and the striatal MF system over time as the agent navigates the linear track. Reliability is computed based on the recent average of SPEs δ^{M} for the hippocampal system- and reward-prediction errors δ^{R} for the striatal system. (*D*) The proportion of influence of the SR system on the value function, P_{SR}, in the linear track environment across trials.

allows for a quick recomputation of value under a new reward distribution.

Arbitration between the two systems was achieved by tracking their reliability in predicting states (HPC) and rewards (DLS) and weighting either systems' action values by this reliability, following Wan Lee et al. (43). We operationalized this as the average recent reward-prediction error for the MF system and as the average successor state-prediction error for the SR system. These reliability measures were then used to compute the proportion of influence the SR system had on the value function, P_{SR} (see Eq. **18** for details). Although not modeled in detail here, we suggest that this arbitration is supported by the medial prefrontal cortex, following previous theoretical and experimental work (2, 52). Fig. 1 *B–D* shows an example of how the arbitrator functions. The agent was trained to find a reward (given with probability 0.8) at the end of a simple linear track, in which each state was uniquely identified by landmarks (Fig. 1*B*). The agent was allowed to explore the environment randomly, so it started with a random-walk SR. Hence, the reliability of the HPC starts out higher than that of the DLS. As the average DLS reward-prediction error goes down, and its reliability catches up with that of HPC, the proportion of HPC influence decreases.

To test the validity of our model, we applied it to spatial and nonspatial decision-making tasks and compared its behavior to that of humans and rodents.

Hippocampal Lesions and Adapted Water-Maze Navigation. An adaptation to the classic Morris water-maze task—in which rodents swim in opaque water to find an invisible platform—involved putting an intramaze landmark into the pool at a fixed offset from the platform and moving both platform and landmark to a different location within the tank at the start of each block of four trials (ref. 44 and Fig. 2*A*). In this version of the task, hippocampally lesioned animals performed *better* than intact animals on the first trial of each session, because intact animals initially lingered at the previous goal location (Fig. 2*B*). However, these animals showed little intrasession learning, while learning across sessions was relatively unimpaired, indicating that they were learning to navigate to the goal location relative to the landmark, since this relationship remained constant across sessions.

In the model, the session-by-session displacement of landmark and platform means that the value function will have to change when using allocentric place-cell features, but not when using egocentric LC features. Hence, when we simulated this task by comparing the performance of the full model to a model with a silenced hippocampal component, our model showed the same effects as in the original experiments (Fig. 2*C*). Fast within-session learning, which relies on the SR's capacity for quick reevaluation of rewards, was impaired after a hippocampal lesion. Between-session learning, which depends on learning the landmark–platform relations, was unimpaired. Finally, control agents performed worse than hippocampally lesioned agents on the first trial after the platform had been moved, because the value function changed in allocentric, but not egocentric, coordinate frames. An inspection of the occupancy maps (Fig. 2 *D–F*) reveals that equivalent errors were made by the agents and by the rats—i.e., lingering at the previous platform location. The hippocampal predictive map guides the agent to the previous platform location because of its allocentric place representation. Only when it reaches that location and the platform is not there does it start unlearning the hippocampal reward representation; Eq. **11**.

Simulating DLS lesions in the task used by Pearce et al. (44) showed the emergence of the opposite pattern to that of HPC lesions: There was little to no learning across sessions for the first trials, while fourth-trial performance was not significantly worse than control performance (*SI Appendix*, Fig. S2*A*). This is consistent with previous findings showing that lesions of the

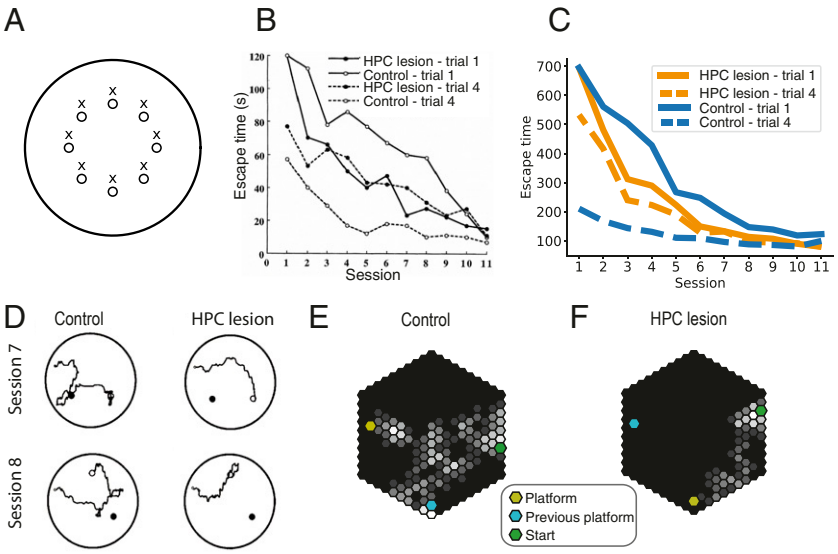

Fig. 2. Results and simulations of the experiment are described in ref. 44. Sessions lasted four trials, and platform and landmark were moved at the beginning of each session. (*A*) Possible locations of the hidden platform (o) and the corresponding landmark (x) in each session. (*B*) Escape latency in the water maze for hippocampal lesioned and control animals on trials 1 (solid lines) and 4 (dashed line) of each session. Hippocampal damage impairs intrasession learning, but preserves learning across sessions. Because animals with hippocampal damage follow a response strategy based on egocentric visual input, they perform better on the first trial of each session than control animals. Reprinted from ref. 15. Copyright (2015), with permission from Elsevier. (*C*) Equivalent plot for the full model (blue) and the model without a hippocampal component, relying solely on MF mechanisms. (*D*) Example trajectories from the first trials of sessions 7 and 8. Animals using a hippocampal place strategy tend to wander around the previous platform location (filled circles) before finding the new platform location (open circles) (adapted from ref. 44). (*E* and *F*) Occupancy maps show a similar effect for simulated agents. Control agents (*E*) linger around the previous platform location, whereas agents that cannot use map-based navigation take a more direct path to the new platform location.

DLS induced a preference for place-guided navigation (53) and that dopamine depletion in the DLS impairs egocentric, but not allocentric, water-maze navigation (54). Our model also accurately captures results from Miyoshi et al. (55), who classified navigation behaviors as cue-guided or place-guided in the cued water-maze task after lesions to both the HPC and the DLS (*SI Appendix*, Fig. S2 *B* and *C*).

These results show that our model captures both landmark-guided and place-memory-guided behavior on the water maze. Furthermore, our model gives a normative perspective on why the animals switch to a landmark-based strategy: Since the striatal system learns about the rewarded location with respect to landmarks, it can use the landmark to navigate directly to the correct location on the first trial of a given session. This gives an advantage to using the striatal system for decision making, which agents learn to exploit. Over the course of multiple sessions, the average prediction error of the striatal system will decrease, causing the reliability-based arbitration mechanism to favor the striatal system, driving lower escape times on first trials of later sessions.

Animals Switch to a Response Strategy on the Plus Maze. The distinct roles of the HPC and dorsal striatum have also been investigated by using the place/response learning task (23, 24). In this task, rats were trained to find a food reward on one arm of a plus maze, starting in the same arm every time, while the opposite arm was blocked (Fig. 3). After training, a probe trial was performed, in which the animal started 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 egocentric turn). If 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).

Fig. 3 shows the results of the original experiment and our simulations. Early in training, most control rats (injected with saline) used a place strategy, but switched to a response strategy after extensive training. Inactivation of the dorsal striatum with lidocaine prevented this switch. Inactivation of the HPC, by contrast, caused the response strategy to be used more often, even early in training. These results indicate that the dorsal striatum supports response learning, while the HPC supports place learning. We simulated the lidocaine inactivation of HPC and dorsal striatum by partly deactivating the SR and MF components of our model, respectively. Early in training, the control agent showed a preference for actions proposed by the HPC, leading the agent to follow a place strategy. This is because the SR reliability was higher than the MF reliability at the start of training, reflecting the fact that animals have explored the environment without rewards before training. Over the course of training, reward-prediction errors in the striatum decreased, causing the reliability of the MF system to increase, at which point the model switched to the MF strategy because of a bias to use the more computationally efficient system. Inactivation of the dorsal striatal and hippocampal components of the model biases the agent to follow a place or response strategy, respectively.

While the results described above show that the DLS and HPC are involved in egocentric and allocentric navigation, respectively, the navigational strategy alone does not speak to an important aspect of MB learning: flexibility in the face of reward devaluation. In devaluation studies, the value of a reinforcer is decreased by pairing it with an aversive event such as illness or by inducing satiety by prefeeding the animal with

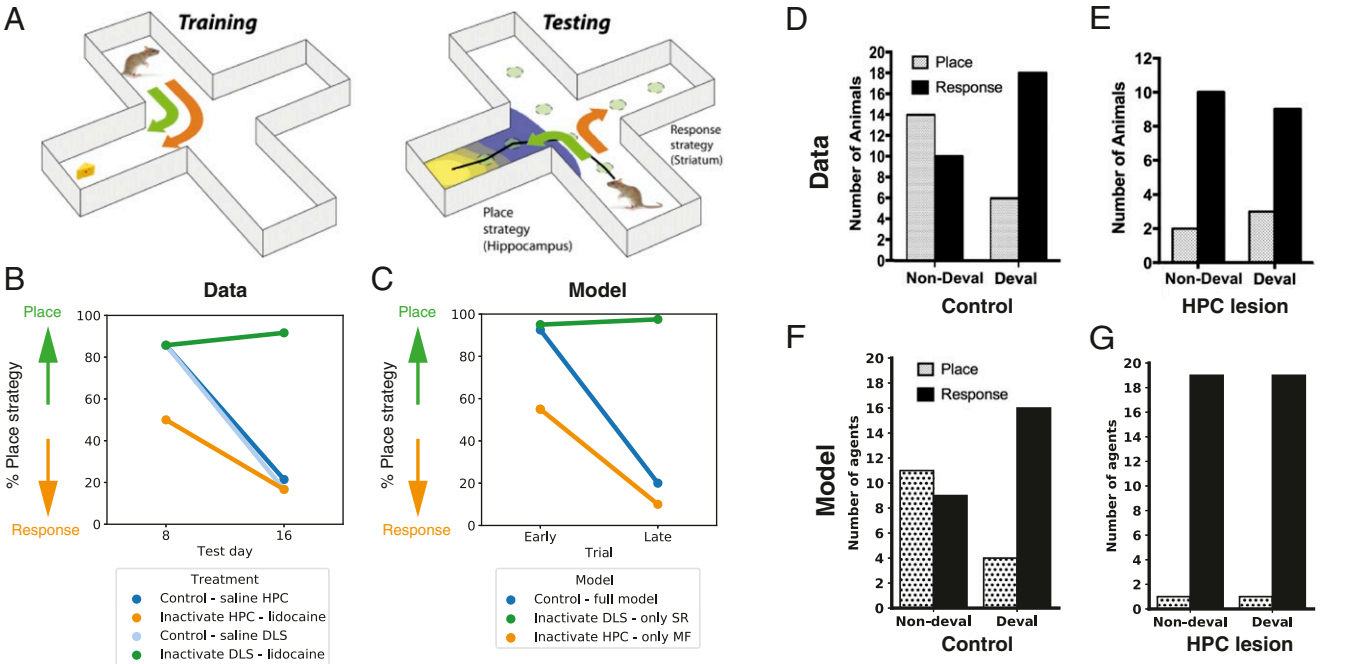

Fig. 3. Navigation in the plus maze. (*A*) Experimental setup used by ref. 23. During training, animals were trained to run from the same starting place to a baited goal arm. During probe trials (on day 8 and day 16), the animal started in the opposite arm. If the animal ran to the same allocentric location as during training, this was labeled as a place strategy (green). Taking the same egocentric turn to end up in the opposite goal arm was classified as a response-learning strategy (orange). (*B*) Behavioral data from ref. 23. Control animals (blue) showed a shift to response learning over the course of training. This was prevented by the inactivation of DLS using lidocaine. The inactivation of HPC using lidocaine caused animals to use a response strategy early on. (*C*) Model results recapitulate these findings. (*D* and *E*) Behavioral data from ref. 56 showing probe-trial behavior before and after the outcome was devalued (deval) by prefeeding the animal with the food reward, for control (*D*) and hippocampally lesioned animals (*E*). *D* and *E* are reprinted from ref. 56, which is licensed under CC BY 4.0. (*F* and *G*) Model-simulation results recapitulate these findings.

the reinforcer (57). Since MF algorithms need to reexperience the state/action leading to the devalued reward to update its value, MF behavior (also referred to as stimulus–response learning) is insensitive to devaluation. MB algorithms, in contrast, can estimate that state/action transitions will lead to a devalued reward without having to reexperience them. This goal-directed, devaluation-sensitive behavior is a hallmark of MB planning (2, 58).

To investigate the relationship between place and response learning on one hand, and goal-directed and stimulus–response learning on the other, we simulated results from Kosaki et al. (56), who studied devaluation on the plus maze. Specifically, they trained rats on the same task as described in Fig. 3*A* (see ref. 59 for a similar study in mice). Subsequently, they devalued the food reinforcer by prefeeding the animals. The results of this devaluation procedure are depicted in Fig. 3*D*. Consistent with the idea that the place strategy is sensitive to the expected value of the outcome, while the response strategy is not, the procedure resulted in a switch from place to response strategies. Furthermore, rats with hippocampal lesions displayed a reliance on the response strategy, regardless of outcome devaluation (Fig. 3*E*), further indicating that the response strategy is insensitive to devaluation. Since sensitivity to reward devaluation is also a property of SR-based learning (60), our model naturally accommodates these results.

Blocking in Landmark But Not Boundary-Related Navigation. A signature of learning stimulus–reward associations using reward-prediction errors is the blocking phenomenon (61). Learning one stimulus–reward association hinders learning of a subsequent association between a different stimulus and the same reward because the prediction error becomes small, reducing further weight updates. In humans, spatial blocking has been shown to occur when learning locations relative to discrete landmarks, but not relative to boundaries (27). Furthermore, learning with respect to landmarks corresponds to increased blood-oxygen-level-dependent (BOLD) signal in the dorsal striatum, whereas learning with respect to boundaries corresponds to activity in the posterior HPC (26).

We aimed to capture these effects by examining the behavior of our agent, following a paradigm similar to ref. 27 (Fig. 4): The agent navigated through an open field to find an unmarked reward location. In order to investigate blocking with respect to boundaries, we explicitly modeled the effect of boundaries on hippocampal place cells, given their dominant role in determining place-cell firing fields (cf. 62 and 63). Rather than learning an SR over a punctate-state representation, the agent learned a matrix of successor features provided by the firing rates of a set of place cells driven by boundary vector cells (BVCs) (64–67).

In the landmark blocking condition (Fig. 4 *A* and *B*), the agent used a landmark to guide navigation. After 10 trials, a second landmark was added, and after 20 trials, the first landmark was removed. Importantly, in this experiment, there were no boundaries, and only one or two landmarks were visible at any time. A single landmark has little effect on place cell firing (63), and, indeed, the presence of a single or two landmarks does not support a reliable place-cell map (64). Therefore, and consistent with BOLD activation results (26), we assume that behavior was controlled by the DLS in this experiment.

As predicted by the TD learning rule, and consistent with the findings of Doeller and Burgess (27), learning about the second

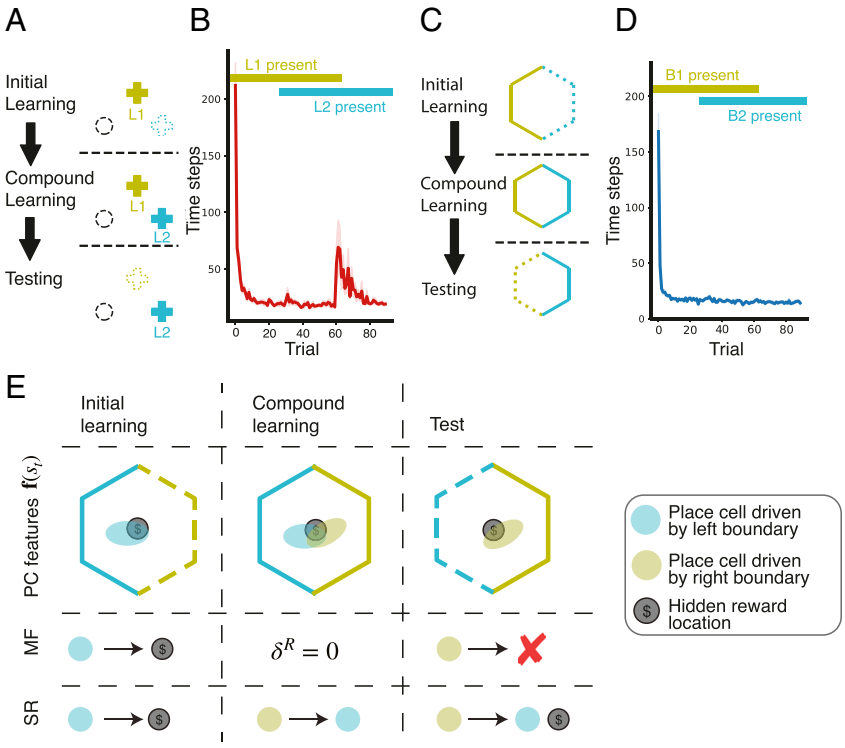

Fig. 4. Boundary versus landmark-blocking experiments, similar to ref. 27. (*A*) Landmark blocking experiment. Agents navigate a virtual water maze to find a hidden platform (dashed circle). During initial learning, one landmark is present (L1). During compound learning, a second landmark is added (L2), after which L1 is removed. (*B*) Average time to find the platform per trial. Increased escape times on removal of L1 indicates blocking of learning about platform location relative to L2 by the prior learning relative to L1. (*C*) Boundary-blocking experiment, following *A*, but with two boundaries (solid green and blue lines). (*D*) Average escape time shows no effect of blocking of learning platform location relative to the right boundary (blue) when the left boundary (green) is removed. (*E*) Illustration of the lack of blocking in boundary-related learning under the SR system, in contrast to an MF system.

landmark was blocked by the prior learning about landmark 1, as evidenced by the drop in performance after its removal.

In the boundary-locking condition (Fig. 4 *C* and *D*), there were no landmarks, meaning that the agent had to rely on its hippocampal system for navigation. The hippocampal system learns a predictive map over boundary-related place-cell activations using successor-prediction errors (SPEs; *SI Appendix*). Prediction-error-based learning like that is susceptible to the blocking effect, and the SR has indeed been used as an explanation for the occurrence of blocking, when learning stimulus–stimulus associations (60). However, when we subjected the agent to a boundary-related blocking paradigm, no blocking occurred (Fig. 4 *C* and *D*).

To understand why this happens, consider the situation in Fig. 4*E*, in which one example place cell was active at the rewarded location, driven by the left boundary. During initial learning, an association between that place cell and the reward was learned. During compound learning, a second boundary drove the activity of another place cell at the rewarded location. In an MF system, the learned value associated to the previous place cell means there was zero prediction error, preventing learning of an association between the second place cell and the reward. In an SR system, however, the agent learns a predictive relationship between the two place cells. Thus, while there is no reward-prediction error, and the reward vector remains unchanged, the newly firing place cell comes to predict the firing of the first place cell (that is associated with reward), mitigating its reduction in firing when the first boundary is removed. This means that, when the first boundary and its associated firing are removed, the agent still predicts reward at the correct location. Thus, consistent with behavioral evidence (26, 27), our model shows no blocking effect during the boundary-related navigation paradigm. This result speaks to the utility of structure learning: The hippocampal SR system learns a multitude of relations, such that its policies are more robust to change in cues and rewards.

Two-Step Task. Outside of the spatial domain, the distinction between MF and MB RL has been heavily investigated by using sequential decision tasks. Here, we describe how our model solves a cognitive decision task of this type—the task of Daw et al. (46) (Fig. 5*A*).

In the two-step decision task designed by Daw et al. (46), human participants were shown a pair of symbols and asked to choose one (Fig. 5*A*). Left or right choices lead to different corresponding second-stage states with high probability (common transitions), but there was a small probability (rare transitions) that the agent transitions to the opposite state. For example, in Fig. 5*A*, 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). During the second stage, participants made another left-or-right choice, resulting in either receiving a reward or not, before starting the next trial. Each of the four outcomes was associated with a reward probability that varied over time as a Gaussian random walk limited between 0.25 and 0.75.

The rewards received or not received on a given trial modify the participants' value estimates for the different actions taken during the two stages, but different RL strategies lead to different behaviors on the next trial. MF learners increased the likelihood of repeating their first-stage action following a reward, regardless of whether a common or rare transition was made. In contrast, MB learners used knowledge of the task's transition structure, such that rewards obtained after a rare transition lead to the opposite choice on the next trial (to maximize the likelihood of reaching the same second state). The key finding of Daw et al. (46) was that human choices reflect both MB and MF influences (Fig. 5*B*).

Our model recapitulates these findings and suggests the HPC could support MB choice in this task, as well as another two-step decision task with deterministic transitions (*SI Appendix*, Fig. S3 and ref. 45). The model DLS, implementing an MF RL system, increased stay probability after rewards, regardless of whether a rare or common transition was made (Fig. 5*C*). In contrast, the HPC uses the SR to generalize value over the graph. When a goal state is reached and a reward is obtained, value is generalized over the graph, according to the degree to which states predict each other. Therefore, on the next trial, the actions were taken that will most likely lead to the recent goal state. Separating transition dynamics from reward estimates thus recapitulates true MB behavior. Combining the two systems results in behavior that is similar to that of human participants in this task.

It has been shown that other, simpler models than pure MB systems can look like MB agents on the two-step task (68). Here, we show that the SR can mimic MB behavior. Because the transition structure is unchanging, caching future state predictions is sufficient for flexible behavior.

Relationship Between Spatial and Two-Step Tasks. A central principle of our model is that MB reasoning and allocentric navigation strategies both rely on the same hippocampal structures. The most direct evidence for this comes from Vikbladh et al. (29), in which both healthy participants and patients with hippocampal damage performed the two-step planning task (46), as well as a landmark versus boundary spatial memory task (26). This allowed the authors to show that, in healthy participants, the degree of MB planning on the sequential decision task correlated with the contribution of allocentric, boundary-driven place memory on the spatial task (reflected in smaller errors from the location predicted by the boundary; Fig. 6*A*). Notably, this correlation cannot be accounted for by variation

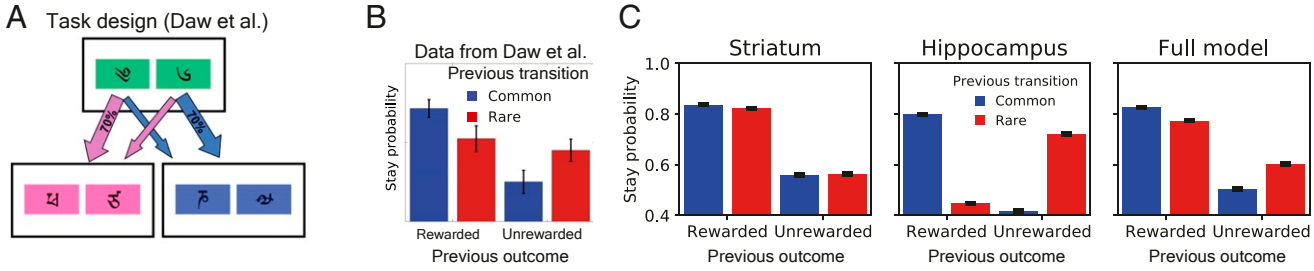

Fig. 5. A nonspatial two-step task. (*A*) Task employed by Daw et al. (46). Here, a single start state led probabilistically to one of either two second states, depending on the action chosen and whether by chance a rare (70%) or common (30%) transition was made. (*B*) Data from Daw et al. (46) showing that human performance lies in between MF and MB. *A* and *B* are reprinted from ref. 46, which is licensed under CC BY 3.0. (*C*) Simulation results for the striatal (*Left*), hippocampal (*Center*), and full (*Right*) models.

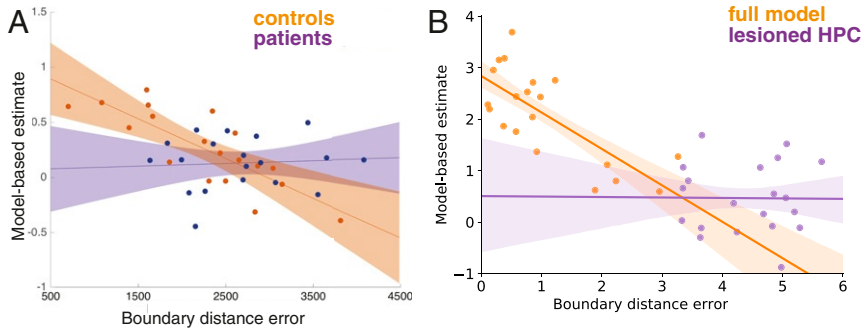

Fig. 6. Relationship between MB planning and allocentric spatial memory. Error bars indicate 80% CIs of the regression in both panels. (*A*) Data from healthy control participants and anterior temporal lobectomy patients, from ref. 29. Allocentric place memory is reflected by responses close to the boundary-predicted location after the landmark has moved (i.e., smaller boundary-distance errors). Dots indicate MB estimates for individual participants, calculated from a mixed-effects logistic regression. Reprinted from ref. 29. Copyright (2019), with permission from Elsevier. (*B*) Simulation data for the full model and agents for which the HPC component was turned off. Here, allocentric place memory is reflected by the average distance between the previous platform location and the location of the maximum of the agent's value function at the start of the next session. Dots represent estimates for individual agents, estimated by a mixed-effects logistic regression.

in general intelligence (intelligence quotient). In patients with hippocampal damage, however, this relationship was significantly reduced.

To test for this effect in our model, we sampled a set of 20 agents with different values for the parameters governing the hippocampal–striatal tradeoff, as well as 20 agents with a partially lesioned hippocampal component (*SI Appendix*). Each agent performed the two-step decision task (46) and the water-maze task of Pearce et al. (44), depicted in Fig. 2. MB planning was quantified as the interaction between effects of reward and transition type in the previous trial on staying with the same action or switching in the next trial (*SI Appendix* and cf. refs. 29 and 46). We quantified the degree of allocentric place memory as the average distance between the previous platform location and the location of the maximum of the agent's value function at the start of the next session. This is akin to the boundary distance error employed by ref. 29. We found a significant correlation ($z = 1.89, p < 0.001$) between model based and allocentric planning (Fig. 6*B*). Agents with hippocampal lesions did not show a significant correlation ($z = -0.02, p = 0.97$), and the difference between these correlation coefficients was significant ($z = 5.44, p < 0.001$), recapitulating the result found by Vikbladh et al. (29).

Discussion

We presented a model of hippocampal and dorsolateral striatal contributions to learning across both spatial navigation and nonspatial decision making. Our simulations support the view that the HPC serves both allocentric place learning and flexible decision making by supplying a predictive map of the underlying structure of the task or environment, whereas the DLS underlies MF learning based on (egocentric) sensory features and actions and that these systems combine weighted by their relative reliability in predicting outcomes.

The involvement of the HPC in abstract nonspatial tasks raises questions about its role throughout evolution. Did the system evolve initially in the spatial domain, but become recruited more generally (14), or was spatial decision making always part of a more general ability (69)? The role of the HPC in MB decision making is much debated. On one hand, lesions of the HPC have not affected hallmarks of MB planning, such as outcome devaluation in lever-pressing tasks (32, 33), although a recent study showed that HPC is involved in devaluation sensitivity of lever pressing immediately after acquisition (when pressing is context-dependent; ref. 70). On the other hand, hippocampal lesions led to a loss of devaluation-sensitivity on the plus maze (Fig. 3 and ref. 56) and impair MB behavior on the two-step task (Fig. 5 and refs. 28 and 29). One crucial difference between the lever-pressing tasks and the tasks simulated here is that the lever-pressing tasks required only one action–outcome association, whereas solving the two-step task and many spatial tasks require chaining multiple action–outcome associations together. Perhaps then, as suggested by Miller et al. (28), the HPC is specifically required when planning requires linking actions to outcomes over multiple steps. By storing temporal abstractions of future states separately from a representation of reward, the SR is particularly well suited for this task of rapidly propagating novel reward information to distant states. That property of the SR has previously inspired models of temporal context memory (71) and might also relate to the role of relational memory tasks more broadly, as they require chaining multiple stimulus–stimulus associations together (37, 39). In line with this role, our simulations showed the hippocampal SR as driving a correlation between spatial-memory performance and MB behavior (Fig. 6 and ref. 29).

Consistent with our model, dorsal striatal neurons showed a great degree of spatial coding in spatial tasks (72), but not in tasks where reward locations were explicitly dissociated from space (73) or where multiple locations were equivalently associated with rewards (74). Indeed, dorsal striatum selectively represents those task aspects, which computational accounts suggest are important for gradual, MF learning (72).

We specifically associate our striatal model with the DLS. Lesion and inactivation studies have shown that the dorsal striatum is functionally very heterogeneous (75). Lesions of the dorsomedial striatum (DMS) result in a switch to response strategies on the plus maze (76) and to cue-based responding in the water maze, while the DLS underlies response learning (77). Furthermore, the DMS has been implicated in learning action–outcome contingencies outside the spatial domain (21, 75). Anatomical connectivity supports this functional dissociation in the dorsal striatum (53, 75). Whereas the DLS receives inputs mostly from sensorimotor cortex and dopaminergic input from the substantia nigra, the DMS receives input from several mesocortical and allocortical areas including the HPC. Indeed, cells encoding route and heading direction have been found in the DMS (78, 79). It is, therefore, likely that the dorsal HPC and the DMS are part of a single circuit involved in flexible goal-directed decision making, whereby the HPC provides map-based information, and the DMS is involved in action selection.

Our work follows several models of spatial decision making by hippocampal and striatal systems (15, 48, 49, 80, 81). Dollé and colleagues (48, 49) used a similar hippocampo-striatal model to explain behavior on the adapted water-maze task (44), presented

in Fig. 2. Our model differs in two important ways. Firstly, in their model, place cells connected to "graph cells" that formed an explicit topological graph of the spatial environment, used to explicitly plan a path to the goal. In the present model, by contrast, the topological structure of the environment is implied in the predictive SR, following a theoretical proposal by Stachenfeld et al. (51) and neuroimaging (40, 41) and behavioral findings (82). Thus, our agent mimicked true MB behavior (explicit graph search) by using an intermediate SR-based strategy. Secondly, their model used another expert network that learned whether to take striatal or hippocampal outputs using TD learning. In contrast, our model arbitrates between systems based on their reliability. This arbitration mechanism predicts that on trials with high reward-prediction error, control should shift away from the MF system. In contrast, a low predictability of state transitions leads to higher average errors in the SR system and should, therefore, lead to a higher degree of MF control. Evidence for this comes from Wan Lee et al. (43), who, furthermore, showed that the prefrontal cortex encodes neural correlates of arbitration based on reliability.

As noted above, the hippocampal results we simulated are also consistent with a fully MB system, which is strictly more flexible. An interesting question is how to disambiguate between animals using an MB strategy versus the SR. One weakness of the temporal-difference SR model used here is that it cannot respond flexibly when the transition structure changes. Momennejad et al. (83) have shown that humans are better at revaluating when the reward function changes than when the transition structure changes, consistent with use of an SR. In addition, hippocampal replay has been suggested to perform offline updates of the hippocampal predictive map to incorporate these kinds of transition changes (84, 85). As an alternative, tracking input covariances and using these for updating the SR allow it to solve certain kinds of transition-revaluation problems without requiring forward simulation (86). A second weakness of the SR, compared to MB systems, is that the SR is policy-dependent. This means that the SR corresponding to an optimal policy for one reward setting is of limited use for problems with a different reward function (87). Piray and Daw (88) have recently proposed that the hippocampal system might resolve this latter weakness using a *default representation*, corresponding to a default policy. Alternatively, the HPC might represent a set of multiple distinct SR maps corresponding to different policies (89). Taken together, these two failure modes of the SR provide interesting avenues for experiments probing animals' behavioral strategies and for theoretical work on computational tradeoffs between these strategies.

In addition to the HPC, the orbitofrontal cortex (OFC) has been hypothesized to be important for representing states in RL problems. Wilson, Niv, and colleagues (90) introduced a model in which OFC plays a critical role in identifying states that are perceptually similar. This corresponds to data showing that OFC is specifically necessary for decision making in partially observable environments (91). Evidence for this theory comes from human functional MRI research showing that unobservable task states can be decoded from OFC and that this relates to task performance (92). This proposed role of the OFC is distinct from, and possibly complementary to, our proposed role for the HPC. In our model, the HPC encodes a predictive map based on observable features that can be used for rapid, flexible decision making. The OFC, on the other hand, is crucial for a general state representation that can be used for downstream MB or MF processes. Whether and how the OFC and the HPC can interact to allow SR learning in partially observable environments is an interesting avenue for further research (see also ref. 93).

Our explanation for the absence of boundary-related blocking (Fig. 4) relies on BVC inputs to hippocampal place cells. BVCs can respond to intramaze landmarks as well as to boundaries (although, in contrast to DLS LCs, BVCs fire irrespective of object identity; ref. 67). This means that a sufficient number of landmarks could drive a reliable place-cell representation of space, allowing hippocampal control and the prevention of blocking. However, in the experiments simulated here, there were only one or two landmarks present. Single landmarks have little influence on firing relative to extended boundaries (63), consistent with the BVC model. Because BVCs fire proportionally to the angle subtended by the stimulus (94), place cells do not provide a reliable representation of space when there is only a single landmark (64). Thus, we predict that the addition of greater numbers of landmarks should allow construction of a reliable place-cell map, thereby leading to increased hippocampal influence and a reduction of blocking effects.

Our model reflects the assumption, driven by our knowledge of the neural representations, that in spatial tasks, the hippocampal SR system uses allocentric representations, while the MF system uses egocentric representations. This allowed us to fit the behavioral data well and raised the question of why the goal-directed system is allocentric, while the stimulus–response system is egocentric? Perhaps an answer lies in the time scale of learning: The allocentric layout of a large environment is stable, irrespective of your changes in location or direction, making it suitable for learning long-term relationships between stimuli. Consistent with this idea, "slow feature analysis" produces grid and place-cell representations from visual inputs because they vary slowly (95). On the other hand, egocentric representations are more suited to mapping sensory inputs to physical actions, both of which are specified egocentrically.

In conclusion, dorsal HPC and DLS support qualitatively different strategies for learning about reward in spatial as well as nonspatial contexts, as captured by the model presented here. The fact that the same model explains behavior in both types of tasks implies that the hippocampal–striatal system is a general-purpose learning device that adaptively combines MB and MF mechanisms.

Materials and Methods

Hippocampal and Striatal Systems for Decision Making. Our model combines a hippocampal RL module based on the SR with a striatal model based on MF value learning (Fig. 1*A*). It arbitrates between these modules based on their relative reliability, which can be computed by using the average of recent prediction errors. Model details are outlined below.

Dorsal Striatal System. The DLS module was implemented as an MF 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 nonspatial tasks. Sensory input was coded by a set of egocentric landmark vector cells coding for the presence or absence of a landmark in a particular egocentric direction, at a particular distance from the landmark to the agent, analogous to the egocentric BVCs recently reported (96). Specifically, the activation of each LC was modeled as a bivariate Gaussian in a space defined by the egocentric angle θ and distance d of the landmark to the agent:

$$f^{LC}(d, \theta) \propto \mathcal{N}([d, \theta]; [d^*, \theta^*], \Sigma), \quad [3]$$

where d^* and θ^* are the preferred distance and orientation of the LC, respectively, and $\Sigma = diag([\sigma_d, \sigma_\theta])$ is the covariance matrix with the tuning width and length of the receptive field on the diagonal entries. We assumed that LCs are sensitive to the identity of the landmark, meaning that a different set of LCs will respond to a different landmark in our model. An example egocentric LC is shown in *SI Appendix*, Fig. S1. In the nonspatial tasks, states were encoded as "one-hot" vectors containing ones for their state indexes, reflecting the fact that states were uniquely identifiable as different images.

LCs in the sensory layer project to neurons in the dorsal striatum in an all-to-all connected way:

$$x_a^{DLS} = Q_{DLS}(s, a) = \sum_{i=1}^{N} w_{i,a} f_i^{LC}(s), \quad [4]$$

where f_i^{DLS} is the activity of LC i, x_a^{DLS} is the firing rate of the dorsolateral striatal neuron corresponding to striatal estimated value Q^{DLS} of action a given state s, N is the total number of sensory neurons, u_i^{LC} is the firing rate of LC i, and $w_{i,a}$ is the weight from sensory neuron i to striatal neuron a.

Learning in the striatal network is mediated by a Q-learning rule (50). This allows the model to compute a TD reward-prediction error δ_t^r:

$$\delta_t^r = r_{t+1} + \gamma \max_{a'} Q_{DLS}(s_{t+1}, a') - Q_{DLS}(s_t, a_t), \quad [5]$$

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

$$\Delta w_{i,a} = \alpha_Q \delta_t^r e_{i,a}, \quad [6]$$

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

$$e_{i,a}(t+1) = f_i^{LC} x_a^{DLS} + \lambda e_{i,a}(t), \quad [7]$$

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 (1).

Hippocampal System. The hippocampal place-cell system was modeled as encoding the SR, following work by Stachenfeld et al. (51). The SR is a predictive representation employed in machine learning (11, 13, 97, 98), containing the discounted future occupancy of each state s' from current state s (Eq. **2**). In the hippocampal SR model, a row of the SR—i.e., $M^\pi(s, :)$—constitutes the current population activity vector—i.e., the activity of every place cell in the current state. A column of M^π contains the activity of a single place cell in all possible locations (states)—i.e., a rate map (*SI Appendix*, Fig. S1). In addition to the SR matrix, the agent will learn a vector with the expected reward $R(s)$ for each states. The agent combines these to compute state value:

$$V_{HPC}^{\pi}(s) = \sum_{s'} M(s, s') R(s'). \quad [8]$$

The factorization of value into the SR and reward confers more flexible behavior because if one term changes, it can be relearned, while the other term remains intact (11). The agent used one-step lookahead to compute the value of each action $Q(s, a)$, combining direct reward and the next state's value:

$$Q_{HPC}(s_t, a_t) = r(s_t) + \gamma \mathbb{E}_{s_{t+1}|s_t,a_t}[V_{HPC}(s_{t+1})]. \quad [9]$$

The SR satisfies a Bellman equation, meaning that any RL method can be used to learn the SR. Here, learning was achieved by using a TD update:

$$\Delta \hat{M}(s_t, s') = \alpha_M \delta_t^M(s'), \quad [10]$$

where $\delta_t^M(s') = \left[\mathbb{I}(s_t = s') + \gamma \hat{M}(s_{t+1}, s') - \hat{M}(s_t, s')\right]$ is a TD SPE pertaining to state s' and α_M is a learning rate. For the spatial-navigation studies modeled in this paper, animals were allowed to freely explore the environment without any reward before starting the task (23, 44). Hence, for these tasks, the SR was initialized as the SR associated to a random-walk policy M^{RW} over a uniform spatial discretization of the environment. This was not the case for the task graphs of the two-step decision tasks (45). Therefore, in these tasks, we initialized the SR as the identity matrix I, encoding no other knowledge than the fact that every state predicts itself. Finally, the reward vector $\hat{R}$ was learned by using a simple delta rule:

$$\Delta \hat{R}(s_t) = \alpha_R \left(r_t - \hat{R}(s_t)\right). \quad [11]$$

Although the SR is often introduced as above (in terms of discrete state counts), accurately estimating the SR for every state is infeasible in very large state spaces. This is known as the *curse of dimensionality*, and it necessitates the use of function approximation (1). The agent observes states through a vector of features $\mathbf{f}(s)$, which, if chosen rightly, will be of much smaller dimension than the number of states, allowing the agent to generalize to states that are nearby in feature space. The feature-based SR [also referred to as Successor Features (13)], rather than encoding the discounted number of state visits, encodes the expected discounted future activity of each feature:

$$\psi^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \mathbf{f}(s_t) | s_0 = s\right]. \quad [12]$$

As in the tabular case, the feature-based SR can be used to compute value when multiplied with a vector of reward expectations per feature, $\mathbf{u}$: $V^\pi(s) = \psi^\pi(s)^T \mathbf{u}$. In the case of linear-function approximation, these Successor Features ψ in Eq. **12** are approximated by a linear function of the features $\mathbf{f}$:

$$\hat{\psi}(s) = W^T \mathbf{f}(s), \quad [13]$$

where W is a weight matrix which parameterizes the approximation. Intuitively, W encodes how much each feature predicts every other feature. As in the tabular case, TD learning can be used to update the SR weights (*SI Appendix*). Thus, at every state s (corresponding to a location) in the environment, the agent observed a population vector $\mathbf{f}(s)$ of BVC-driven place cells. It then computed its estimated Successor Features ψ using its current estimate of weights W and Eq. **13**, which encode the discounted sum of future population firing-rate vectors $\mathbf{f}$ of the input place cells. In terms of circuitry, W might correspond to the Schaffer collaterals projecting from CA3 to CA1 neurons, corresponding to $\mathbf{f}$ and ψ, respectively.

In the context of HPC, the feature-based SR allows us to represent states as population vectors of place cells with overlapping firing fields (the features), rather than having a one-to-one correspondence between place cells and states. Then, we are free to model the dependence of the place cell firing on specific environmental features (boundaries). This dependence has been extensively characterized by computational models of BVCs (64, 65, 99–101), which were shown to exist in the subiculum (66). Accordingly, we modeled a set of hippocampal place cells, whose activity $\mathbf{f}_i(s_t)$ was the thresholded sum of a set of BVC inputs (see ref. 64 for details on how BVC and place-cell maps were calculated).

Crucially, modeling place cells as driven by BVCs allows us to explain the puzzling experimental finding by Doeller and Burgess (27) that learning to navigate to a location relative to a landmark, but not relative to a boundary, is sensitive to the blocking effect (61). In an accompanying neuroimaging paper, the authors showed that landmark learning was associated to BOLD activity in the dorsal striatum, whereas boundary-related navigation was associated to activity in the HPC (26).

Arbitration Process. The agent has access to both its MF DLS component and its hippocampal component employing the SR. Both systems estimate the same value function, but might make different types of errors, and the agent has to arbitrate between them.

Rational arbitration should reflect the relative uncertainty (2), requiring the posterior distribution over values, rather than just the values themselves. Here, we used a convenient proxy for uncertainty, introduced by Wan Lee et al. (43)—namely, the recent average of prediction errors: the reward-prediction error for the MF component and the SPE for the SR component. If the SPE is low, this means that the SR system has a good estimate of the world. Similarly, if reward-prediction errors are low, this means the MF system has a reliable estimate of the value function. The reliability can be tracked by using a Pearce–Hall-like update rule (102), computing the recent average of absolute prediction errors Ω:

$$\Delta\Omega = \eta(|\delta| - \Omega), \quad [14]$$

where $|\delta|$ is the absolute reward-prediction error and η is a learning rate. The reliability is defined as:

$$\chi = (\delta_{MAX} - \Omega)/\delta_{MAX}, \quad [15]$$

with δ_{MAX} being the upper bound of the prediction error, which was set to one. Since in our model both systems are trained by a prediction error, we can apply this to both the MF and SR systems. Following Wan Lee et al. (43), we used the reliability measure for arbitration. These authors computed transition rates α and β for transitioning from MF to MB states, and vice versa, as follows. Here, we used the same terms, but for transitions between MF and SR. These transition rates are functions of the reliability of the respective systems:

$$\alpha(\chi_{MF}) = \frac{A_\alpha}{1 + \exp(B_\alpha \chi_{MF})}, \quad [16]$$

$$\beta(\chi_{SR}) = \frac{A_\beta}{1 + \exp(B_\beta \chi_{SR})}, \quad [17]$$

where the A and B parameters in both equations determine the transition rate and the steepness of these curves, respectively. These parameters were fitted to behavioral data by Wan Lee et al. (43), and we matched their parameter values (*SI Appendix*, Table S1). At each time step, the rate of change of the proportion of influence of the SR system P_{SR} was computed by using the following differential equation, generating a push–pull mechanism between HPC and DLS influence over behavior:

$$\frac{dP_{SR}}{dt} = \alpha(\chi_{MF})(1 - P_{SR}) - \beta(\chi_{SR})P_{SR}. \quad [18]$$

Note that, consistent with behavioral data from human subjects (43), this arbitration mechanism resulted in a weighted influence of both systems in the final value estimates (Fig. 1), rather than a discrete choice. Note that the arbitrator combines the action values, not the actions. Thus, the agent will not end up with a midway action when the two systems encode different preferences. Lesions or partial inactivations of either the DLS or the HPC were achieved by setting limits on P_{SR} (see *SI Appendix* for more details).

Code Availability. The results were generated by using code written in Python. Code is available on ModelDB (accession no. 266836) (103).

ACKNOWLEDGMENTS. We thank Dan Bush, Will de Cothi, Changmin Yu, and Kevin Miller for useful comments on the manuscript; Oliver Vikbladh and Maté Lengyel for discussions; and our anonymous reviewers for insightful suggestions. This work was supported by the European Union's Horizon 2020 research and innovation program under Grant Agreement 785907 Human Brain Project SGA2; European Research Council Advanced Grant NEUROMEM; the Wellcome Trust; and the Gatsby Charitable Foundation.

1. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998), p. 1054.
2. N. D. Daw, Y. Niv, P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
3. E. C. Tolman, Cognitive maps in rats and man. *Psychol. Rev.* **55**, 189–208 (1948).
4. E. Tulving, Episodic and semantic memory. *Organization of memory* **1**, 381–403 (1972).
5. D. L. Schacter, D. R. Addis, R. L. Buckner, Remembering the past to imagine the future: The prospective brain. *Nat. Rev. Neurosci.* **8**, 657–661 (2007).
6. A. Bicanski, N. Burgess, A neural-level model of spatial memory and imagery. *eLife* **7**, e33752 (2018).
7. R. A. Rescorla, A. R. Wagner, "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement" in *Classical Conditioning II: Current Research and Theory*, A. H. Black, W. F. Prokasy, eds. (Appleton-Century-Crofts, New York, NY, 1972), vol. 2, pp. 64–99.
8. R. S. Sutton, Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
9. L. R. Squire, S. Zola-Morgan, The medial temporal lobe memory system. *Science* **253**, 1380–1386 (1991).
10. P. R. Montague, P. Dayan, T. J. Sejnowski, A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
11. P. Dayan, Improving generalisation for temporal difference learning: The successor representation. *Neural Comput.* **5**, 613–624 (1993).
12. L. Lehnert, M. L. Littman, Transfer with model features in reinforcement learning. arXiv:1807.01736 (4 July 2018).
13. A. Barreto, R. Munos, T. Schaul, D. Silver, Successor features for transfer in reinforcement learning. arXiv:1606.05312 (16 June 2016).
14. J. O'Keefe, L. Nadel, *The Hippocampus as a Cognitive Map* (Clarendon Press, Oxford, UK, 1978).
15. F. Chersi, N. Burgess, The cognitive architecture of spatial navigation: Hippocampal and striatal contributions. *Neuron* **88**, 64–77 (2015).
16. NM. White, The role of stimulus ambiguity and movement in spatial navigation: A multiple memory systems analysis of location discrimination. *Neurobiol. Learn. Mem.* **82**, 216–229 (2004).
17. J. O'Keefe, J. Dostrovsky, The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
18. J. S. Taube, R. U. Muller, J. B. Ranck, Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J. Neurosci.* **10**, 420–435 (1990).
19. T. Hafting, M. Fyhn, S. Molden, M. Moser, E. I. Moser, Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).
20. R. Poldrack, M. Packard, Competition among multiple memory systems: Converging evidence from animal and human brain studies. *Neuropsychologia* **41**, 245–251 (2003).
21. H. H. Yin, S. B. Ostlund, B. J. Knowlton, B. W. Balleine, The role of the dorsomedial striatum in instrumental conditioning. *Eur. J. Neurosci.* **22**, 513–523 (2005).
22. H. H. Yin, B. J. Knowlton, B. W. Balleine, Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* **19**, 181–189 (2004).
23. M. G. Packard, J. L. McGaugh, Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiol. Learn. Mem.* **72**, 65–72 (1996).
24. M. G. Packard, Glutamate infused posttraining into the hippocampus or caudate-putamen differentially strengthens place and response learning. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12881–12886 (1999).
25. R. J. Mcdonald, N. M. White, Parallel information processing in the water maze : Evidence for independent memory systems involving dorsal striatum and hippocampus. *Behav. Neural. Biol.* **270**, 260–270 (1994).
26. C. F. Doeller, J. A. King, N. Burgess, Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5915–5920 (2008).
27. C. F. Doeller, N. Burgess, Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5909–5914 (2008).
28. K. J. Miller, M. M. Botvinick, C. D. Brody, Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* **20**, 1269–1276 (2017).
29. O. M. Vikbladh *et al.*, Hippocampal contributions to model-based planning and spatial memory. *Neuron* **102**, 683–693.e4 (2019).
30. D. P. Kimble, R. BreMiller, Latent learning in hippocampal-lesioned rats. *Physiol. Behav.* **26**, 1055–1059 (1981).
31. D. P. Kimble, W. P. Jordan, R. BreMiller, Further evidence for latent learning in hippocampal-lesioned rats. *Physiol. Behav.* **29**, 401–407 (1982).
32. L. H. Corbit, B. W. Balleine, The role of the hippocampus in instrumental conditioning. *J. Neurosci.* **20**, 4233–4239 (2000).
33. L. H. Corbit, S. B. Ostlund, B. W. Balleine, Sensitivity to instrumental contingency degradation is mediated by the entorhinal cortex and its efferents via the dorsal hippocampus. *J. Neurosci.* **22**, 10976–10984 (2002).
34. J. Ward-Robinson *et al.*, Excitotoxic lesions of the hippocampus leave sensory preconditioning intact: Implications for models of hippocampal functioning. *Behav. Neurosci.* **115**, 1357–1362 (2001).
35. S. Gaskin, S. Chai, NM. White, Inactivation of the dorsal hippocampus does not affect learning during exploration of a novel environment. *Hippocampus* **15**, 1085–1093 (2005).
36. W. B. Scoville, B. Milner, Loss of recent memory after bilateral hippocampal lesions. *J Neurol. Neurosurg. Psychiatry.* **20**, 11–21 (1957).
37. J. A. Dusek, H. Eichenbaum, The hippocampus and memory for orderly stimulus relations. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 7109–7114 (1997).
38. L. M. DeVito, H. Eichenbaum, Memory for the order of events in specific sequences: Contributions of the hippocampus and medial prefrontal cortex. *J. Neurosci.* **31**, 3169–3175 (2011).
39. M. Bunsey, H. Eichenbaum, Conservation of hippocampal memory function in rats and humans. *Nature* **379**, 255–257 (1996).
40. A. C. Schapiro, N. B. Turk-Browne, K. A. Norman, M. M. Botvinick, Statistical learning of temporal community structure in the hippocampus. *Hippocampus* **26**, 3–8 (2016).
41. M. M. Garvert, R. J. Dolan, T. E. Behrens, A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. *eLife* **6**, e17086 (2017).
42. F. Vargha-Khadem *et al.*, Differential effects of early hippocampal pathology on episodic and semantic memory. *Science* **277**, 376–380 (1997).
43. S. Wan Lee, S. Shimojo, J. P. O'Doherty, Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **81**, 687–699 (2014).
44. J. M. Pearce, A. D. L. Roberts, M. Good, Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors. *Nature* **62**, 1997–1999 (1998).
45. B. B. Doll, K. D. Duncan, D. A. Simon, D. Shohamy, ND. Daw, Model-based choices involve prospective neural activity. *Nat. Neurosci.* **18**, 767–772 (2015).
46. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
47. F. Chersi, N. Burgess, "Hippocampal and striatal involvement in cognitive tasks : A computational model" in *Proceedings of the 6th International Conference on Memory ICOM16* (2016), pp. 24–28.
48. L. Dollé, D. Sheynikhovich, B. Girard, R. Chavarriaga, A. Guillot, Path planning versus cue responding: A bio-inspired model of switching between navigation strategies. *Biol. Cybern.* **103**, 299–317 (2010).
49. L. Dollé, R. Chavarriaga, A. Guillot, M. Khamassi, Interactions of spatial strategies producing generalization gradient and blocking: A computational approach. *PLoS Comput. Biol.* **14**, e1006092 (2018).
50. C. J. C. H. Watkins, P. Dayan, Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
51. K. L. Stachenfeld, M. M. Botvinick, S. J. Gershman, The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
52. S. Killcross, E. Coutureau, Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb. Cortex* **13**, 400–408 (2003).
53. B. D. Devan, N. M. White, Parallel information processing in the dorsal striatum: Relation to hippocampal function. *J. Neurosci.* **19**, 2789–2798 (1999).
54. A. A. Braun *et al.*, Dopamine depletion in either the dorsomedial or dorsolateral striatum impairs egocentric Cincinnati water maze performance while sparing allocentric Morris water maze learning. *Neurobiol. Learn. Mem.* **118**, 55–63 (2015).
55. E. Miyoshi *et al.*, Both the dorsal hippocampus and the dorsolateral striatum are needed for rat navigation in the Morris water maze. *Behav. Brain Res.* **226**, 171–178 (2012).
56. Y. Kosaki, J. M. Pearce, A. McGregor, The response strategy and the place strategy in a plus-maze have different sensitivities to devaluation of expected outcome. *Hippocampus* **28**, 484–496 (2018).

57. C. D. Adams, A. Dickinson, Instrumental responding following reinforcer devaluation. *Q. J. Exp. Psychol. B* **33**, 109–121 (1981).
58. P. Dayan, K. C. Berridge, Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognit. Affect Behav. Neurosci.* **14**, 473–492 (2014).
59. E. De Leonibus *et al.*, Cognitive and neural determinants of response strategy in the dual-solution plus-maze task. *Learn. Mem.* **18**, 241–244 (2011).
60. M. P. H. Gardner, G. Schoenbaum, S. J. Gershman, Rethinking dopamine as generalized prediction error. *Proc. Biol. Sci.* **285**, 20181645 (2018).
61. L. J. Kamin, "Predictability, surprise, attention, and conditioning" in *Punishment and Aversive Behavior*, B. A. Campbell, R. M. Church, Eds. (Appleton-Century-Crofts, New York, 1969), pp. 279–296.
62. J. O'Keefe, N. Burgess, Geometric determinants of the place fields of hippocampal neurons. *Nature* **381**, 425–428 (1996).
63. A. Cressant, R. U. Muller, B. Poucet, Failure of centrally placed objects to control the firing fields of hippocampal place cells. *J. Neurosci.* **17**, 2531–2542 (1997).
64. C. Barry *et al.*, The boundary vector cell model of place cell firing and spatial memory. *Rev. Neurosci.* **17**, 71–98 (2006).
65. T. Hartley, N. Burgess, C. Lever, F. Cacucci, J. O'Keefe, Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus* **10**, 369–379 (2000).
66. C. Lever, S. Burton, A. Jeewajee, J. O'Keefe, N. Burgess, Boundary vector cells in the subiculum of the hippocampal formation. *J. Neurosci.* **29**, 9771–9777 (2009).
67. A. Bicanski, N. Burgess, Neuronal vector coding in spatial cognition. *Nat. Rev. Neurosci.* **21**, 453–470 (2020).
68. T. Akam, R. Costa, P. Dayan, Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS Comput. Biol.* **11**, e1004648 (2015).
69. H. Eichenbaum, T. Otto, N. J. Cohen, The hippocampus: What does it do?. *Behav. Neural. Biol.* **57**, 2–36 (1992).
70. L. A. Bradfield, B. K. Leung, S. Boldt, S. Liang, B. W. Balleine, Goal-directed actions transiently depend on dorsal hippocampus. *Nat. Neurosci.* **23**, 1194–1197 (2020).
71. S. J. Gershman, C. D. Moore, M. T. Todd, K. A. Norman, P. B. Sederberg, The successor representation and temporal context. *Neural. Comput.* **24**, 1553–1568 (2012).
72. M. A. A. van der Meer, A. Johnson, N. C. Schmitzer-Torbert A. D. Redish, Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron* **67**, 25–32 (2010).
73. N. C. Schmitzer-Torbert, A. D. Redish, Task-dependent encoding of space and events by striatal neurons is dependent on neural subtype. *Neuroscience* **153**, 349–360 (2008).
74. J. D. Berke, J. T. Breck, H. Eichenbaum, Striatal versus hippocampal representations during win-stay maze performance. *J. Neurophysiol.* **101**, 1575–1587 (2009).
75. H. H. Yin, B. J. Knowlton, The role of the basal ganglia in habit formation. *Nat. Rev. Neurosci.* **7**, 464–476 (2006).
76. H. H. Yin, B. J. Knowlton, Contributions of striatal subregions to place and response learning. *Learn. Mem.* **11**, 459–463 (2004).
77. B. D. Devan, R. J. McDonald, N. M. White, Effects of medial and lateral caudate-putamen lesions on place- and cue-guided behaviors in the water maze: Relation to thigmotaxis. *Behav. Brain Res.* **100**, 5–14 (1999).
78. E. Tabuchi, A. B. Mulder, S. I. Wiener, Neurons in hippocampal afferent zones of rat striatum parse routes into multi-pace segments during maze navigation. *Eur. J. Neurosci.* **19**, 1923–1932 (2004).
79. K. Ragozzino, S. Leutgeb, S. Mizumori, Dorsal striatal head direction and hippocampal place representations during spatial navigation. *Exp. Brain Res.* **139**, 372–376 (2001).
80. D. J. Foster, R. G. Morris, P. Dayan, A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* **10**, 1–16 (2000).
81. N. J. Gustafson, N. D. Daw, Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS Comput. Biol.* **7**, e1002235 (2011).
82. J. L. S. Bellmund *et al.*, Deforming the metric of cognitive maps distorts memory. *Nat. Hum. Behav.* **4**, 177–188 (2019).
83. I. Momennejad *et al.*, The successor representation in human reinforcement learning. *Nat. Hum. Behav.* **1**, 680–692 (2017).
84. E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol.* **13**, e1005768 (2017).
85. T. Evans, N. Burgess, Coordinated hippocampal-entorhinal replay as structural inference, *Adv. Neural Information Processing Systems* **32**, 1729–1741 (2019).
86. J. P. Geerts, K. L. Stachenfeld, N. Burgess, "Probabilistic successor representations with Kalman temporal differences" in *Conference on Computational Cognitive Neuroscience* (2019).
87. L. Lehnert, S. Tellex, M. L. Littman, Advantages and limitations of using successor features for transfer in reinforcement learning. arXiv:1708.00102 (31 July 2017).
88. P. Piray, ND. Daw, A common model explaining flexible decision making, grid fields and cognitive control. bioRxiv: 856849 (10 December 2019).
89. T. J. Madarasz, T. E. Behrens, Better transfer learning with inferred successor maps. *Adv. Neural Inf. Process. Syst.* arXiv:1906.07663 (18 June 2019).
90. R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, Y. Niv, Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–278 (2014).
91. L. A. Bradfield, A. Dezfouli, M. Van Holstein, B. Chieng, B. W. Balleine, Medial orbitofrontal cortex mediates outcome retrieval in partially observable task situations. *Neuron* **88**, 1268–1280 (2015).
92. N. W. Schuck, M. B. Cai, R. C. Wilson, Y. Niv, Human orbitofrontal cortex represents a cognitive map of state space. *Neuron* **91**, 1402–1412 (2016).
93. E. Vertes, M. Sahani, A neurally plausible model learns successor representations in partially observable environments. arXiv:1906.09480 (22 June 2019).
94. N. Burgess, T. Hartley, Orientational and geometric determinants of place and head-direction, *Adv. Neural Information Processing Systems* **14**, 165–172 (2002).
95. M. Franzius, H. Sprekeler, L. Wiskott, Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Comput. Biol.* **3**, e166 (2007).
96. J. R. Hinman, G. W. Chapman, M. E. Hasselmo, Neuronal representation of environmental boundaries in egocentric coordinates. *Nat. Commun.* **10**, 2772 (2019).
97. A. Barreto, S. Hou, D. Borsa, D. Silver, D. Precup, Fast reinforcement learning with generalized policy updates. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.1907370117 (2020).
98. T. D. Kulkarni, A. Saeedi, S. Gautam, S. J. Gershman, Deep successor reinforcement learning. arXiv:1606.02396 (8 June 2016).
99. N. Burgess, A. Jackson, T. Hartley, J. O'Keefe, Predictions derived from modeling the hippocampal role in navigation. *Biol. Cybern.* **83**, 301–312 (2000).
100. R. M. Grieves, É. Duvelle, P. A. Dudchenko, A boundary vector cell model of place field repetition. *Spatial Cognit. Comput.* **18**, 217–256 (2018).
101. W. de Cothi, C. Barry, Neurobiological successor features for spatial navigation. *Hippocampus*, 10.1002/hipo.23246 (2020).
102. J. M. Pearce, G. Hall, A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
103. R. A. McDougal *et al.*, Twenty years of ModelDB and beyond: Building essential modeling tools for the future of neuroscience. *J. Comput. Neurosci.* **42**, 1–10 (2017).

PNAS
www.pnas.org

Supplementary Information for

A general model of hippocampal and dorsal striatal learning and decision making

Jesse P. Geerts, Fabian Chersi, Kimberly L. Stachenfeld & Neil Burgess

Neil Burgess
E-mail: n.burgess@ucl.ac.uk

This PDF file includes:

Supporting Information Text

We describe arbitration in our model in more detail. We then describe task-specific adaptations that were made to the model, and some additional experiments.

Arbitration between hippocampal and striatal systems

We implemented arbitration between the hippocampal and dorsal striatal systems in our model using a rule introduced by Wan Lee et al. (1). These authors suggested that arbitration between model-based and model-free systems was done based on a *reliability* signal. They also used fMRI to show that inferior lateral prefontal and frontopolar cortex encode such reliability signals, as well as the output of a comparison between these signals. Furthermore, they showed evidence that the connectivity between these regions and model-free value areas is negatively modulated by the degree of model-based control.

Here, we applied their method to arbitration between a hippocampal system based on the Successor Representation and a striatal system based on model-free learning. The idea is that the reliability of both systems is tracked by computing the recent average of prediction errors of both systems. The Pearce-Hall update rule for tracking average prediction error is:

$$\Delta\Omega = \eta(|\delta| - \Omega) \quad [1]$$

where $|\delta|$ is the absolute RPE and η is a learning rate. The reliability is defined as:

$$\chi = (\delta_{MAX} - \Omega)/\delta_{MAX} \quad [2]$$

with δ_{MAX} being the upper bound of the prediction error, which was set to 1. After each episode, the reliability of each system was updated using the following rule:

$$\Delta\chi = \eta\left[\left(1 - \frac{|\delta|}{\delta_{MAX}}\right) - \chi\right] \quad [3]$$

This measure goes to zero as the average prediction error increases ($\Omega \to \delta_{MAX}$), and goes to one as the average prediction error decreases ($\Omega \to 0$).

Following Wan Lee et al. (1), we can use the reliability measure for arbitration. These authors computed transition rates α and β for transitioning from MF to MB states and vice versa as follows. Here we use the same terms but for transitions between MF and SR. These transition rates are functions of the reliability of the respective systems:

$$\alpha(\chi_{MF}) = \frac{A_\alpha}{1 + \exp(B_\alpha \chi_{MF})} \quad [4]$$

$$\beta(\chi_{SR}) = \frac{A_\beta}{1 + \exp(B_\beta \chi_{SR})} \quad [5]$$

where the A and B parameters in both equations determine transition rate and the steepness of these curves, respectively. These parameters were fitted to behavioural data by Wan Lee et al. (1) and we matched their parameter values (see Table S1).

At each time step, the rate of changes of the probability of choosing the SR system P_{SR} was computed using the following differential equation:

$$\frac{dP_{SR}}{dt} = \alpha(\chi_{MF})(1 - P_{SR}) - \beta(\chi_{SR})P_{SR} \quad [6]$$

Although not explored here (but see 1), this means that there is a certain "stickiness" to the model: if the model is currently choosing MF actions, it will take some time to move weight to the MB system.

Following Wan Lee et al. (1), state action value estimates were given by a weighted average of the two model components:

$$Q(s,a) = P_{SR}Q_{HPC}(s,a) + (1 - P_{SR})Q_{DLS}(s,a) \quad [7]$$

Thus, the degree to which a system contributes to the value estimate is influenced by its reliability. Given these full-model state-action values, the agent chose actions following a softmax policy:

$$\pi(a|s) = \frac{e^{\tau^{-1}Q(s,a)}}{\sum_{a'} e^{\tau^{-1}Q(s,a')}} \quad [8]$$

where τ^{-1} is an inverse temperature parameter which sets the balance between exploration and exploitation. The higher the inverse temperature, the more the agent chooses higher-valued actions.

Task-specific adaptations

Although the general model architecture remained the same throughout all simulations, different adaptations were made to the model described above such that it could be used in the differen state spaces defined by the tasks.

Plus maze. For the Plus Maze task described in Fig.3, landmark cells were tuned to the ends of the maze. We assumed that the landmark cells could not distinguish between the two ends of the maze such that, from the point of view of the striatal system, probe trials and training trials looked the same.

Blocking. For the blocking simulations (Fig.4), we adapted the hippocampal controller (that worked with a tabular state representation as input) to incorporate the effects of boundaries on place cell firing. To that end, we defined the hippocampal SR system using linear function approximation. The agent observes states through a vector of features $\mathbf{f}(s)$ which, if chosen rightly, will be of much smaller dimension than the number of states, allowing the agent to generalise to states that are nearby in feature space. The feature-based SR (2) encodes the expected discounted future activity of each feature:

$$\psi^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} \mathbf{f}(s_t) | s_0 = s\right] \quad [9]$$

As in the tabular case, the feature-based SR can be used to compute value when multiplied with a vector of reward expectations per feature, $\mathbf{u}$: $V^{\pi}(s) = \psi^{\pi}(s)^{T}\mathbf{u}$. In the case of linear function approximation, these Successor Features ψ in Equation 9 are approximated by a linear function of the features $\mathbf{f}$:

$$\hat{\psi}(s) = W^{T}\mathbf{f}(s), \quad [10]$$

where W is a weight matrix which parameterises the approximation.

In the context of hippocampus, the feature-based SR allows us to represent states as population vectors of place cells with overlapping firing fields (the features), rather than having a one-to-one correspondence between place cells and states. Then we are free to model the dependence of the place cell firing on specific environmental features (boundaries). This dependence has been extensively characterised by computational models of boundary vector cells (BVCs) (3–7), which were shown to exist in the subiculum (8). Accordingly, we modelled a set of hippocampal place cells whose activity $\mathbf{f}_i(s_t)$ was the thresholded sum of a set of BVC inputs (see 5, for details on how BVC and place cell maps were calculated).

Thus, at every state s (corresponding to a location) in the environment, the agent observed a population vector $\mathbf{f}(s)$ of BVC-driven place cells (see Fig.S1 for an example). It then computed its estimated Successor Features ψ using its current estimate of weights W and Equation 10, which encode the discounted sum of future population firing rate vectors $\mathbf{f}$ of the input place cells. In terms of circuitry, W might correspond to the Schaffer collaterals projecting from CA3 to CA1 neurons, corresponding to $\mathbf{f}$ and ψ, respectively.

As in the tabular case, temporal difference learning can be used to update the SR weights:

$$\Delta W = \alpha\left[\mathbf{f}(s_t) + \gamma\psi(s_{t+1}) - \psi(s_t)\right]\mathbf{f}(s_t)^{T} \quad [11]$$

Note that the algorithm has not changed with respect to the one-hot state encoding mentioned earlier – it is easy to see that the function approximation version reduces to the tabular case when $\mathbf{f}$ is a one-hot vector. The reward expectation vector $\mathbf{u}$ was updated using a simple delta rule:

$$\Delta\hat{\mathbf{u}} = \alpha\left(r_t - \hat{\mathbf{u}}^{T}\mathbf{f}\left(s_t\right)\right)\mathbf{f}\left(s_t\right) \quad [12]$$

Two-step tasks. For the non-spatial two-step tasks (Fig.5 and Fig.S3), the DLS cells were assumed to provide a one-hot representation of the task states. While this is significantly different from the landmark cell representation used in the spatial navigation studies, this representation reflected the fact that states were uniquely identifiable as different images. Furthermore, this is consistent with experimental evidence showing that dorsal striatum represents reward-predictive cues (9).

Hippocampal damage in the two-step and spatial tasks. In order to mimic the individual differences between participants found by Vikbladh et al. (10), we sampled 20 different agents with varying values for the parameters governing the transition from SR to MF and vice versa (see Equations 4 and 5). Specifically, we sampled A_{α} values (steepness of the transition from MF to SR) uniformly between .5 and 5, and A_{β} (steepness of the transition from SR to MF) values uniformly between 2 and .5. In addition to the 20 "full agents", we sampled 20 agents for which the hippocampal component was partially inactivated by setting a maximum to the P_{SR}. To mimic variability in the size of the lesion that was present in the dataset of Vikbladh et al. (10), we sampled max P_{SR} values from a uniform distribution between 0 and 0.35.

Quantification and statistical analysis

To investigate the relationship between the agents' spatial navigation and non-spatial decision making strategies, we quantified the agents' degree of MB planning, as well as their degree of using an allocentric strategy, and computed their correlation.

For quantifying MB planning, we followed earlier studies (10, 11) and analysed the agents' choices using a mixed-effects logistic regression (estimated using the *statsmodels* Python package, (12)). For each trial, the dependent variable (stay with the same first-level action or switch) was explained in terms of whether there was a reward on the previous trial, whether the previous transition was of the rare or common type, and the interaction between these factors. The logic of the two-step task is that an MB learner will stay with the same action if it was rewarded after a common transition, but will be more likely to switch if it gets rewarded after a rare transition. Thus, the degree of MB planning can be quantified as the interaction between previous reward and trial type.

For quantifying the degree of allocentric place memory, we computed the average distance between the previous platform location and the location of the maximum of the agent's value function at the start of the next session. This is akin to the boundary distande error employed by (13).

After computing the correlation between allocentric place memory and MB planning for both the "healthy" and "lesioned" groups of agents, we asked whether the two correlation coefficients were significantly different from each other by applying the Fisher z-transform (14) to the coefficients, and testing whether the difference between the transformed coefficients was significantly different from zero.

For the cued water maze task described in Fig.S2, the differences among the groups in relation to the number of agents that chose a place or a cue strategy were analysed by the Fisher exact test as implemented in R (15).

Additional tasks

Cue versus place Water Maze. In addition to the hippocampal lesion described in Fig.2, we simulated a DLS lesion in the task used by Pearce et al. (16). Fig.S2A shows the simulation results: there is little to no learning across sessions for the first trials of each session, indicating impaired acquisition of the landmark-platform association. Fourth-trial performance is not significantly worse than control performance, which is a sign of intact place learning as agents still learn during a session in which the platform has a fixed location. This is consistent with a previous finding showing that dopamine depletion in the DLS impairs egocentric but not allocentric Water Maze navigation (17). Fig.S2B shows results from a study by Myoshi et al. (18) that investigated the effects of bilateral lesions of the hippocampus, DLS or both in a cue on a probe test in the Water Maze. Animals were trained to swim to a given location in the Water Maze, that was indicated by the presence of a landmark. Then, during a probe trial, the landmark was placed elsewhere in the maze, and the animals' behaviour wass classified as cue-guided if the animal swam directly to the cued platform, as place-guided if it swam directly to the place the hidden platform was the day before, or as thigmotaxic if the animals swam around the edge of the pool. This dual-solution probe trial is akin to the first trial of each session in Pearce et al. (16). Fig.S2C shows that our simulations accurately capture these results, where we classified behaviour as "cue" or "place" guided if the agent reached the platform as indicated by the landmark or previous location within a given number of time teps (60), and as "neither" otherwise.

Deterministic two-step task. In the experiment designed by Doll et al. (19), human participants where shown a pair of two pictures from one of two categories (faces or tools) and were asked to choose one. This was defined as the start state. The participants' initial choice determined which of two second-stage states they would transition to. These second stage states corresponded to a choice from a pair of pictures from one of two new categories (scenes or body parts; see Figure S3A). Each second-stage option (the 'outcome') was either rewarded with money or not rewarded. The reward probability for each outcome drifted slowly and randomly such that participants continuously learned by trial and error which second-stage choices were 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 choices 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 dissociates behaviour consistent with MB and MF learning. A model-based learner represents transition probabilities, and uses this transition model to compute the best action. Thus, when a model-based learner encounters a reward, this should affect its behaviour in the next trial regardless of whether it 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 MF 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. Consistent with humans making use of both strategies, Doll and colleagues showed that human performance on this task lies somewhere in between these strategies (Figure S3B).

Our model recapitulates the main effects found by Doll and colleagues. The SR model mimics model-based behaviour by separating reward information from information about the transition structure. When the goal is reached, value is generalised to states that predict the goal states. Thus, following reward, the hippocampal model will learn to take actions to end up in the same second stage state in the next trial, regardless of whether it has the same or different starting state (Figure S3C). 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 S3C). Combining these two models gives a pattern of behaviour in between model-based and model-free, akin to human performance. However, in contrast to our model, human participants showed a higher stay probability for the "same starting state" condition than for the "different starting state" condition. This propensity to stay with the same action does not follow directly from a MF/MB trade-off.

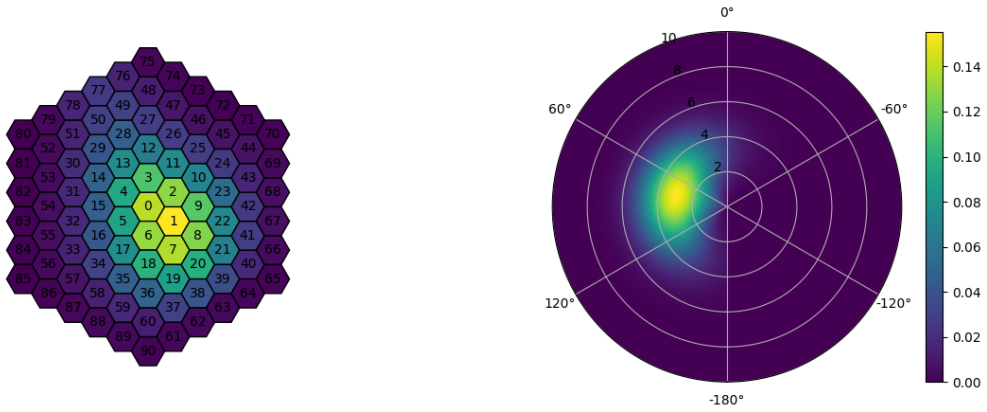

Fig. S1. Example receptive fields. Left panel: Example SR place cell map in a discretised maze. Right panel: Example landmark cell receptive field plotted in polar coordinates.

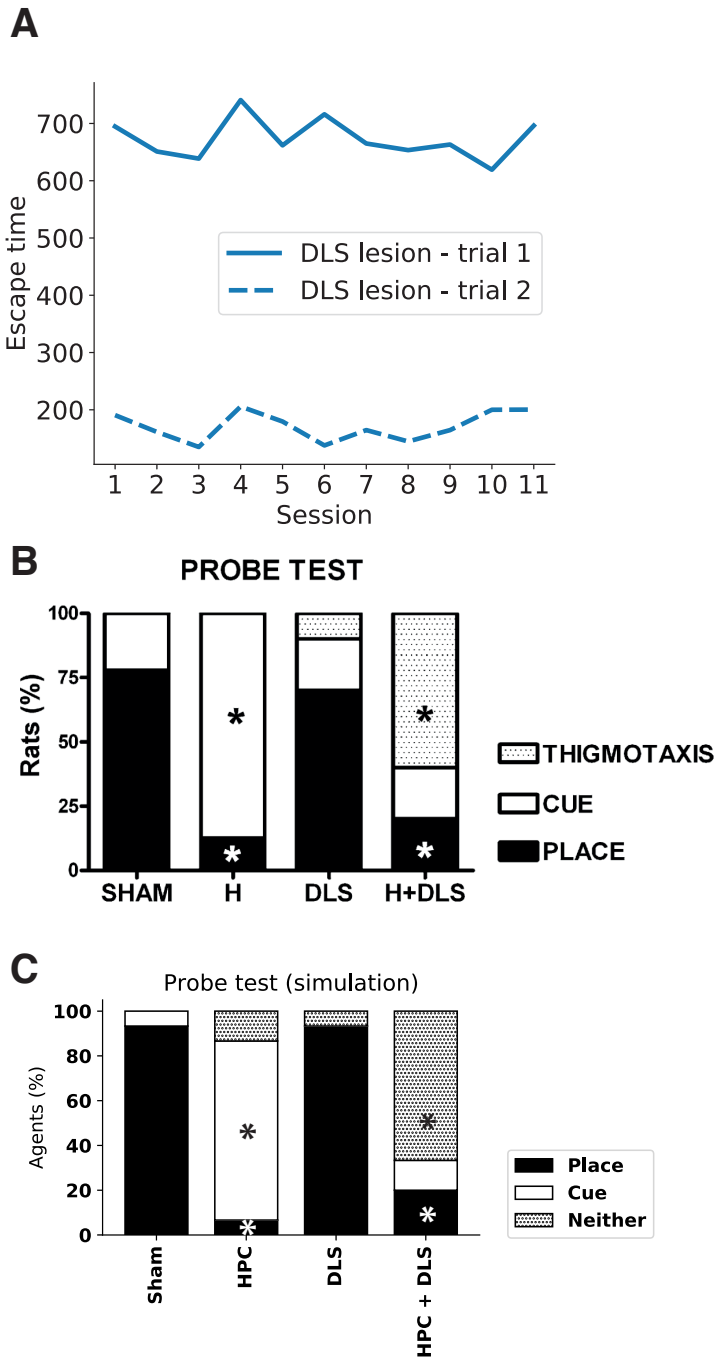

Fig. S2. (A) Simulation result of a DLS lesion in the Pearce et al. (16) study, showing escape time on the first and fourth trial of each session. Landmark and platform were moved together after every session. (B) Data from Myoshi et al. (18) showing the effects of bilateral lesions of the dorsal hippocampus (H) and/or the dorsolateral striatum (DLS) on a probe test carried out after 5 days of training on the Morris Water Maze. Data express the proportion of rats that (i) swam directly to the cued platform, (ii) to the place the hidden platform was the day before, or (iii) exhibited thigmotaxic swimming behaviour (swimming around the edges of the pool) in the first trial in the cued version. *P < 0.05 compared to SHAM animals; Fisher test. (C) Simulation results showing the effects of ablating the HPC and/or DLS model components on the task described in (B). * correspond to P < 0.05 in a Fisher test compared to SHAM animals/agents in both (B) and (C).

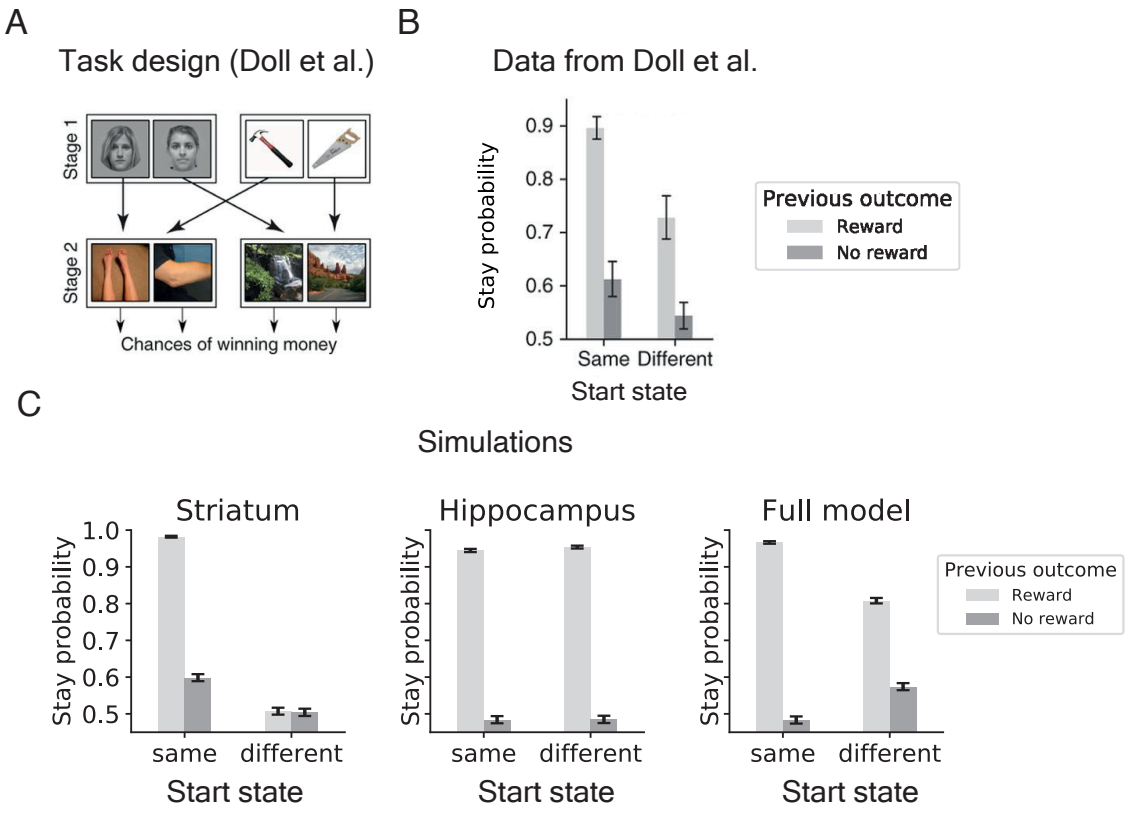

Fig. S3. (A) Task stucture employed by (19). (B) The probability that human participants in Doll et al. (2015) chose the same first-stage action as on the previous trial binned by whether the previous choice was rewarded, and whether they started in the same state. (C) Simulation results. The hippocampal model mimics the true MB agent presented in the original paper. The striatal model shows MF behaviour. Combining the two models results in a behavioural pattern that shows both effects. As in (19), MB behaviour was quantified as the main effect of previous reward on choice behaviour (estimate=.96, $Z = 4.3$, $P = 1.66 \times 10^{-5}$). This effect is greater when the current state is the same as the previous one (estimate=2.37, Z=6.64, $P = 3.18 \times 10^{-11}$), indicating the presence of MF behaviour.

Name	Symbol	Value
SR learning rate	α_M	0.07
Q learning rate	α_Q	0.07
Softmax inverse temperature (exploration parameter)	τ^{-1}	5
Discount parameter	γ	0.95
Reliability learning rate	η	0.03
Maximum prediction error	δ_{MAX}	1
Steepness of transition curve MF to SR	A_α	3.2
Steepness of transition curve SR to MF	A_β	1.1

Table S1. Parameters

References

1. S Wan Lee, S Shimojo, JP O'Doherty, Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron* **81**, 687–699 (2014).
2. A Barreto, R Munos, T Schaul, D Silver, Successor Features for Transfer in Reinforcement Learning. *arXiv*, 1–13 (2016).
3. N Burgess, A Jackson, T Hartley, J O'keefe, Predictions derived from modelling the hippocampal role in navigation. *Biol. cybernetics* **83**, 301–312 (2000).
4. T Hartley, N Burgess, C Lever, F Cacucci, J O'Keefe, Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus* **10**, 369–379 (2000).
5. C Barry, et al., The boundary vector cell model of place cell firing and spatial memory. *Rev. Neurosci.* **17**, 71–98 (2006).
6. RM Grieves, É Duvelle, PA Dudchenko, A boundary vector cell model of place field repetition. *Spatial Cogn. & Comput.* **18**, 217–256 (2018).
7. W de Cothi, C Barry, Neurobiological successor features for spatial navigation. *Hippocampus*, 1–9 (2020).
8. C Lever, S Burton, A Jeewajee, J O'Keefe, N Burgess, Boundary vector cells in the subiculum of the hippocampal formation. *J. Neurosci.* **29**, 9771–9777 (2009).
9. MAA van der Meer, A Johnson, NC Schmitzer-Torbert, AD Redish, Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron* **67**, 25–32 (2010).
10. OM Vikbladh, et al., Hippocampal Contributions to Model-Based Planning and Spatial Memory. *Neuron* **102**, 683–693.e4 (2019).
11. ND Daw, SJ Gershman, B Seymour, P Dayan, RJ Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
12. S Seabold, J Perktold, statsmodels: Econometric and statistical modeling with python in *9th Python in Science Conference*. (2010).
13. O Vikbladh, et al., Two Sides of the Same Coin: The Hippocampus as a Common Neural Substrate for Model-Based Planning and Spatial Memory. *bioRxiv* (2018).
14. RA Fisher, Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507–521 (1915).
15. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), (2013) ISBN 3-900051-07-0.
16. JM Pearce, ADL Roberts, M Good, Hippocampal lesions disrupt navigation based on cognitive maps but not heading vectors. *Nature* **62**, 1997–1999 (1998).
17. AA Braun, et al., Dopamine depletion in either the dorsomedial or dorsolateral striatum impairs egocentric Cincinnati water maze performance while sparing allocentric Morris water maze learning. *Neurobiol. Learn. Mem.* **118**, 55–63 (2015).
18. E Miyoshi, et al., Both the dorsal hippocampus and the dorsolateral striatum are needed for rat navigation in the Morris water maze. *Behav. Brain Res.* **226**, 171–178 (2012).
19. BB Doll, KD Duncan, DA Simon, D Shohamy, ND Daw, Model-based choices involve prospective neural activity. *Nat. Neurosci.* **18**, 767–772 (2015).