Chapter 12: Predictive maps

There's more to life than reward. And even if you're only trying to maximize reward, there are situations where it's still useful to know about more than just reward. These knowledge representations are useful (in the sense of reward maximization) to the extent that they can be deployed efficiently for action selection. This chapter studies the flexibility-efficiency trade-off entailed by different representations and the algorithms that operate on them. We develop a generalization of the value function concept that predicts a multiplicity of different targets (future states or features). This *predictive map* enables more flexibility than the value function alone, with only a modest efficiency loss. Evidence suggests that this predictive map is represented in the hippocampus and updated by dopaminergic temporal difference errors or a temporally asymmetric learning rule. A low-dimensional representation in entorhinal cortex may regularize learning toward predictions that are smooth across the state space.

The reinforcement learning (RL) models described in the last two chapters focus on learning representations (value functions and policies) that facilitate optimal prediction and behavior with relatively little computation. In the simplest case, value functions and policies are stored in look-up tables, requiring no computation beyond finding the right entry for a given state-action pair. In a more complex variation, these functions are neural networks, still relatively cheap. The trade-off is that these representations are relatively inflexible: if the world changes, the representations may need to be relearned.

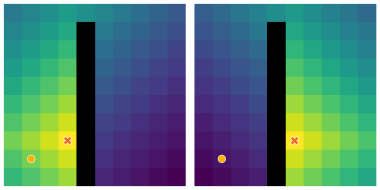

Figure 1: **Gridworld illustration of how local changes have non-local effects on the value function**. The goal is shown as an X and the agent is shown as a circle. Black squares are walls. The shading of the other squares shows the value function. The two gridworlds differ only by the goal placement.

For example, a TD learning agent can estimate a tabular value function $\hat{V}(s)$ by interacting with the environment. This may be a very time-consuming process if the state space is large, because the agent will have to sample each state many times. Critically, a local change in the environment has non-local changes on the value function (Figure 1) because the value of each state depends on the values of other states (through the Bellman equation; see Chapter 10). As a consequence, the states may need to be comprehensively resampled in order to update the value function. In settings where experience is scarce, relearning may be prohibitively costly. Importantly, animals don't always need to relearn. They are sometimes capable of rapid adaptation. How do they do it?

Under uniform sampling of states, an accurate estimate requires a total number of samples that is at least quadratic (and possibly worse) in the number of states (Samsonov et al., 2024).

This chapter will discuss one approach to this problem (another will be discussed in the next chapter). We will show how the concept of a value function can be generalized to encompass predictions of many different targets (not just cumulative reward). These "generalized predictions" can be used to support rapid adaptation. We will

review behavioral and neural evidence for generalized predictions, as well as for generalized prediction errors that drive learning.

1 *Latent learning*

To build some intuition, let's return to the phenomenon of *sensory preconditioning* introduced in Chapter 10. In the first stage of training, one CS (B) is preceded by another CS (A) in the absence of reward. Subsequently, B is paired with reward. Finally, a test (again in the absence of reward) reveals conditioned responding to A, despite never being paired with reward. In Chapter 10, we noted that a probabilistic version of TD learning could explain the sensory preconditioning phenomenon, due to the posterior covariance matrix that guides indirect credit assignment.

Notice that this is essentially the same design as second-order conditioning, except that the A-B trials are confined to the first phase.

One problem with this account is that it requires the offset of A to coincide with the onset of B in order to learn a positive covariance (the basis of the sensory preconditioning effect in the model). If A and B are simultaneously presented, the model learns a *negative* covariance; this is what underlies retrospective revaluation phenomena like backward blocking discussed in Chapter 10. Contrary to this model prediction, sensory preconditioning still occurs with simultaneous presentation (Rescorla, 1980). Moreover, some studies find that sensory preconditioning is stronger with shorter interstimulus intervals (e.g., Hoffeld et al., 1958; Lavin, 1976), again at odds with the probabilistic TD model's predictions.

Intuitively, sensory preconditioning reflects the formation of predictive links between arbitrary stimuli, which should be stronger when the temporal separation between stimuli is small. These links can then be mobilized to support reward prediction and action selection. To formalize this idea, we will describe an alternative model—the *successor representation* (SR)—which doesn't involve Bayesian machinery like covariance matrices. Instead, it is based on learning long-range state predictions. Collectively these state predictions form a predictive map of the environment.

The predictive map can explain a broader class of "latent learning" phenomena where an animal is exposed to mazes, stimulus sequences, or other environments in the absence of reinforcement; subsequent tests reveal that the animal has acquired knowledge about the structure of these environments. For example, Tolman (1948) described experiments in which rats were allowed to freely explore a maze; after a week, the animals started receiving food when they reached a particular location of the maze. These animals learned to find the food much more quickly compared to the learning curves of naive animals, indicating that they acquired knowledge about the

structure of the maze which they could use to guide their reward-seeking.

Another striking form of latent learning is observed in contextual fear conditioning experiments, where a shock is paired with a context (e.g., a particular box). If the shock is delivered immediately after placing the animal in the context, a subsequent test of conditioned fear (freezing in the context) is *weaker* compared to the same procedure with a time delay between context entry and shock delivery—a phenomenon known as the *immediate shock deficit* (Fanselow, 1990). One way to think about this phenomenon is that the additional time is required for the animal to explore the box and form a predictive map, which then allows the shock delivery to propagate across the whole state space.

2 *The successor representation*

Our starting point is the following decomposition of the value function (Dayan, 1993):

$$V^\pi(s) = \sum_{\tilde{s}} M^\pi(s,\tilde{s}) R(\tilde{s}), \tag{1}$$

where

$$M^\pi(s,\tilde{s}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathbb{I}[s_t = \tilde{s}] | s_0 = s, \pi\right] \tag{2}$$

is the SR, a matrix that keeps track of the expected discounted number of times an agent visits state s' on a trajectory starting in state s. It can be thought of intuitively as a kind of "predictive map" specifying where the agent will be in the near future (with a predictive horizon set by the discount factor γ). An agent currently occupy state s can inspect the corresponding row of M to see what the future holds in store.

In matrix form, the SR can be computed analytically: $M^\pi = (I - \gamma T^\pi)^{-1}$, where I is the identity matrix and $T^\pi(s,s') = \sum_a \pi(a|s) T(s'|s,a)$ is the state transition matrix.

Eq. 1 says that the value of state s is a linear combination of immediate expected rewards in other states (s') weighted by the frequency of visiting each state in the near future starting from s. This is a powerful result: given the SR, values can be computed by a linear operation on the reward function.

The SR obeys a Bellman equation analogous to the one for values:

$$M^\pi(s,\tilde{s}) = \mathbb{I}[s = \tilde{s}] + \gamma \sum_{s'} T^\pi(s,s') M^\pi(s',\tilde{s}). \tag{3}$$

This means that TD learning can be used to update an estimate of the SR, $\hat{M}^\pi$, from observed state transitions ($s \to s'$):

$$\Delta \hat{M}^\pi(s,\tilde{s}) \propto \mathbb{I}[s = s'] + \gamma \hat{M}^\pi(s',\tilde{s}) - \hat{M}^\pi(s,\tilde{s}). \tag{4}$$

The right-hand side is the TD error applied to long-range state prediction. The immediate rewards can also be learned by a simple error-driven update after observing reward r in state s:

$$\Delta\hat{R}(s) \propto r - \hat{R}(s). \tag{5}$$

Note that reward estimates can be learned more quickly than value or SR estimates, because they don't rely on bootstrapping future predictions, as in the TD update.

The key feature of this model is the separation of state prediction and reward prediction. This allows the model to flexibly reuse state predictions when the reward changes.

Let's now examine how this model handles sensory preconditioning. After the initial stimulus-stimulus training, $\hat{M}^\pi(A, B) = \gamma^{\text{ISI}}$, where ISI is the interstimulus interval between stimuli A and B. This formalizes the idea that the predictive link between stimuli diminishes with temporal separation. When B is paired with reward, the reward expectation is propagated to A via the predictive link, $\hat{V}^\pi(A) = \gamma^{\text{ISI}}\hat{R}(B)$. Thus, the SR model explains the key features of sensory preconditioning.

Note that the policy π is irrelevant for Pavlovian protocols, but we retain it for notational consistency.

A similar approach can be used to explain other forms of latent learning. For example, context pre-exposure in a contextual fear conditioning paradigm provides the SR with data to form a predictive map of the environment, such that the delivery of shock (a change to the reward function coincident with the agent's current location) is propagated to other states in the environment. This is why eliminating this pre-exposure attenuates conditioned fear to the context: upon return to the context, the animal is likely placed in a location different from the one that coincided with the shock, so it will only produce a fear response if the shock expectation has been propagated across the state space.

Latent learning is fundamentally at odds with the "model-free" algorithms described in Chapters 10 and 11, which provide no mechanism for learning in the absence of reinforcement (the exception is the probabilistic TD model, though arguably this is not purely model-free). The SR account is not fully model-based, since one cannot in general recover the transition function from the SR. The account of sensory preconditioning and latent learning depends on a partial model (the reward function) combined with a model-free representation of state predictions (the SR). Below we will discuss fully model-based approaches to RL.

Here "model-free" is used to denote any algorithm that does not use a model of the environment, which in this setting corresponds to the transition and reward function.

3 *The flexibility-efficiency trade-off*

The SR lies somewhere in between pure model-free and pure model-based algorithms in terms of its flexibility-efficiency trade-off. Compared to model-free algorithms, it is (as we've already seen) more flexible in its ability to adapt to changes in the reward function. Be-

cause it represents the full pairwise predictive map over states, the computational complexity of state evaluation (Eq. 1) is linear in the number of states. Thus, it is potentially less efficient compared to a model-free algorithm with a lower-dimensional representation, though this is not a fundamental limitation once we consider feature-based generalizations of the SR (Section 7). Model-based evaluation achieves maximal flexibility, since it can adapt to any change in the reward and transition functions, but at the cost of higher computational complexity (depending on how exactly values are computed; see next chapter).

One implication of using the SR for state evaluation is that it should be differentially sensitive to changes in the transition function compared to the reward function. This is because the SR achieves its efficiency by compiling transition information in long-range state predictions. A change in the transition function induces non-local changes in the SR, analogous to the non-local changes induced in the value function. In contrast, changes to the reward function induce local changes, because this information is represented separately. Model-based evaluation should be equally sensitive to both transition and reward function changes, whereas model-free evaluation should be equally insensitive.

Momennejad et al. (2017) tested these predictions in a human RL task. The experiment consisted of three phases (Figure 2): learning, relearning, and test (in extinction). At the end of Phase 1 (learning) and Phase 3 (test), participants provided a continuous valued rating indicating which of the two starting states they preferred. Revaluation was greater in the reward revaluation condition compared with the transition revaluation condition, consistent with the prediction of SR-based evaluation. Nonetheless, humans were not completely insensitive to changes in the transition structure, indicating that they may rely partially on model-based evaluation in this task. One hypothesis is that they rely on SR-based evaluation for an initial approximation, which can be refined through additional model-based computation.

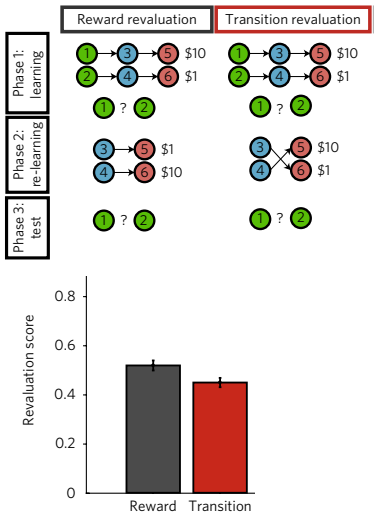

Figure 2: **Comparing transition and reward revaluation procedures**. (Top) Experimental design. Circles denote states and arrows denote transitions. (Bottom) Behavioral results. Revaluation score denotes the change in preference rating after versus before the relearning phase. Adapted from Momennejad et al. (2017).

4 *A predictive map in the hippocampus*

If the SR account is correct, we should be able to find predictive maps in the brain. One place to look is the hippocampus, because lesions of the hippocampus impair several forms of latent learning (e.g., Port et al., 1987; Talk et al., 2002; Wiltgen et al., 2006), and stimulus-stimulus predictions are encoded in hippocampal activity during latent learning (Garvert et al., 2017; Wang et al., 2020).

There is a long tradition of thinking about the hippocampus as

The evidence for hippocampal dependence of latent learning is controversial, with some studies failing to find deficits following lesions (Kimble and BreMiller, 1981; Ward-Robinson et al., 2001).

Although we focus on the hippocampus in this section, we note that there is evidence for predictive maps in other regions, such as visual cortex (Ekman et al., 2023). See Chapter 3 for more discussion of predictive representations in the brain.

representing some form of "cognitive map" (O'Keefe and Nadel, 1978), originally motivated by the discovery of place cells in the hippocampus tuned to an animal's location. It later became evident that place cells were not simply registering an animal's location. For example, when rats repeatedly traverse a linear track, place cells gradually expand their tuning in the direction opposite of travel (Mehta et al., 2000). In other words, a cell that initially responded when the rat was in a particular location on the track eventually began to respond to earlier locations. Another example comes from a study of rats in a circular water maze, where the rats needed to find a submerged platform to escape the water (Hollup et al., 2001). Place fields (tuning functions) were concentrated near the platform location (i.e., a disproportionate number of place cells were tuned to locations near the platform). Both of these studies illustrate asymmetries in the structure of place cell tuning that can't be explained by a purely spatial account.

The predictive map offers a different account of these phenomena (Stachenfeld et al., 2017). Assume that each state corresponds to a location, and the place field for a single cell corresponds to one column of the SR estimate, $\hat{M}^\pi(\cdot, \tilde{s})$; we will refer to this as the *SR field* tuned to location $\tilde{s}$ (Figure 3). An SR field is a *retrospective representation* in the sense that activity is maximal for states leading to $\tilde{s}$. A row of the SR, $\hat{M}^\pi(s, \cdot)$, corresponds to the population code for state s. This is a *predictive representation* in the sense that activity is maximal for cells tuned to states leading from s.

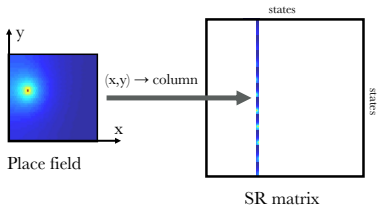

Figure 3: **Place field corresponding to a single column of the SR**. Reproduced from Gershman (2018).

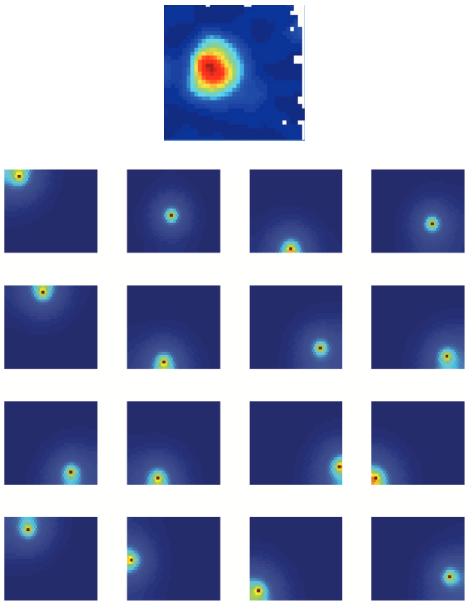

Figure 4: **Real and simulated place fields in a random foraging paradigm**. (Top) Typical place field, reproduced from Moser et al. (2015). (Bottom) SR place fields, reproduced from Stachenfeld et al. (2017).

Place cells are traditionally identified using a random foraging paradigm in which rats are placed in a box and food pellets are randomly distributed, which induces the rat to explore the whole box. In this setting, with no systematic asymmetries in the direction of travel, SR fields will resemble classical place fields—approximately symmetric around the preferred location (Figure 4; though see Section 7 for further nuances on this point). If rats are instead trained on the linear track or circular water maze, systematic asymmetries in the direction of travel will bias the SR fields backwards in the direction of travel (reflecting the retrospective nature of the representation; Figure 5) and concentrate them near the platform (reflecting the shift in tuning toward frequently visited locations). Thus, the predictive map can explain both classical and non-classical aspects of place fields.

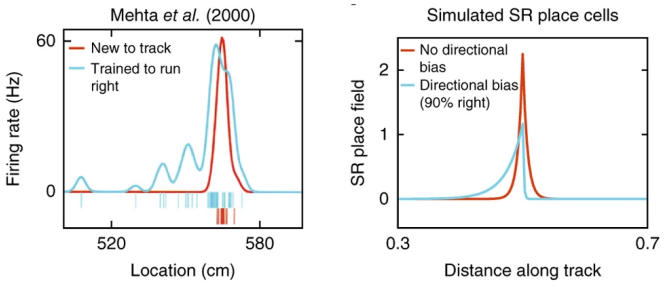

Figure 5: **Real and simulated place cells skew backwards relative to the direction of travel**. Reproduced from Stachenfeld et al. (2017).

The SR depends on the discount factor γ, which controls the predictive horizon. Thus, in principle one could have multiple predictive maps with different choices of γ, yielding a multi-scale representation. This could be useful for adaptively solving tasks with diverse temporal demands—i.e., where some tasks require longer-range prospection than others. In fact, the hippocampus contains multiple maps at different scales, arranged topographically along its dorsoventral axis (Kjelstrup et al., 2008), where more dorsal neurons have smaller place fields. This gradient can be captured by changing γ, as shown by Stachenfeld et al. (2017).

5 *Learning the predictive map with dopamine*

Since we've already shown that the SR can be estimated using TD learning, a natural hypothesis is that dopamine provides the TD errors needed for this learning mechanism. This "generalized prediction error" hypothesis can resolve a number of puzzles concerning dopamine (Gardner et al., 2018).

The picture of phasic dopamine developed in the last two chapters conceptualizes it purely in terms of reward prediction. However, there are phenomena that seem to contradict this conceptualization.

For example, some dopamine neurons respond to prediction errors not only in the *amount* of reward but also in the *identity* of reward, such as different flavors with similar desirability (Takahashi et al., 2017). At the population level, dopamine neurons encode information about reward identity that diminishes over the course of learning, consistent with a prediction error signal (Stalnaker et al., 2019). If states correspond to reward identities, then the TD update in Eq. 4 can capture the sensitivity of dopamine neurons to sensory prediction errors. If dopamine neurons are tuned to different states, then reward identity will be encoded at the population level as long as the sensory prediction errors are non-zero.

The generalized prediction error hypothesis can also explain the results of experiments in which dopamine activity was manipulated. For example, Sharpe et al. (2017) attenuated the sensory preconditioning effect by inhibition of dopamine neurons, demonstrating the necessity of dopamine signaling for stimulus-stimulus learning. This finding is consistent with recordings of dopamine release specifically in response to sensory prediction errors during sensory preconditioning (Costa et al., 2025).

If the hippocampus encodes the SR and dopamine is used to update the SR, then it stands to reason that dopamine inputs to the hippocampus should be critical for the acquisition of predictions. The evidence for this hypothesis is suggestive but somewhat indirect. The over-representation of high-reward locations in the hippocampus can be attenuated by inhibition of dopamine inputs to the hippocampus (Mamad et al., 2017; Krishnan et al., 2022). It's also well-established that hippocampal plasticity depends on dopamine signaling (Tsetsenis et al., 2023). What's still lacking is direct evidence that sensory prediction errors encoded by dopamine drive plasticity of the hippocampal predictive map.

6 *A recurrent neural network implementation*

Rather than relying on an exogenous error signal, another possibility is that mechanisms endogenous to the hippocampus could compute and learn the SR. In this section, we will focus on one implementation of this idea proposed by Fang et al. (2023). Let x_t denote an input to a linear recurrent neural network (RNN) with recurrent synaptic connectivity matrix Ω, where Ω_{ij} denotes the strength of the synapse connecting neuron i to neuron j. Assuming a time step of $\Delta t = 1$, the firing rate dynamics of the recurrent units (with activity z) can be written in discrete time as:

Both x and z are row vectors.

$$z_{t+1} = x_t + \gamma z_t \Omega, \tag{6}$$

where the discount factor γ plays the role of a synaptic gain. If $x_t = x$ is a constant one-hot encoding of the state (i.e., $x(s) = 1$ whenever the system is in state s, otherwise $x(s) = 0$), and the connectivity matrix corresponds to the state transition matrix, $\Omega = T^\pi$, then at steady state the recurrent units will converge to the row of the SR corresponding to x (i.e., the predictive map for state s encoded by x):

This is actually an old idea, originating in the early neural network literature (Sutton and Pinette, 1985) and later revisited by Russek et al. (2017).

$$\lim_{t\to\infty} z_t = xM^\pi = M^\pi(s,\cdot). \tag{7}$$

Intuitively, the RNN is progressing the predictions forward at each iteration and adding up the results.

The question now becomes how to learn Ω such that it corresponds to T^π. This requires that a learning rule associate consecutive states together and normalize the synaptic strengths so that they define a valid probability distribution. A learning rule satisfying both requirements is given by:

$$\Delta\Omega \propto z_{t-1}^\top z_t - z_{t-1}^\top z_{t-1}\Omega. \tag{8}$$

The first term implements a form of temporally asymmetric potentiation, similar to spike-timing dependent plasticity (see Chapter 9). The second term is a form of anti-Hebbian depotentiation that ensures normalization. Importantly, the update rule uses only information local to a synapse, and converges to the transition matrix asymptotically. Fang et al. (2023) showed that the update rule still works with more biologically realistic assumptions and can recover quantitative features of recorded place fields.

Anti-Hebbian depotentiation has been observed at hippocampal inhibitory synapses (Lamsa et al., 2007), though these are different from the recurrent synapses (putatively in the CA3 subregion) that would most naturally implement the RNN architecture.

The idea that recurrent synapses learn a transition matrix is appealing for another reason, explored further in the next chapter. While the steady state response corresponds to long-range prediction, the initial response corresponds approximately to the next-state prediction. Thus, the same network could potentially serve as a simulator of the transition model, supporting model-based planning and imagination (see next chapter).

7 *Successor features*

The SR is tabular—it assumes that the state space can be discretely enumerated. Learning a tabular representation requires experiencing each state many times. This is clearly prohibitive in large or continuous state spaces. To address this problem, we develop a feature-based generalization of the SR along the same lines as feature-based value function approximators. In Chapter 10 we introduced a feature vector $x = f(s)$ generated by a population of tuning functions defined over s. If expected rewards can be represented as a linear function of the

features, $R(s) = \sum_d w_d x_d$, then the value function can be represented as a linear function of *successor features* (SFs), $\psi^\pi(s)$:

$$V^\pi(s) = \sum_d w_d \psi_d^\pi(s), \qquad \psi^\pi(s) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t f(s_t) | s_0 = s, \pi\right]. \quad (9)$$

SFs are analogous to value functions, but accumulating general features rather than just reward. This allows an agent to generalize across states, and even across tasks (as explained below). SFs also obey a Bellman equation:

$$\psi^\pi(s) = f(s) + \gamma \sum_{s'} \psi^\pi(s'), \quad (10)$$

and hence can be estimated using TD learning:

$$\Delta\hat{\psi}^\pi(s) \propto f(s) + \gamma\hat{\psi}^\pi(s') - \hat{\psi}^\pi(s). \quad (11)$$

The weights can likewise be updated using an error-driven update (essentially the Rescorla-Wagner model discussed in Chapter 10):

$$\Delta\hat{w} \propto f(s)[r - \hat{w}]. \quad (12)$$

If different subpopulations of dopamine neurons signal TD errors for different features, this model may explain how a single canonical computation (the TD update) is compatible with the diversity of feature sensitivity across dopamine neurons (Engelhard et al., 2019).

De Cothi and Barry (2020) have proposed a particular choice of features based on studies of the hippocampus and adjacent structures. An influential theory (Hartley et al., 2000) posits that place cells are constructed by taking weighted combinations of *boundary vector cells* (Figure 6), which are tuned to the distance and angle to a boundary. Such cells have been identified in subiculum (Lever et al., 2009), classically considered an output region of the hippocampus, though it also provides reentrant input back into the hippocampus via the entorhinal cortex. De Cothi and Barry (2020) showed that the SFs corresponding to boundary vector cells closely resemble hippocampal place fields. Notably, they capture the elongation of place fields along environmental boundaries (Muller et al., 1987), a fact that is not captured by the tabular SR model of place fields. SFs defined over boundary vector cells can also capture the stretching or compression of boundary-adjacent place fields when the environment is enlarged or shrunk along the corresponding axis (O'Keefe and Burgess, 1996).

Figure 6: **Tuning functions of boundary vector cells**.Reproduced from Lever et al. (2009).

Functionally similar *border cells* have been identified in the entorhinal cortex (Solstad et al., 2008).

8 *Spectral regularization by entorhinal grid cells*

The entorhinal cortex provides the principal input to the hippocampus. It contains another important class of neurons, *grid cells*, which

respond periodically across space, organized into hexagonal grids (Hafting et al., 2005). One hypothesis about the function of these cells is that they help stabilize place fields (Bush et al., 2014). This hypothesis is motivated by the fact that place cell activity is driven primarily by sensory inputs (in addition to boundary vector cells), and is thus vulnerable to accumulating corruption by sensory noise. For example, place fields are gradually destabilized when a rat navigates in darkness (Zhang et al., 2014). Grid fields, in contrast, can persist in the darkness (Hafting et al., 2005), possibly because they can make use of proprioceptive velocity information (i.e., sensing self-motion based on signals from motor effectors). Thus, grid cells are well-designed to serve a stabilization function. This is consistent with the observation that entorhinal input is not necessary for the appearance of place fields, but entorhinal lesions disrupt the precision and stability of place fields (Hales et al., 2014). We will now formalize how grid fields could be used for stabilizing the predictive map through a form of *spectral regularization* (Stachenfeld et al., 2017).

Spectral regularization takes its name from the spectrum of a matrix (the set of eigenvalues). When ordered from smallest to largest, the decay of eigenvalues provides information about the intrinsic dimensionality of the matrix—i.e., how many eigenvectors we would need to accurately reconstruct the matrix. If the spectrum has a fast decay, this means its intrinsic dimensionality is low and we only need to retain the eigenvectors with the largest eigenvalues. Applied to the SR (and to SFs), this occurs when the predictive map varies smoothly over the state space, which will generally be true for any environment in which states can be accessed from other states with a small number of steps.

Recall the eigendecomposition of a matrix: $M = \Phi\Lambda\Phi^{-1}$, where each column of Φ is an eigenvector and Λ is a diagonal matrix with eigenvalues on the diagonal.

The "regularization" part of spectral regularization refers to algorithms that bias an estimator towards the low-dimensional reconstruction derived from the eigendecomposition. For example, we could modify the SF update (Eq. 11) to include a bias term:

$$\Delta\hat{\psi}^{\pi}(s) \propto (1-\eta)[f(s) + \gamma\hat{\psi}^{\pi}(s') - \hat{\psi}^{\pi}(s)] + \eta[\tilde{\psi}^{\pi}(s) - \hat{\psi}^{\pi}(s)], \quad (13)$$

where $\tilde{\psi}^{\pi}(s)$ is the low-dimensional reconstruction, and $\eta \in [0,1]$ is a regularization parameter. The update will pull $\hat{\psi}^{\pi}(s)$ more strongly towards $\tilde{\psi}^{\pi}(s)$ when η is close to 1. Stachenfeld et al. (2017) showed that this form of regularization is effective at improving the convergence speed of learning by suppressing noise.

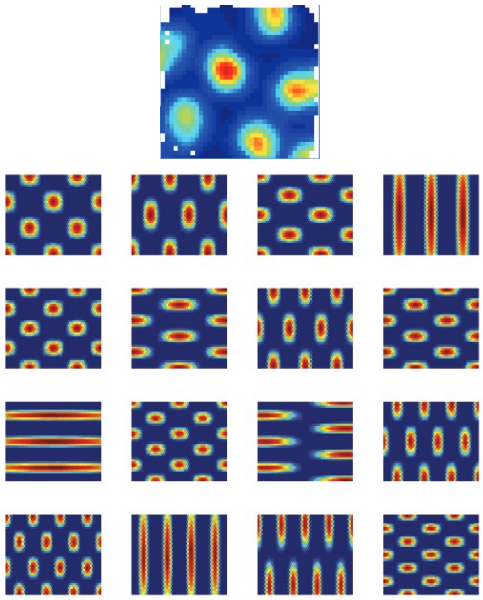

Figure 7: **Real and simulated grid fields in a random foraging paradigm**. (Top) Typical grid field, reproduced from Moser et al. (2015). (Bottom) SR eigenvector grid fields, reproduced from Stachenfeld et al. (2017).

Connecting this back to grid cells, we can show that the eigenvectors of the predictive map are grid-like (Figure 7). If the environment has a toroidal topology (i.e., it is topologically equivalent to a donut) with uniform transition probabilities, then it can be shown that the eigenvectors of the SR are periodic. Obviously real environments are not truly toroidal, but quasi-periodic eigenvectors are still obtained in sufficiently large open spaces. Thus, the eigenvectors capture the essential periodic nature of grid fields. Importantly, they exhibit systematic distortions when barriers are introduced. For example, in a hairpin maze, the eigenvectors (like grid fields) align with the arms of the maze and tend to repeat across alternating arms (Figure 8).

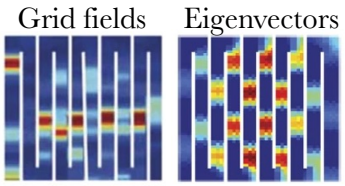

Figure 8: **Fragmented grid fields and eigenvectors**. (Left) Grid fields in a hairpin maze, reproduced from Derdikman et al. (2009). (Right) Eigenvectors of the SR, reproduced from Stachenfeld et al. (2017).

Another feature of grid cells that is captured by this framework is their multi-scale structure: the spatial scale is organized into a small number of discrete modules, arranged topographically along the dorsoventral axis of the entorhinal cortex (Figure 9). This mirrors the topography of the hippocampus dorsoventral axis. Although the entorhinal topography is discrete and the hippocampal topography is a continuous gradient, they may be linked through a continuous gradient in connection strengths. Theoretically, these strengths should match the eigenvalues of the corresponding eigenvector fields. Unlike the eigenvectors (which are invariant to the discount factor γ), the eigenvalue spectrum decays more quickly with higher discount factors, reflecting the fact that less spatial precision is required to reconstruct maps when γ is large. This implies that the neurons with the largest spatial scales in the entorhinal cortex (large eigenvalues)

Stachenfeld et al. (2017) showed that discrete modular organization also arises from the eigenvector decomposition, due to the discrete nature of the eigendecomposition.

should be preferentially connected to the neurons with the largest spatial scales in the hippocampus (large discount factors), consistent with anatomical data (Strange et al., 2014).

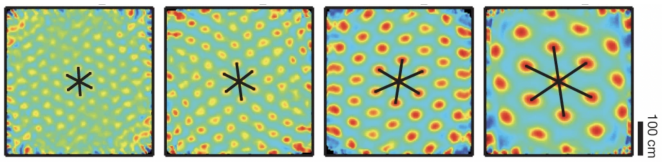

Figure 9: **Grid fields for cells recorded at different positions along the dorsoventral axis of the entorhinal cortex**. Reproduced from Stensola et al. (2012).

How are the eigenvectors computed? One possibility is that they are computed by a hippocampus-to-entorhinal feedback pathway. Such a pathway has been identified (Butola et al., 2025), consistent with the degradation of grid fields following hippocampal inactivation (Bonnevie et al., 2013). In this scenario, the entorhinal cortex might implement one of several neural algorithms for eigendecomposition (Cichocki and Unbehauen, 1992). Alternatively, Stachenfeld et al. (2017) showed how the eigenvectors could be directly computed from state transitions by stochastic gradient descent. This has the computational advantage of only relying on feedforward input, although it may be less compatible with experimental evidence.

9 *Multi-task learning*

We earlier argued that the SR confers flexibility across different reward functions, at least when the transition function is stable. However, there's a snag: the SR is policy-dependent, and the optimal policy typically changes when the reward function changes. Because the SR can't easily adapt in response to policy changes, the resulting value estimates will be suboptimal. The same issue applies to SFs.

Computer scientists have developed techniques to address this problem. To set the stage, we first extend SFs over state-action pairs:

The tuning function $f(s)$ can also be generalized to $f(s,a)$.

$$\psi^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t) | s_0 = s, a_0 = a\pi\right]. \tag{14}$$

This allows us to represent the state-action value function as:

$$Q^{\pi}(s,a) = \sum_d w_d \psi_d^{\pi}(s,a). \tag{15}$$

Now suppose that we have a new reward function parametrized by the vector w'. Barreto et al. (2020) showed that a deterministic policy $\pi'(a|s) = \mathbb{I}[a = a'(s)]$ can be constructed that improves upon (or matches) the policies optimized for earlier tasks, by choosing the action that maximizes the state-action value across all previous tasks:

$$a'(s) = \operatorname*{argmax}_a \max_i Q^{\pi_i}(s,a), \tag{16}$$

The improvement guarantee requires that the new weight vector is within the span of the weight vectors for previous tasks—i.e., there exists a vector α such that $w' = \sum_i \alpha_i w_i$.

where i indexes previous tasks. This algorithm, *generalized policy iteration* (GPI), assumes that the SFs for different policies, $\{\psi^{\pi_i}(s,a)\}$, are stored in memory and reused to solve new tasks. Does the brain do something like this?

First, we can examine the behavioral implications of GPI. Tomov et al. (2021) designed a human RL task in which three different resources could be collected and redeemed for particular prices. The resource levels correspond to the features, $f(s)$, and the prices correspond to the weight vector, w. The critical question was what people would do when tested on a new set of prices (w'). The task was designed such that different patterns of behavior would be predicted under GPI compared to pure model-based and model-free algorithms. GPI favors collecting resources that were encountered frequently in previous tasks, provided that they have relatively large positive weights in the test task. This was also true of people in the experiment, who failed to pursue the optimal policy, which required them to visit a state with relatively less of a resource that was frequently encountered during training.

Next, we can more directly ask whether past predictive maps are stored and reused in the brain. Using brain imaging of humans, Hall-McMaster et al. (2025) showed that neural activity associated with training policies was reactivated during generalization to new tasks. Furthermore, the strength of reactivation predicted the probability of reuse. Surprisingly, the reactivated neural activity in this study was not found in the hippocampus, but rather in visual and prefrontal areas. As noted earlier, there is evidence for predictive maps in visual cortex (Ekman et al., 2023), and action-coding signals can be found there as well (Monaco et al., 2020). There may not be a simple division of labor between feature representations in sensory areas and predictive representations in downstream areas.

10 *Prioritized replay for offline updating*

Another role for memory in learning and generalization arises in the context of "offline" updating, where learning is applied to predictions for previously experienced states and actions. Suppose we have an estimate of the state-action value function, $\hat{Q}^\pi$, which we just updated after a state-action-reward sequence. Because the values are interdependent, we would ideally like the newly updated value estimates to propagate across the state-action space. The simplest approach is to sample random state-action pairs (s,a) and use the Bellman equation to determine the updated value:

$$\hat{Q}^\pi(s,a) \leftarrow R(s) + \gamma \sum_{s'} T(s'|s,a) \max_{a'} \hat{Q}^\pi(s',a'). \tag{17}$$

This is known as a *Bellman backup*. Intuitively, it propagates information backward along the chain of events.

Repeating this update for different state-action pairs (an algorithm known as *value iteration*) is guaranteed to converge to the correct values under the optimal policy π^* (Sutton and Barto, 2018). However, it has two major drawbacks. First, it requires a model (T and R), which might not be available (see next chapter). Second, it might be very inefficient to do this repeatedly for every state-action pair if the space is vast (e.g., chess). We can address the lack of a model by replacing Eq. 17 with a stochastic approximation based on samples $e = (s, a, s', r)$ drawn from memory, leading to a form of TD learning:

$$\Delta\hat{Q}^{\pi}(s,a) \propto r + \gamma \max_{a'} \hat{Q}^{\pi}(s',a') - \hat{Q}^{\pi}(s,a), \tag{18}$$

We will use "backup" to refer generically to both TD updates and Bellman backups. In this case, the TD update is known as *Q-learning*.

where for simplicity we are assuming a tabular approximation. Sampling from memory also has the advantage that it focuses updating on parts of the state-action space that the agent is likely to visit (assuming the past is representative of the future).

The particular way in which memory is sampled can have a big impact on the speed of learning. Mattar and Daw (2018) showed this by deriving the optimal prioritization strategy and comparing it to uniformly sampling memory (Figure 10).

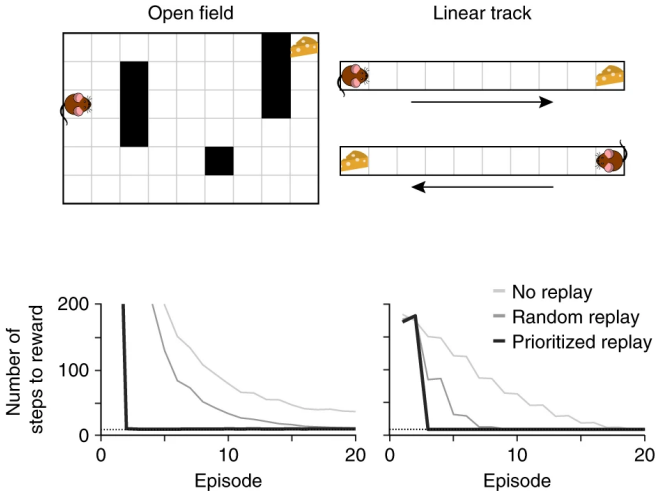

Figure 10: **Learning speed under different memory prioritization strategies for two environments**. Reproduced from Mattar and Daw (2018).

The optimal prioritization strategy is defined by the *expected value of backup* (EVB):

$$\text{EVB}(k) = \hat{V}^{\pi_k}(s) - \hat{V}^{\pi}(s), \tag{19}$$

where s is the agent's current state, k indexes memories $\{e_k\}$, and π_k is the new policy after applying a backup to the retrieved memory. The EVB quantifies the improvement in the policy that could be attained by applying a backup to a particular memory. The optimal prioritization strategy schedules these updates based on decreasing

EVB. Mattar and Daw showed that the EVB can be expressed in the following way:

$$\text{EVB}(s,k) = \text{Gain}(k) \cdot \text{Need}(s,k), \tag{20}$$

where

$$\text{Gain}(k) = V^{\pi_k}(s_k) - V^{\pi}(s_k) \tag{21}$$

is the improvement local to the retrieved state, and

$$\text{Need}(s,k) = M^{\pi}(s, s_k) \tag{22}$$

is how often the memory is "needed" in the sense of how often the retrieved state will be revisited in the future, quantified by the SR. Thus, in this setting the predictive map serves to guide memory retrieval.

Mattar and Daw used this model to explain patterns of place cell reactivation, which tends to happen during periods of quiescence or sleep. The main idea is that place cell reactivation reflects the retrieval of past memories, allowing us to directly examine the predictions of the optimal prioritization strategy. For example, the model explains why place cells tend to be reactivated sequentially based on the sequence of experienced states, but critically makes different predictions about whether the sequence will be in forward or reverse order depending on the situation (Figure 11). If the agent encounters an unexpected reward, the gain term will be greatest for the current state's predecessors, leading to backward replay. This is precisely what is observed on the linear track (Foster and Wilson, 2006). In the absence of a prediction error, the need term dominates, leading to forward replay (states expected in the near future have higher need). This is why forward replay tends to be observed before an animal starts a run on the linear track (Diba and Buzsáki, 2007).

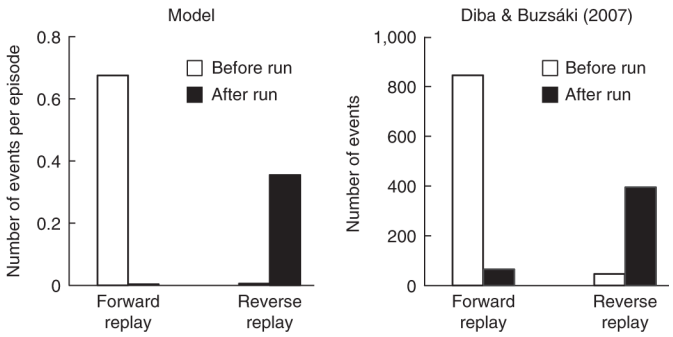

Figure 11: **Frequency of forward vs. reverse replays**. Reproduced from Mattar and Daw (2018).

11 Conclusion

Predictive maps enable the brain to (partially) escape the inflexibility that curses model-free algorithms. We have shown how a canonical RL computation (TD learning of values) can be generalized to construct predictive maps. These maps can also be constructed using a recurrent neural network architecture. Behavioral and neural evidence for predictive maps supports the claim that these representations are used for evaluation, decision making, and multi-task generalization. In the next chapter, we venture towards the pinnacle of flexibility: the use of mental models to plan and simulate.

Study questions

1. In what ways does the SR provide a better account of latent learning than the probabilistic TD learning model?
2. Why might there be multiple predictive maps in the brain (and even within the hippocampus)?
3. How might you incorporate predictive maps into the kind of policy gradient algorithms described in the last chapter?

References

Barreto, A., Hou, S., Borsa, D., Silver, D., and Precup, D. (2020). Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences*, 117:30079–30087.

Bonnevie, T., Dunn, B., Fyhn, M., Hafting, T., Derdikman, D., Kubie, J. L., Roudi, Y., Moser, E. I., and Moser, M.-B. (2013). Grid cells require excitatory drive from the hippocampus. *Nature Neuroscience*, 16:309–317.

Bush, D., Barry, C., and Burgess, N. (2014). What do grid cells contribute to place cell firing? *Trends in Neurosciences*, 37:136–145.

Butola, T., Hernandez-Frausto, M., Blankvoort, S., Flatset, M. S., Peng, L., Hairston, A., Johnson, C. D., Elmaleh, M., Amilcar, A., Hussain, F., et al. (2025). Hippocampus shapes entorhinal cortical output through a direct feedback circuit. *Nature Neuroscience*, 28:811–822.

Cichocki, A. and Unbehauen, R. (1992). Neural networks for computing eigenvalues and eigenvectors. *Biological Cybernetics*, 68:155–164.

Costa, K. M., Shimbo, A., Stalnaker, T., Raheja, N., Mirani, J., Sercander, C., and Schoenbaum, G. (2025). Striatal dopamine signals

errors in prediction across different informational domains. *Science Advances*, 11:eadq9684.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5:613–624.

De Cothi, W. and Barry, C. (2020). Neurobiological successor features for spatial navigation. *Hippocampus*, 30:1347–1355.

Derdikman, D., Whitlock, J. R., Tsao, A., Fyhn, M., Hafting, T., Moser, M.-B., and Moser, E. I. (2009). Fragmentation of grid cell maps in a multicompartment environment. *Nature Neuroscience*, 12:1325–1332.

Diba, K. and Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, 10:1241–1242.

Ekman, M., Kusch, S., and de Lange, F. P. (2023). Successor-like representation guides the prediction of future events in human visual cortex and hippocampus. *elife*, 12:e78904.

Engelhard, B., Finkelstein, J., Cox, J., Fleming, W., Jang, H. J., Ornelas, S., Koay, S. A., Thiberge, S. Y., Daw, N. D., Tank, D. W., et al. (2019). Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature*, 570:509–513.

Fang, C., Aronov, D., Abbott, L., and Mackevicius, E. L. (2023). Neural learning rules for generating flexible predictions and computing the successor representation. *elife*, 12:e80680.

Fanselow, M. S. (1990). Factors governing one-trial contextual conditioning. *Animal Learning & Behavior*, 18:264–270.

Foster, D. J. and Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, 440:680–683.

Gardner, M. P., Schoenbaum, G., and Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B*, 285:20181645.

Garvert, M. M., Dolan, R. J., and Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *elife*, 6:e17086.

Gershman, S. J. (2018). The successor representation: its computational logic and neural substrates. *Journal of Neuroscience*, 38:7193–7200.

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

Hales, J. B., Schlesiger, M. I., Leutgeb, J. K., Squire, L. R., Leutgeb, S., and Clark, R. E. (2014). Medial entorhinal cortex lesions only partially disrupt hippocampal place cells and hippocampus-dependent place memory. *Cell Reports*, 9:893–901.

Hall-McMaster, S., Tomov, M. S., Gershman, S. J., and Schuck, N. W. (2025). Neural evidence that humans reuse strategies to solve new tasks. *PLoS Biology*, 23:e3003174.

Hartley, T., Lever, C., Cacucci, F., and O'Keefe, J. (2000). Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus*, 10:369–379.

Hoffeld, D. R., Thompson, R. F., and Brogden, W. (1958). Effect of stimuli time relations during preconditioning training upon the magnitude of sensory preconditioning. *Journal of Experimental Psychology*, 56:437–442.

Hollup, S. A., Molden, S., Donnett, J. G., Moser, M.-B., and Moser, E. I. (2001). Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *Journal of Neuroscience*, 21:1635–1644.

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

Kjelstrup, K. B., Solstad, T., Brun, V. H., Hafting, T., Leutgeb, S., Witter, M. P., Moser, E. I., and Moser, M.-B. (2008). Finite scale of spatial representation in the hippocampus. *Science*, 321:140–143.

Krishnan, S., Heer, C., Cherian, C., and Sheffield, M. E. (2022). Reward expectation extinction restructures and degrades CA1 spatial maps through loss of a dopaminergic reward proximity signal. *Nature Communications*, 13:6662.

Lamsa, K. P., Heeroma, J. H., Somogyi, P., Rusakov, D. A., and Kullmann, D. M. (2007). Anti-Hebbian long-term potentiation in the hippocampal feedback inhibitory circuit. *Science*, 315:1262–1266.

Lavin, M. J. (1976). The establishment of flavor-flavor associations using a sensory preconditioning training procedure. *Learning and Motivation*, 7:173–183.

Lever, C., Burton, S., Jeewajee, A., O'Keefe, J., and Burgess, N. (2009). Boundary vector cells in the subiculum of the hippocampal formation. *Journal of Neuroscience*, 29:9771–9777.

Mamad, O., Stumpp, L., McNamara, H. M., Ramakrishnan, C., Deisseroth, K., Reilly, R. B., and Tsanov, M. (2017). Place field assembly distribution encodes preferred locations. *PLoS Biology*, 15:e2002365.

Mattar, M. G. and Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21:1609–1617.

Mehta, M. R., Quirk, M. C., and Wilson, M. A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, 25:707–715.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., and Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, 1:680–692.

Monaco, S., Malfatti, G., Culham, J. C., Cattaneo, L., and Turella, L. (2020). Decoding motor imagery and action planning in the early visual cortex: overlapping but distinct neural mechanisms. *NeuroImage*, 218:116981.

Moser, M.-B., Rowland, D. C., and Moser, E. I. (2015). Place cells, grid cells, and memory. *Cold Spring Harbor perspectives in biology*, 7:a021808.

Muller, R., Kubie, J., and Ranck Jr, J. (1987). Spatial firing patterns of hippocampal complex-spike cells in a fixed environment. *Journal of Neuroscience*, 7:1935–1950.

O'Keefe, J. and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, 381:425–428.

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford University Press.

Port, R. L., Beggs, A. L., and Patterson, M. M. (1987). Hippocampal substrate of sensory associations. *Physiology & Behavior*, 39:643–647.

Rescorla, R. (1980). Simultaneous and successive associations in sensory preconditioning. *Journal of Experimental Psychology. Animal Behavior Processes*, 6:207–216.

Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., and Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, 13:e1005768.

Samsonov, S., Tiapkin, D., Naumov, A., and Moulines, E. (2024). Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4511–4547. PMLR.

Sharpe, M. J., Chang, C. Y., Liu, M. A., Batchelor, H. M., Mueller, L. E., Jones, J. L., Niv, Y., and Schoenbaum, G. (2017). Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nature Neuroscience*, 20:735–742.

Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B., and Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science*, 322:1865–1868.

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

Stalnaker, T. A., Howard, J. D., Takahashi, Y. K., Gershman, S. J., Kahnt, T., and Schoenbaum, G. (2019). Dopamine neuron ensembles signal the content of sensory prediction errors. *Elife*, 8:e49315.

Stensola, H., Stensola, T., Solstad, T., Frøland, K., Moser, M.-B., and Moser, E. I. (2012). The entorhinal grid map is discretized. *Nature*, 492:72–78.

Strange, B. A., Witter, M. P., Lein, E. S., and Moser, E. I. (2014). Functional organization of the hippocampal longitudinal axis. *Nature Reviews Neuroscience*, 15:655–669.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Sutton, R. S. and Pinette, B. (1985). The learning of world models by connectionist networks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 7.

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., and Schoenbaum, G. (2017). Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron*, 95:1395–1405.

Talk, A. C., Gandhi, C. C., and Matzel, L. D. (2002). Hippocampal function during behaviorally silent associative learning: Dissociation of memory storage and expression. *Hippocampus*, 12:648–656.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55:189–208.

Tomov, M. S., Schulz, E., and Gershman, S. J. (2021). Multi-task reinforcement learning in humans. *Nature Human Behaviour*, 5:764–773.

Tsetsenis, T., Broussard, J. I., and Dani, J. A. (2023). Dopaminergic regulation of hippocampal plasticity, learning, and memory. *Frontiers in Behavioral Neuroscience*, 16:1092420.

Wang, F., Schoenbaum, G., and Kahnt, T. (2020). Interactions between human orbitofrontal cortex and hippocampus support model-based inference. *PLoS Biology*, 18:e3000578.

Ward-Robinson, J., Coutureau, E., Good, M., Honey, R., Killcross, A., and Oswald, C. (2001). Excitotoxic lesions of the hippocampus leave sensory preconditioning intact: implications for models of hippocampal function. *Behavioral Neuroscience*, 115:1357–1362.

Wiltgen, B. J., Sanders, M. J., Anagnostaras, S. G., Sage, J. R., and Fanselow, M. S. (2006). Context fear learning in the absence of the hippocampus. *Journal of Neuroscience*, 26:5484–5491.

Zhang, S., Schönfeld, F., Wiskott, L., and Manahan-Vaughan, D. (2014). Spatial representations of place cells in darkness are supported by path integration and border information. *Frontiers in Behavioral Neuroscience*, 8:222.