nature neuroscience

REVIEW ARTICLE

https://doi.org/10.1038/s41593-022-01153-y

How to build a cognitive map

James C. R. Whittington[1,2,5] ✉, David McCaffary[2,5], Jacob J. W. Bakermans[2] and Timothy E. J. Behrens[2,3,4]

Learning and interpreting the structure of the environment is an innate feature of biological systems, and is integral to guiding flexible behaviors for evolutionary viability. The concept of a cognitive map has emerged as one of the leading metaphors for these capacities, and unraveling the learning and neural representation of such a map has become a central focus of neuroscience. In recent years, many models have been developed to explain cellular responses in the hippocampus and other brain areas. Because it can be difficult to see how these models differ, how they relate and what each model can contribute, this Review aims to organize these models into a clear ontology. This ontology reveals parallels between existing empirical results, and implies new approaches to understand hippocampal–cortical interactions and beyond.

Since the 1950s, the hippocampal formation has been implicated in functions from episodic memory to spatial and abstract cognition[1–5]. Neuroscientists have attempted to characterize, and provide normative explanations for, the neural representations that support these functions. This has been particularly fruitful in the spatial domain, where several cell types, including hippocampal place cells and entorhinal grid cells, provide a neural instantiation of Tolman's (and Turner's) cognitive map[3,4,6,7] (Fig. 1a).

Cognitive maps were proposed as internal neural representations that enable flexible behavior, such as planning routes or taking novel shortcuts[6–8]. More recent descriptions formalized the fundamental role of cognitive maps as organizing knowledge for generalization[2,3,9] to enable the rapid inference from sparse observations that characterizes biological intelligence[10]. This relates to psychologists' schemas[11]; mental frameworks for understanding new information, and learning sets[12]; and learning common task rules, which enables faster learning in new tasks. These broad concepts encompass domains from social to logical cognition[6], but most neural evidence for cognitive maps is from studies of space[3,13].

Recent evidence, however, suggests parallels between spatial and nonspatial cognition[9] (Fig. 1b). For instance, hippocampal place cells fire not only to locations in space but also to 'location(s)' in sound frequency[14], value[15] or sensory evidence space[16]. Similarly, a putative functional magnetic resonance imaging marker for the hexagonal firing patterns of entorhinal grid cells, developed for physical space[17], is also found when animals are presented with stimuli that vary along two abstract dimensions (for example, neck and leg length of birds[18], odors[19], social hierarchies[20], reward probability and value[21]). These parallels in representation suggest that the mechanism for constructing spatial cognitive maps is an instance of a general coding mechanism that is capable of building abstract cognitive maps covering any domain.

Understanding how the brain represents these different domains of cognition in the same way requires a formalism that connects physical and abstract space. Several hippocampal models have attempted to do this, but it is unclear how these models differ, what each model contributes and how they relate. In this Review, we organize these models into a clear ontology. This ontology reveals that to understand representations in the hippocampal formation, we need to understand how to model sequences—both individual sequences and the statistical structure of sequences. Many of the reviewed models are stated in the reinforcement learning (RL) framework. However, rather than learning from reinforcement, they learn from sensory (or state) predictions to make a good state representation; the RL (or graph) framework just provides convenient mathematical forms. The models learn to turn a sequence of observations (with no rewards) into a useful representation for when rewards do come along later. This is exactly what O'Keefe and Nadel[3] (and Tolman before them[6]) were proposing, and is one kind of latent learning. The ontology also reveals a common understanding behind many existing cellular representations, and suggests new ways to understand hippocampal–cortical interactions. We end by discussing how these models may help to understand neural representations of higher-order cognitive domains, such as language, logical operators and mathematics, thereby providing a pathway toward cognitive maps as Tolman envisaged: the basis of reasoning across all domains of cognition.

A recent review discussed similar ideas of learning structures from sequences[22]. Both ref. [22] and this Review say that the cortex and/or the hippocampus learns (potentially arbitrary topologies) from sequences, with continuous attractor networks involved in learning. We note some theoretical accounts of cognitive maps do not address representation learning[23–25]. Although these models provide mechanistic insights, we do not discuss them in detail.

The cognitive mapping problem

Cognitive maps organize knowledge to enable flexible behavior[3,6,9,26]. Enabling behavior means cognitive maps must contain information relevant to behavioral tasks. Enabling flexibility means the map must enable new behaviors in the face of new challenges, and be built as fast as possible for any new world. The aim for cognitive maps, then, is to learn as much as possible ahead of time, so that online learning and computations are minimized. To achieve this, neural representations of cognitive maps must satisfy certain requirements. Here we describe these computational considerations and the resulting models that have had many recent successes in predicting neuronal representations.

Reinforcement learning and planning. To enable successful behavior, cognitive maps must represent state (that is, a

[1]Department of Applied Physics, Stanford University, Stanford, CA, USA. [2]Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK. [3]Wellcome Centre for Human Neuroimaging, University College London, London, UK. [4]Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London, London, UK. [5]These authors contributed equally: James C.R. Whittington, David McCaffary. ✉e-mail: jcrwhittington@gmail.com

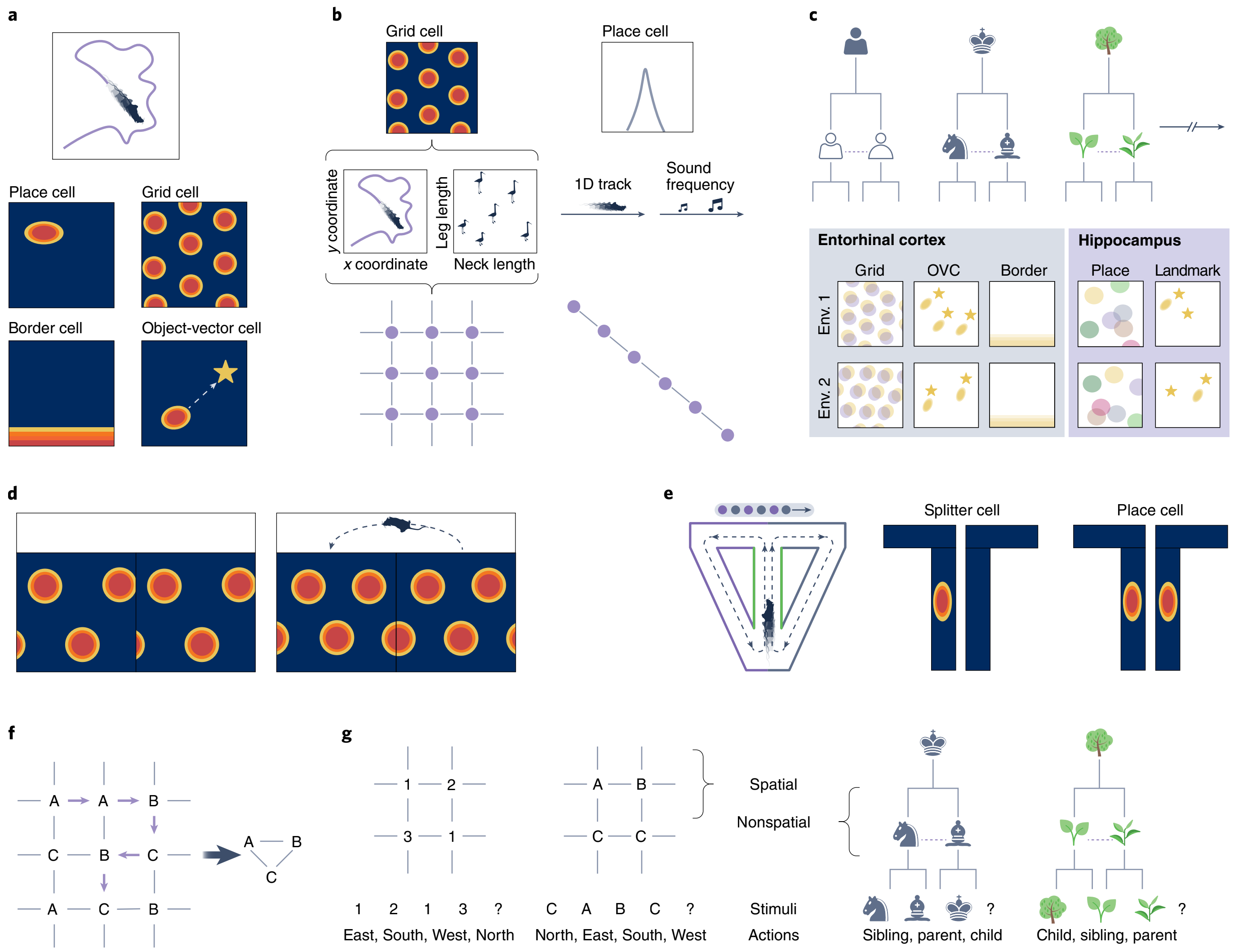


Fig. 1 | The cognitive mapping problem: generalization and latent states. a, When navigating naturalistic environments, a range of cell representations are found in the cognitive map of the hippocampal–entorhinal system. **b**, Recent evidence has implicated these same representations (top) in coding abstract or conceptual spaces (middle; for example, 'bird space'[18] or sound frequencies[14]), with subsequent theoretical accounts suggesting that a single coding mechanism underlies both physical and conceptual spaces (bottom)[9,61,127]: understanding how states of the world (for example, locations in physical/bird space) relate to each other (for example, using graphs) of the world. **c**, Top, understanding the common abstraction among relational structures (for example, families and chess pieces) allows understanding of other structures (plant kingdom) in the same light. The abstraction requires the sensory particularities to be generalized over. Bottom, cell representations of the entorhinal cortex map relational spaces and generalize across environments (Envs. 1 and 2) more than hippocampal cells. **d**, Because the sensory world is aliased (the same observation can happen in different locations), representations must be latent (not a simple function of the current observation). Rodents exhibit latent state representations when they traverse two sensorially identical rooms[43]. Initially, an identical grid cell code represents both rooms, but as the animal realizes these two rooms are connected by a corridor, a global grid cell code predominates; the latent representations separate states with differing sensory futures. **e**, In a T-maze alternation task[40], where rodents take alternating left and right turns, 'splitter cell' representations form (in addition to spatial place cells), which fire preferentially on left or right trials. These are nonspatial latent state representations, because they disambiguate the same spatial location (central trunk) depending on whether it is predicting 'go left' or 'go right'. **f**, The aliasing problem in graphs: if states are represented just by observations, then the left graph is equivalent (shown via black arrow) to the right graph, thus the state-space of the left graph cannot be fully represented by observations alone. **g**, Sequence prediction tasks are sufficient to learn latent state representations, because identical observations can have different neighbors. Sensory sequences, and the associated actions, can come from both space and non-space (for example, families). Some sensory predictions can be done only by knowing (generalizing) certain rules; for example, `North`+`East`+`South`+`West`=0 or `Parent`+`Sibling`+`Niece`=0.

configuration of the world). Deciding when to turn while driving requires knowing how the road curves, where the steering wheel is, where other cars are, and what the road signs say. RL[27] formalizes this concept: actions are taken based on the current world state (for example, turning right when the road bends right). Representing the entire world state is generally infeasible, as it can contain information along countless, and often task-irrelevant, dimensions. Not only is this problematic for representational capacity, it also impedes learning efficiency (switching cars shouldn't require relearning how to drive). This 'curse of dimensionality'[28] can be mitigated by appropriate state abstractions (for example, ignoring car colors). Learning, or attending to, an appropriate abstraction is a central issue of cognitive mapping[23,29,30].

Box 1: Reinforcement learning state-spaces, graphs and graph representations

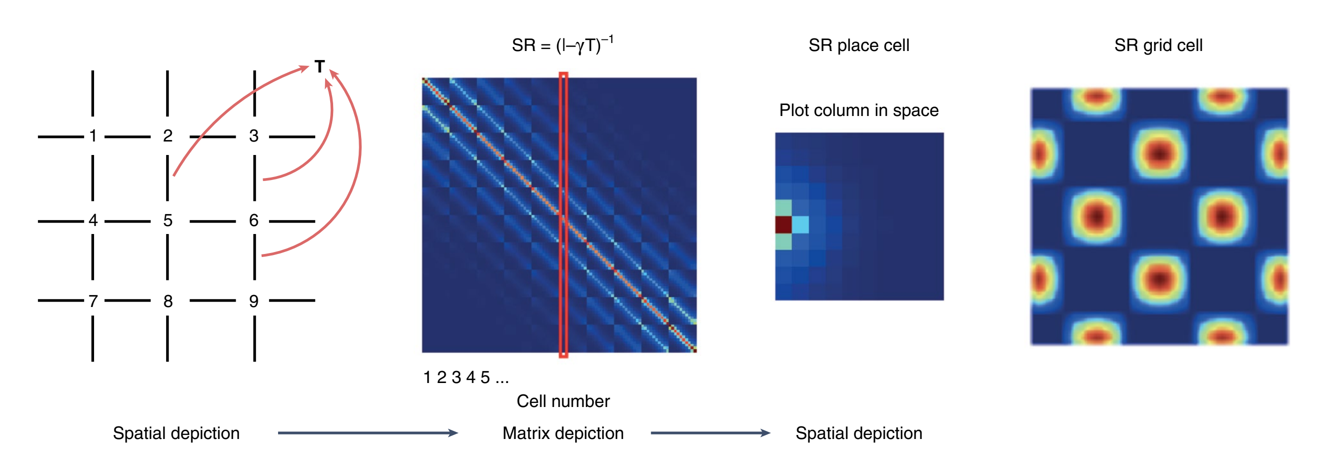
The problem of building graphs for cognitive maps is the same problem as building state-spaces in RL. Crucially, the state-space in RL is tightly linked to behavior (through rewards, values and policies). However, once the state-space is defined there is a further choice of how each state is actually represented. Clever choice of representation can reduce online value/policy computations. This has allowed normative mathematical theories to predict neural representations.

RL is concerned with taking appropriate actions at specific states (s) to maximize the expected (discounted by γ) sum of future rewards $v(s) = \mathbb{E}\left[r(s) + \gamma r(s\prime) + \gamma^2 r(s\prime\prime)\cdots\right]$, where s' and s'' are states following s. Bellman[28] realized that this is a recursive equation, as the right-hand side contains the left-hand side but one step in the future: $v(s) = r(s) + \gamma \sum_t P(s\prime \mid s, \pi)\, v(s\prime)$, where $P(s_{t+1} \mid s_t, \pi)$ is the transition probabilities between states under a policy π. In essence, Bellman's equation says the value of the current state is the reward at that state plus the average value of states you can transition to. If you can assign credit to each state (like these equations do), then taking good actions is easy: just go to the neighboring state with the highest value $v(s')$.

RL state-spaces define graphs with transition matrix elements $T_{ij} = P\left(s_j \mid s_i, \pi\right)$. One graph representation, the sucessor representation (SR)[129] (see the figure), is particularly relevant to cognitive maps[36,37]. The SR is a (discounted) sum of n-step transition matrices; $\mathbf{S} = \sum_n \gamma^n \mathbf{T}^n$. Elements of this matrix, S_{ij}, describe connectedness via all possible paths between two locations. Critically, if we represent connections between states in the world in terms of the SR distance, then computing the value is easy, as the SR is one-half of the value computation[129] ($\mathbf{v} = \mathbf{Sr}$ where $\mathbf{v}$ and $\mathbf{r}$ are vectors whose elements are values and reward at each state, respectively).

Stachenfeld and colleagues[37] noticed that the columns of **S** look like hippocampal place cells (see the figure, center), and that some eigenvectors of **S** resemble entorhinal grid cells (see the figure, right; cell thresholded at zero), similarly to work demonstrating that some eigenvectors of place cell covariance matrices resemble grid cells[79]. Notably, SR makes many predictions about how both grid and place cells behave in different environments and tasks[37,130–132]. Critically, it also makes predictions of representations in nonspatial tasks[37,133,134]. Because it derives from a theory of learning, it can also account for behavioral phenomena that are otherwise hard to explain[94].

One prominent issue with SR, however, is its policy dependence[135]. This means that when rewards move, or, worse, when obstacles appear, value calculations using SR are no longer optimal[135]. A recent model addresses this problem[38] using linear RL[136]. This model builds a default representation (DR) for default behaviors that can be linearly updated when rewards change to approximate the new optimal policy. The required DR resembles the SR, and can therefore be computed from grids cells. The model further provides a new account of how to build world representations compositionally out of component cells representations (for example, how grid and border cells interact to represent the insertion of a barrier)[62]. We return to this important issue in Box 4 and related text.



Classic (model-free) RL learns the value of states, or which actions are good in which states, and therefore requires no knowledge of how states relate to each other. Although this is optimal in the long term[31], value-based learning is often inflexible and slow to learn[27]. Knowing relationships between states (state-space structure) enables flexible planning between any start and goal state, for example, taking a new route home if the regular route is blocked[32]. Unfortunately, traditional planning mechanisms (for example, tree search) are computationally costly, although a clever representation of state-space (see below) can reduce the cost of planning, sometimes completely. This is a powerful way to formalize a central goal of cognitive maps: solving problems in representation, not by exhaustive computation.

Space as a state-space. To understand what this means, let's consider physical space. Here, the state-space comprises physical locations like a literal map. This abstraction alone clearly profoundly helps the spatial planning problem. However, location can be represented in various ways; for example, by a unique identifier (A, B, C,…), or by x and y coordinates. The choice of representation has major consequences. First, consider finding the shortest routes. In the former, you must search through a series of neighbors. In the latter, you can simply compute a vector from start to end representations. Second, consider adding a new location. In the former, a new identifier (for example, a cell) is required along with new relationships (for example, synapses) to neighboring identifiers. In the latter, nothing new is required because (x, y) extends to new locations. These two representation types are analogous to place and grid cells; individual place cells encode unique locations, and new locations therefore require new place cells, whereas grid cells enable vector calculations[33,34] and naturally extend to new locations (albeit

Box 2: Building latent state representations from sequences

State-spaces must be inferred from observations. Because the sensory world is aliased—the same observation can occur more than once—states cannot be inferred from sensory appearance alone. Instead, sequences of observations uniquely identify states because two states with the same sensory observation will have different futures. States inferred via sequences are known as latent states, and building a latent state-space map can be used to enable different behaviors in sensorially identical situations.

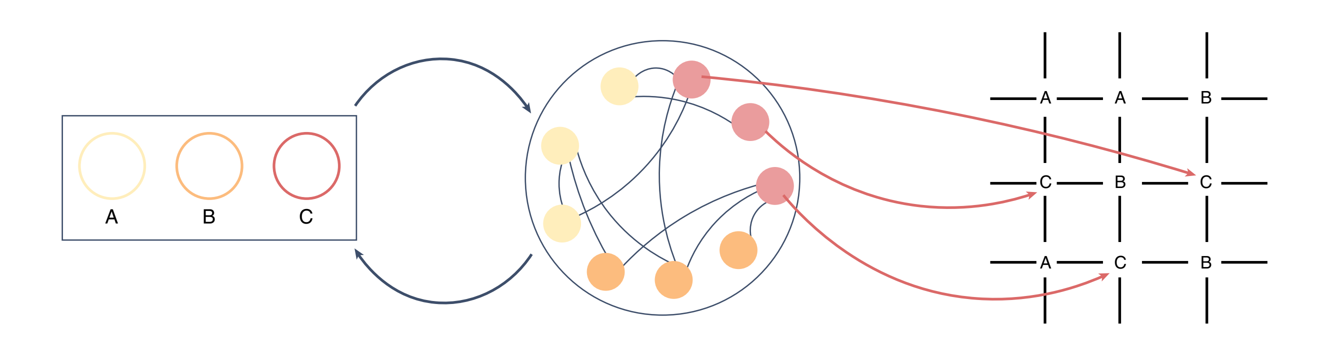
The clone structured cognitive graph (CSCG) model[74] is an elegant approach for building de-aliased state-spaces. Here, the hippocampus contains multiple 'clone' cells for each sensory observation[74,137]. Now, one hippocampal 'frog' clone cell responds to a frog in one location, and another responds if a frog appears elsewhere (see the figure). The model uses Bayes to (1) infer which hippocampal clone cells should be active for each sensory observation and (2) learn an appropriate set of transition weights between clone cells. These transition weights are analogous to the transition matrix for graphs, but critically the state-space is learned, rather than provided by the modeler.

Many hippocampal findings can be understood in terms of representing latent states, from basic phenomena, such as place cells, through to complex representations which vary as a function of animal behavior. These predictions are in common between the CSCG model and more complicated models that follow, and we show a number of these in detail in Fig. 2. A critical difference between CSCG and the following models is that CSCG infers the entire latent space within the hippocampus (as opposed to the cortical input to the hippocampus). This enables learning rules to be local, biologically plausible and fast. However, the CSCG model has to learn each map de novo and cannot benefit from having learnt similar maps before. It is exciting to think how these benefits may be combined ('Complementary maps in hippocampus and cortex'; Fig. 3b).

CSCG is easily expressed in mathematics, and is closely related to hidden Markov models. From a sequence of sensory observations $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \cdots, \mathbf{x}_T\}$ and actions $\mathbb{A} = \{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \cdots, \mathbf{a}_T\}$, we infer discrete latent states $\mathbb{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \cdots, \mathbf{z}_T\}$. Now, the same sensory observation, $\mathbf{x}$, can be linked to different latent states (clones) $\mathbf{z}$, via an 'emission' distribution $p(\mathbf{x}|\mathbf{z})$, naturally accounting for the aliasing problem. Along with predicting sensory observations, CSCG latent states predict future latent states and actions $p(\mathbf{z}_t, \mathbf{a}_t \mid \mathbf{z}_{t-1})$. Modeling the full sequence of observation is then:

$$p(\mathbb{X}, \mathbb{Z}, \mathbb{A}) = p(\mathbf{z}_0) \prod_t p(\mathbf{x}_t \mid \mathbf{z}_t)\, p(\mathbf{z}_t, \mathbf{a}_t \mid \mathbf{z}_{t-1})$$

Here, each element of $\mathbf{z}$, z_i, is a 'clone' of a sensory observation (see the figure); note that we use t for vectors in time and i for indexing elements of each vector. Concretely, if there are 4 possible sensory observations, and 5 clones for each observation, there will be 20 elements to $\mathbf{z}$. The probability of observing a 'frog' given a 'frog clone' is defined as 1, but 0 given a 'snail clone'; $p(\mathbf{x} \mid z_i \in C(\mathbf{x})) = 1$ whereas $p(\mathbf{x} \mid z_i \notin C(\mathbf{x})) = 0$ if $C(\mathbf{x})$ are the clones of $\mathbf{x}$. CSCG marginalizes over $\mathbf{z}$ and uses the expectation–maximization algorithm to train the model[138], that is, learn an appropriate set of transition probabilities $p(\mathbf{z}_t, \mathbf{a}_t \mid \mathbf{z}_{t-1})$ and infer $\mathbf{z}_t$. Once trained, this model can be used for planning by inferring a sequence of actions and observations conditioned on a start and end clone.



periodically). By a clever choice of representation, grid cells abolish the need for computation.

Nonspatial state-spaces. Although it is easy to intuit good state-spaces in physical space, it is harder in non-space. One approach, derived from RL[35,36], is to cast spatial learning as understanding relationships on a graph (Fig. 1b). In space, nodes of a graph define physical locations, and so edges between nodes exist if two locations are directly connected. Importantly, graphs also formalize nonspatial problems (Fig. 1b,c). Family trees, social networks and atoms in molecules, all consist of relationships between entities and can be represented with graphs. Nodes in the graph thus represent nonspatial locations, for example, Alice is Bob's grandparent in a family tree.

Graphs define state-spaces and so enable value-based RL (Box 1). They also enable planning: starting with Bob (characterized by a vector $\mathbf{v}$ with all elements set as zeros except for the Bob node element whose value is 1; each person/state/node is defined by another vector element), and multiplying $\mathbf{v}$ by $\mathbf{T}$ ($\mathbf{Tv}$; $\mathbf{T}$ is the transition matrix, where T_{ij} is the transition probability from state j to i), gives a distribution over future states (people) after one step. Similarly, multiplying again by $\mathbf{T}$ ($\mathbf{T}^2\mathbf{v}$) gives the distribution after two steps. Repeating this process until a nonzero entry appears in Alice's node provides the shortest path between Bob and Alice (two, because Alice is Bob's grandparent).

Several graph-based models of the hippocampal formation have been proposed[37,38], and, intriguingly, their representations of space resemble place and grid cells (Box 1).

Latent states and sequence learning. Graphs can flexibly represent problems, but how do we know which graphs to build? What defines each graph node, or each state in an RL problem? Sensory observations cannot define states, as two identical observations can exist in different locations (aliasing) with very different consequences;

Box 3: Path-integrating state-spaces

Path integration offers a powerful way to build latent state-spaces. It builds maps that embed knowledge of the structure of the space (in physical space, North+East+South+West=0; see the figure, **a**–**c**). This means that path-integration maps are: (1) inherently latent (and abstract), since they follow rules, not sensory observations and (2) allow relational knowledge to be transferred to any situations where the same rules apply. Notably, although path integration is not limited to space, not all graphs can use path integration.

Path-integrating models utilize a particular type of recurrent neural network (RNN) known as continuous attractor neural networks (CANNs[139]; see the figure, **b**), where neurons are recurrently connected via weights, **W**, and receive velocity input, **a**. The neural dynamics are given by:

$$\tau \frac{d\mathbf{g}}{dt} = -\mathbf{g} + f(\mathbf{W}\mathbf{g} + \mathbf{B}\mathbf{a})$$

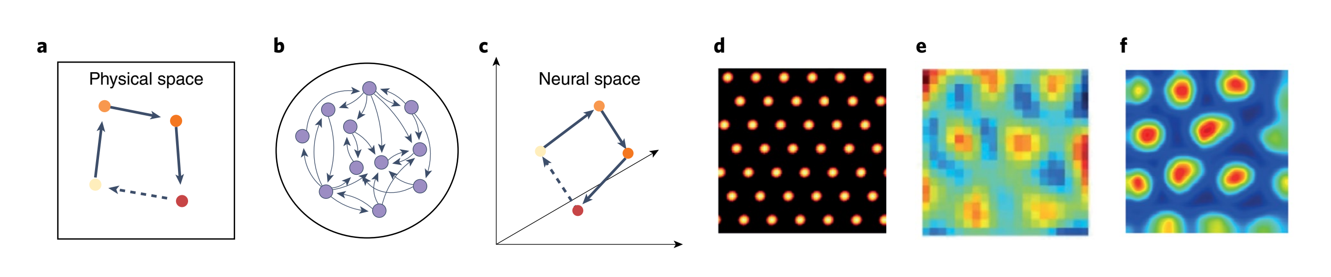
Here, τ is the time constant of neuronal response, f is a nonlinear activation function, **g** is a vector of cells to be path-integrated, and **B** is a matrix projecting velocity inputs, **a**, to cells, **g**. We note an alternative, but less biologically plausible, equation is $\tau\frac{d\mathbf{g}}{dt} = -\mathbf{g} + f(\mathbf{W}_{\mathbf{a}}\mathbf{g})$, where the recurrent matrix $\mathbf{W}_{\mathbf{a}}$ depends on the movement velocity. With an appropriate set of weights, CANNs use path integration, with different cell classes (head direction cells[139,140], place cells[141,142] and grid cells[143]; see the figure, **d**) modeled with different weights. Remarkably, CANNs really exist in nature; ring attractors[144], both in connections and anatomy, are found in flies[145], and attractor manifolds are found in rodents[59,146].

Other path-integrating models exist[84,85]. For example, velocity-coupled oscillators (VCOs) suggest path integration (along an axis) via interference between theta oscillations and velocity-dependent dendritic oscillations, with their phase difference indicating path-integrated distance along an axis (this looks like a plane wave!). Here, grid cells are the sum of three such neurons with preferred axes at $\frac{\pi}{3}$ relative angles.

One major limitation of CANNs and VCOs, however, is that the weights of the recurrent weight matrix, **W**, are carefully selected and not learned from sensory experience. However, it is easy enough to set up path integration as a learning problem via predicting observations **x**: use path integration of the latent state variable **z** and then predict observations **x** from the latent states:

$$p(\mathbb{X}, \mathbb{Z} \mid \mathbb{A}) = p(\mathbf{z}_0) \prod_t p(\mathbf{x}_t \mid \mathbf{z}_t)\, p(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_t)$$

Where the path-integrating part $(p(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{a}_t))$ is now replaced by a discrete-time version, that is, $\mathbf{z}_t = f(\mathbf{W}\mathbf{z}_{t-1} + \mathbf{B}\mathbf{a}) + \text{noise}$. In fact, several models use a deterministic RNN (that is, set the noise term to 0). These models successfully learn to use path integration when tasked with predicting ground truth spatial representations, that is, **x** is either place cells[78], or x and y coordinates[147]. Neural units in both models form periodic representations (see the figure, **e** and **f**), but these are often amorphous, fourfold symmetric grids. An elegant analytic result[148], however, demonstrated that the four- to six-fold symmetry transition is governed by a single property: a third-order regularization term of grid cells. Indeed, this is easily implemented by the biological constraint of ensuring neural activity is positive[79,148]. Reproduced from ref. [143] under a Creative Commons license CC BY 4.0 (**d**). Reproduced with permission from ref. [78], Springer Nature Limited (**e**). Reproduced with permission from ref. [148] (**f**).



crossing a road implies looking right in the United Kingdom, but left in Germany. Formally, our world is not 'fully observable'; instead, we face 'partially observable' problems and must infer latent state[26,29] representations that disambiguate UK and German roads. Although single observations are not enough to infer latent state representations (Fig. 1f), sequences of observations are, because identical observations do not have identical surroundings (for example, having earlier eaten Bratwurst might trigger you to look left when crossing a road).

Indeed, the hippocampal formation learns from sequences and its neural representations disambiguate states using latent representations[16,39–45]. For example, rodent grid cells initially code two identical boxes identically. However, when the animal realizes the boxes are connected by a corridor, the grid representation changes to represent the global two-box-and-corridor space[43] (Fig. 1d). This latent state representation disambiguates sensory aliased boxes due to their different futures. Physical location can also be aliased; in spatial alternation tasks[40,41] (Fig. 1e), the same physical position (for example, the central 'trunk') predicts different futures depending on the animal's previous left/right choice. Splitter cells[40,41], place cells[42], grid cells[43] and lap cells[45] are all examples of the cognitive map disambiguating the world into latent states.

When graphs are given the capacity to learn and infer latent states from sequences, they begin to predict many of the latent state cells described above, for example, via the clone-structured cognitive graph (CSCG) model (Box 2).

Path integration and compression. Inferring latent states is really about understanding where you are in an abstract space. In the two-room task[43], the global grid code uniquely identifies physical locations. For spatial alternation tasks, 'splitter' cells identify location in physical space and location in left/right trials. Working out where you are in physical space is easy; accumulate self-movement vectors (for example, North, South, East and West from head direction cells[46]) to update your location: this is path integration[47] (Box 3 figure). Ants, rodents, birds and humans all use path integration[48–50],

Box 4: Generalizing with memories

We have seen models that build latent state representations, and models that use path integration. If these principles could be combined, we could build a powerful system that learns arbitrary latent states from sensory observations (like CSCG[74]) but additionally generalizes these representations (like path-integration models[78,143]) and composes them arbitrarily. For abstract representations to be reused (generalized) in different sensory environments, the same abstract locations must be 'linked' to different sensory observations. Hippocampal memories offer the ideal substrate for this link; they can rapidly tie sensory observations to specific locations.

Hippocampal models of generalization (the Tolman–Eichenbaum machine, TEM[61], and the spatial memory pipeline, SMP[75]) are tasked with predicting, as fast as possible, sensory observations in novel, but structurally similar, environments (for example, multiple different families or 2D worlds; Fig. 1g). Both models consist of two key components: (1) an abstract path-integration module that is reusable across environments, and (2) a relational memory[2] module that, like an address book, links abstract location representations with sensory representations (see the figure, **a**). These links change from world to world, allowing the same abstractions to apply to multiple worlds.

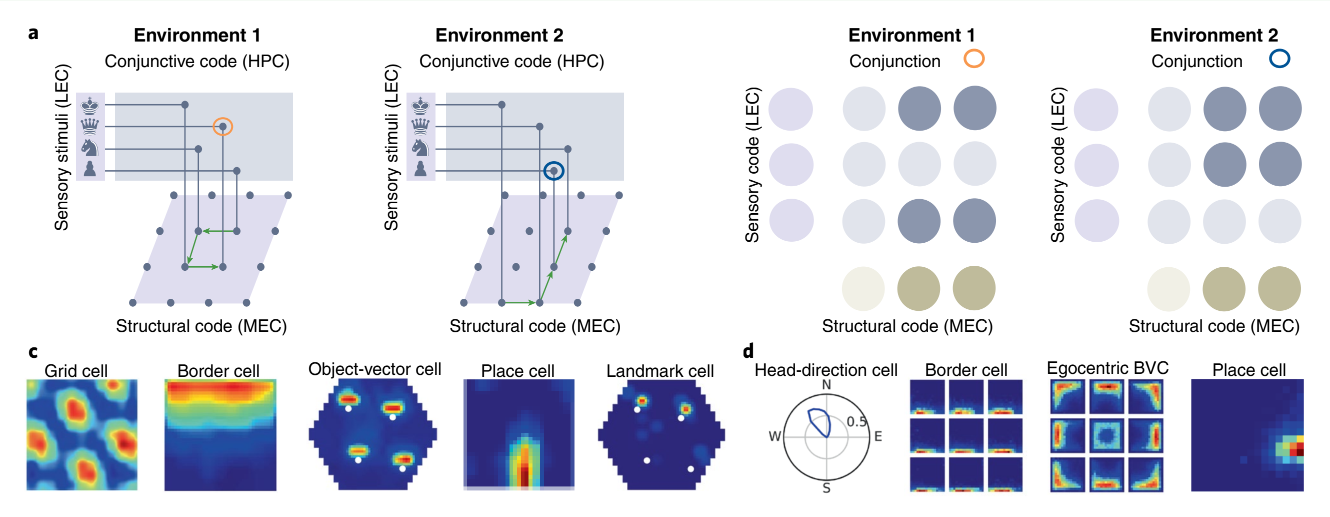
Recall the probabilistic interpretation of path integration:

$$p\,(\mathbb{X}, \mathbb{Z} \mid \mathbb{A}) = p\,(\mathbf{z}_0) \prod_t p\,(\mathbf{x}_t \mid \mathbf{z}_t)\, p\,(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{a}_t)$$

Previously, $p\,(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_t)$ was fixed and each abstract location **z** could only predict a single sensory observation **x**. If, instead, we had an address book of relational memories **M**, we could remember what is where in each environment. To predict upcoming sensory observations, all that is required is to imagine a transition in abstract representation (**z**, via path integration), then retrieve the memory at that location ('what' did I see the last time I was 'here'). Sensory prediction is now a combination of path integration and memory retrieval. But what space are we using path integration in, and how does it get built? Previously, the weights, **W**, in the path integrator $(p\,(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{a}_t)$, where $\mathbf{z}_t = f(\mathbf{W}\mathbf{z}_{t-1} + \mathbf{B}\mathbf{a}) + \text{noise})$ were built from predicting *x* and *y* coordinates or place cells (that is, spatially curated representations). Now, we can predict actual sensory observations. This is more powerful. When sensory objects are arranged in space, the same spatial path-integration mechanism as previous models will be learned, but when the sensory world has more complex dependencies, these will also be learnt. If the best way to predict the sensory future is to learn a complex map of latent states, then these models will learn to use path integration in this latent space (Fig. 2).

Although TEM and SMP are conceptually the same model, they have different implementations. Two critical ones are that (1) TEM is supplied with allocentric actions and object representations, but SMP must infer them from egocentric input and pixels, and (2) SMP implements memory with a memory network from machine learning[149], whereas TEM uses more biologically realistic Hebbian learning[150] and Hopfield networks[151]. This biological constraint means that the link between the abstract and sensory worlds must take place in neuronal units; that is, the same hippocampal neurons must know both the abstract location and the sensory prediction. This type of conjunctive representation is commonly observed real in hippocampal neurons[42,152]. In TEM, this conjunction enables generalization via hippocampal remapping[55–57], as the same cortical representations (LEC and MEC) are reused in different environments, facilitated by different hippocampal combinations (see the figure, **b**).

TEM and SMP are deep artificial neural networks that learn to generalize structural knowledge and recapitulate a host of known representations of the hippocampal cognitive map in doing so (see the figure, **c** and **d**; TEM/SMP). Since SMP works from egocentric inputs, it generates cells involved in the egocentric to allocentric coordinate transformation[66]. Additionally, TEM learns compositional entorhinal representations in spatial and nonspatial tasks, and can solve classical relational memory tasks that are crucially dependent on the hippocampal formation, such as transitive inference[39,153]. Reproduced from ref. [61] under a Creative Commons license CC BY 4.0 (**c**). Reproduced with permission from ref. [75] (**d**).



and in mammals this relies on the hippocampal formation[51]. Entorhinal grid cells are an attractive substrate for two-dimensional (2D) path integration because they extend to all locations, are error correcting[52], require far fewer cells to represent location than place cells[53] and are experimentally driven more from path-integration signals than place cells[54]. Indeed, when neural network models are trained to perform path integration, they learn grid cell representations as their substrate of 2D navigation (Box 3).

Path integrating in graphs and non-space requires a modification. Rather than accumulating self-movement vectors, accumulate abstract-movement vectors instead (`Parent`, `Child`, `Sibling`, `Aunt`, `Nephew`, and so on, for family tree graphs). Just like x and y coordinates versus unique identifying representations for space, with graph representations that use path integration (versus representing every connection separately), adding a new node (Chloe is Bob's `Sibling`) implies other connections (Chloe is Alice's `Grandchild`) without needing to be told them. This is because path integration treats all nodes equally and exploits relational structure, for example, `Sibling`+`Grandparent`=`Grandparent`; take the 'sibling' then the 'grandparent' action, also known as your sibling's grandparent, which is also your grandparent. With path integration you only need to know a few rules rather than every possible relationship; path integration is a compressed representation.

However, not all graphs can use path integration, as it requires the same actions, with the same consequences, to be possible at every location. Consistent actions do not always exist across graphs. For example, on a random graph, relating the connection between nodes 1 and 10 to the connection between nodes 10 and 19 likely makes no sense because there is no common meaning of 'this action' + 'that action' = 'some other action'. The best you can do is 'take the link from 1 to 10' then 'take the link from 10 to 19'. These links have no deeper meaning than the nodes they connect. This is unlike physical space, or families, where `East`, or `Parent`, always has a meaning.

Generalization. Generalization, or the transfer of knowledge between situations, underlies behavioral flexibility. Without it, new situations cannot be understood in the context of existing knowledge and previously learned behaviors cannot be leveraged. Sensory generalization lets you understand that a Pekingese is a type of dog, whereas structural generalization enables deep and powerful inferences: doors often lead to new rooms; addition works for 100 s as it does for 10 s; the same path-integration rules apply in different rooms. These pieces of knowledge have profound effects on behavior.

However, generalizing with graphs is difficult as they require perfect alignment across situations. Perfect alignment is a non-deterministic polynomial-time (NP)-hard problem, which means that the problem is extremely computationally intensive and essentially impractical. By contrast, generalizing with path-integration representations is easy because all positions are treated equally; representations corresponding to the bottom-right in one environment could equally represent the middle of another environment. Furthermore, as path-integration maps are latent (and thus abstract), they chart the relational structure of one family just as well as for another: generalization of relational knowledge.

The hippocampal formation is critical for generalization, as well as for memory, and some forms of imagination[1,2,5]. However, hippocampal representations do not generalize; neighboring place fields are not necessarily neighbors in other environments (remapping[55–57]; Fig. 1c). By contrast, entorhinal representations do generalize; neighboring grid cells (within-module) are also neighbors in other environments even though the overall map can be shifted and/or rotated (realignment[58,59]). Spatial generalization, at least, exists in the entorhinal cortex and is consistent with path integration.

Learning to generalize is often a sequence-learning problem, but with sequences from many different environments (Fig. 1g). When encountering a new family, after observing that Daniel is Emily's parent, and Fran is Daniel's sibling, it is only possible to predict Fran's niece (Emily) if you already know (and generalize) relational knowledge: `Parent`+`Sibling`+`Niece`=0. This is a sequence because actions (for example, `Parent`) are added in order and transitions return you to the starting location (Emily) by path integration.

To actually make sensory predictions, you need to know not just abstract knowledge but also how it interacts with real-world representations (an abstract family tree location may interact (correspond) with Emily for one family, and Chris for another; Box 4, figure **a**). One influential proposal is that hippocampal cells reflect this interaction, with abstract knowledge from medial entorhinal cortex (MEC) and sensory knowledge from lateral entorhinal cortex (LEC) combined in the hippocampus[9,60,61]. This bridges the abstract-to-real divide and permits generalization, since the same abstract map (MEC) can be reused across different sensory (LEC) environments and contexts. Two models that generalize this way are discussed in Box 4.

Composition. Generalization does not always mean transferring a whole map to a new environment; often sub-components, or combinations of sub-components, can be generalized. For example, differently shaped rooms can be understood with two components: an underlying 2D space and walls that can be placed anywhere. Should the cognitive map represent such common structural elements across tasks, then these elements can be composed to understand any given task configuration[38,61,62]. To encourage arbitrary composition, different structural elements (bases) should be represented independently (factorized) from one another[9]. Understanding a task then becomes a structural inference problem; finding the appropriate bases to represent the current task[63].

The hippocampal formation's cognitive map contains many basis representations (Fig. 1a,c). Object-vector cells[64] (OVCs), border-vector cells[65–69] (BVCs), reward cells[70] and goal-vector cells[71] (GVCs) are all examples of local basis representations that encode any object/border/goal, irrespective of its location. By contrast, grid cells are examples of global bases, as they describe information equally across all space. The models discussed in Box 4 learn these compositional cell representations to aid generalization.

New interpretations, integrations and predictions

Although the reviewed models account for a variety of cellular data from spatial and nonspatial tasks, they often do so in different ways. Here we integrate these ideas, leading to a deeper understanding of cognitive maps and new accounts of several other neural phenomena.

Nonspatial hippocampal cells are latent state representations for generalization. Many nonspatial hippocampal representations have been observed[16,40,41,45,72]. Although these cells seemingly represent tasks differently, they can be unified as representing latent state-spaces. We have argued that latent state representations separate states with different futures but also enable generalization, because latent maps can be reused. However, to generalize as fast as possible, every level of abstraction must be represented simultaneously; space in spatial tasks, non-space in nonspatial tasks, and both in interacting spatial–nonspatial tasks.

For example, consider spatial alternation tasks[40,41] where animals cycle left → right → left → right ⋯ at a choice point (Fig. 2a). This task can be 'unrolled' into a 'big-loop' state-space where the first half is going left and the second is going right. This is a latent state-space for the task; it de-aliases the common 'trunk' section depending on whether the animal is going left or right. However, this 'big-loop' ignores space; it does not know you're at the same physical location on return to the common trunk. To do this, space must also be represented and generalized. Indeed, hippocampal cells in this task code for both space (place cells) and big-loop (splitter cells)[40].

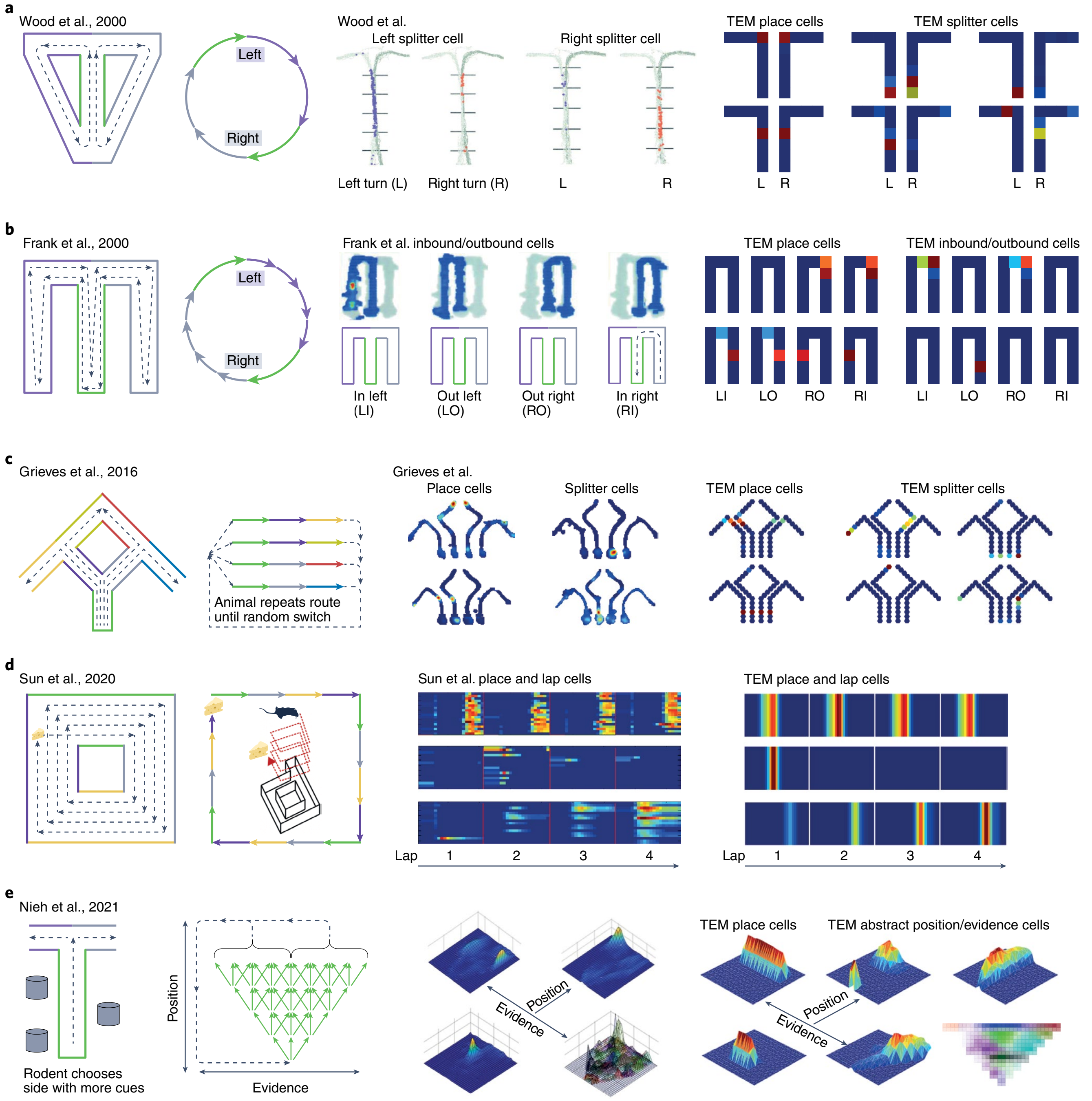


Fig. 2 | Representing latent states. Many apparently different neural phenomena are captured with a unifying computational principle; building state-spaces that can accurately predict different futures (latent states) as fast as possible (generalization). **a**–**e**, For each row, left/center-left are the task and its latent state-space (with colors denoting sensory experience), whereas center-right/right are real/TEM neural representations. **a**, In a T-maze task[40], where animals alternate left/right turns, the state-space is described by a 'big-loop' latent space, since the central trunk predicts different futures depending on a previous left/right turn. Hippocampal cells represent both space (place cells) and the 'big-loop' (splitter cells). Splitter cells are dependent on trajectory, firing at the same spatial location (central trunk) differentially depending on the prospective future (left/right). TEM learns both spatial (place) and nonspatial (splitter) cells when trained on this task; splitter cells to represent latent state in the 'big-loop' and place cells to represent physical location, and thus facilitate spatial generalization. Adapted with permission from ref. [40], Elsevier. **b**,**c**, More complicated spatial alternation tasks[41,44] are also described with 'big-loop' latent state-spaces. Both real and TEM hippocampal representations contain spatial (place) and nonspatial (trajectory-dependent) cell representations. Adapted with permission from ref. [41], Elsevier (**b**). Adapted from ref. [44] under a Creative Commons license CC BY 4.0 (**c**). **d**, Performing four laps to receive a reward is a nonspatial task[45]. It is also described as a 'big-loop' latent state-space. Rodent hippocampus, and TEM, represent both space (place cells; top) and non-space (lap-specific cells; middle/bottom). Adapted with permission from ref. [45], Springer Nature America, Inc. **e**, A T-maze task where rodents choose left/right depending on sensory evidence (as the animal moves along the central trunk) has a latent state-space spanned by position and evidence. Hippocampal cells, and TEM learned hippocampal representations, map this position-evidence latent space that is not just spatial location (bottom-right, a collection of many different cells representations). Adapted with permission from ref. [16], Springer Nature Limited. Code for simulations is available at https://github.com/djcrw/generalising-structural-knowledge. Further learned cell representations are shown in Supplementary Figs. 1 and 2.

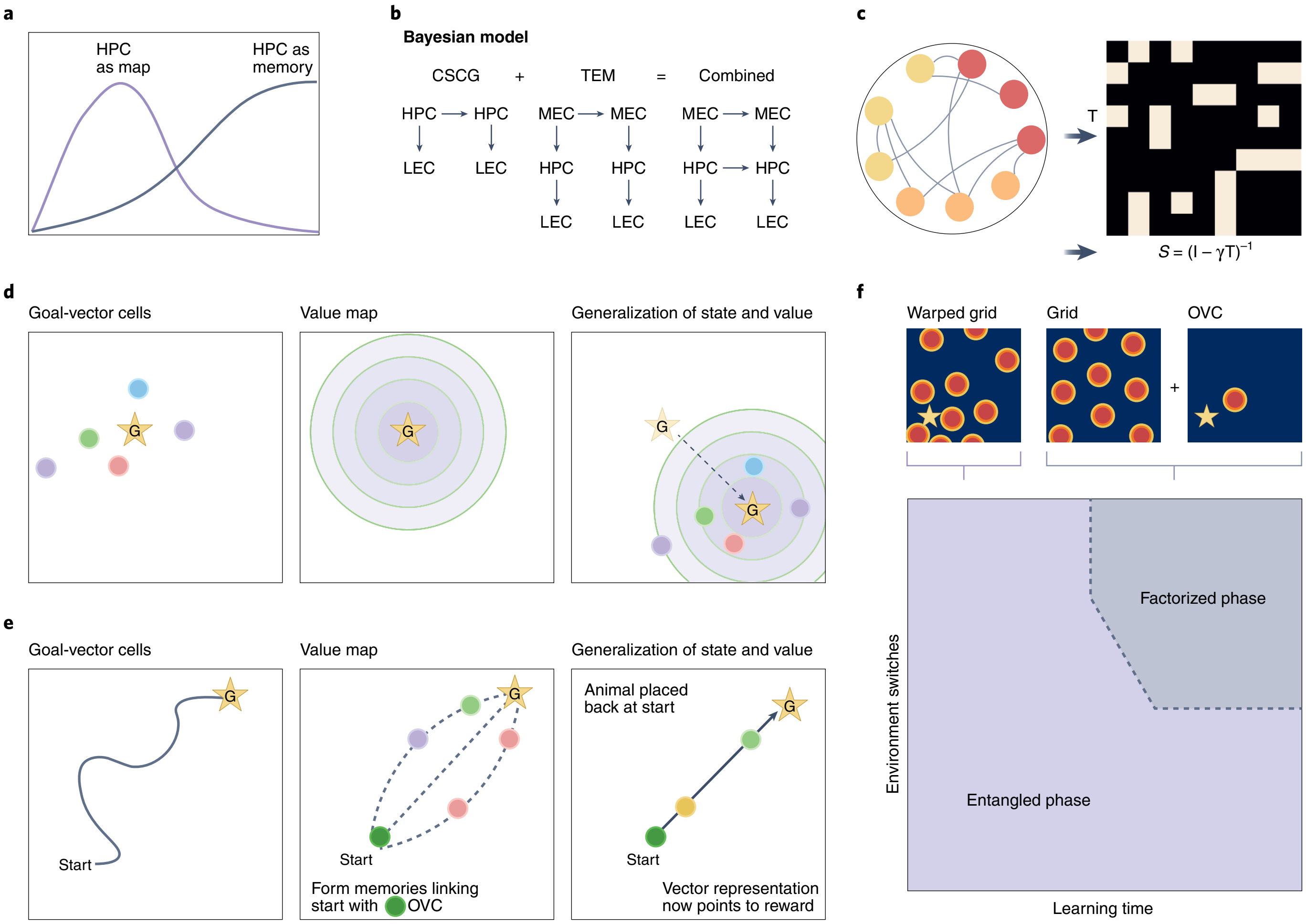


Fig. 3 | Integrating different cognitive map models and new predictions. a,**b**, The reviewed models suggest two roles for the hippocampus (HPC): (1) a map, that is, connections between hippocampal cells encode relationships between states, and (2) memories linking cortical map representations. **a**, We suggest the hippocampus serves both roles, but does so in different situations. In experiences where no previous cortical map is useful, hippocampal representations build a relational map; in familiar experiences where the cortex has learned how to structure experience (for example, by path integration), the hippocampus fulfils the role of memory. We suggest that with increasing experience, there is a transition from the hippocampus as a map to memory, and this will be tied to the behavioral ability to generalize (via cortex). **b**, TEM (HPC as memories) and CSCG (HPC as map) models can be easily integrated (as both are formalized probabilistically) into a model with a hippocampus that can form both maps and memories. **c**–**e**, State-spaces for behavior. **c**, Learned latent state-spaces can be inputted to RL algorithms such as the SR. **d**, On the other hand, compositional representations, such as GVCs, permit rapid generalization of policy. Because these representations already generalize to novel goals in novel environments, all that is required is a pre-computed set of values (or policies) associated with the GVCs. The value map (or policy) is simply transferred along with the GVCs: credit assignment through generalization. **e**, Replay might play a role in this mechanism. After encountering a goal, we want the goal-vector representations to exist across all of space, and especially any start locations. Replay trajectories provide an offline solution; path-integrate (offline) GVCs and bind them (via memory) to important locations such as the start state. Thus, when reentering the same environment, vector representations and the associated value map (or policy) already exist. This is replay as the offline building of maps for credit assignment through generalization. **f**, The aforementioned mechanisms rely on compositional representations, and in particular factorized representations (those that can be linearly de-mixed). Sometimes, however, brain representations are not compositional, but entangled[97]. Since compositional representations are beneficial for generalization, we suggest animals have factorized or entangled representations depending on the pressure to generalize; regularly staying in the same task will encourage entangled representations, whereas regularly switching tasks will encourage factorized representations.

Here we show many nonspatial tasks[16,40,41,45,72] can be understood by these two principles alone (latent states for disambiguation and generalization). We use the TEM, because it learns and generalizes latent states at multiple levels of abstraction. First, training TEM on spatial alternation tasks[40,41,72] (Fig. 2a–c), TEM recapitulates both splitter and place cells—splitter cells for the 'big-loop' and place cells for spatial generalization. Second, when rodents are rewarded every four laps of a loop, hippocampus contains both spatial place cells, and nonspatial cells that care about which lap[45]. TEM learns the same cells: lap cells for the 'big-loop' and place cells for spatial generalization (Fig. 2d). Last, when animals make left/right choices on a T-maze depending on the relative number of left/right sensory cues (Fig. 2e), hippocampal cells form an abstract map spanned by physical space and cue difference ('evidence'). TEM learns exactly this; physical space to predict choice, and cue difference to predict reward left/right.

Complementary maps in hippocampus and cortex. The reviewed models mirror an old debate in cognitive mapping: is the hippocampus building maps[3] or memories[1,73]? In particular, some models (for example, SR and CSCG) propose that the hippocampus stores the map with its neurons representing state/location; other models (for example, TEM and SMP) propose that the entorhinal cortex (or other cortical areas) represent the map, and the hippocampus binds map locations to real-world experiences via memories. These models are not just conceptually different, they are functionally different. 'Hippocampus-as-a-map' models (SR and CSCG[37,38,74]) quickly learn any map but the maps do not generalize, whereas 'hippocampus-as-memory and cortex-as-a-map' models (TEM and SMP;[61,75]) slowly learn the cortical map but, when learned, can immediately generalize it.

We suggest the hippocampus might combine these functional elements. This is profitable as the hippocampus can provide a usable state-space for each environment (successor representation (SR)/CSCG) before the cortex has learned a generalizable map (SMP/TEM; Fig. 3a). Furthermore, the hippocampal maps expedite cortical learning since they can be replayed to the cortex offline, and provide higher-fidelity training signals than observations alone because they are de-aliased. This integrated approach could be formulated as a TEM/SMP model, but one in which the hippocampus is predictive of future hippocampal states (Fig. 3b).

Cognitive maps and behavior. The discussed models relate to behavior in different ways. Models formulated using RL (SR and default representation (DR)[37,38]) provide a state-space for model-free and model-based learning, and suggest the hippocampus constructs a predictive cognitive map for RL[36,37]. Because these models only require well-separated state-space as input, any model that constructs such state-spaces could act as input (for example, CSCG, or TEM, hippocampal cells can input to the SR; Fig. 3c). The sequential models offer an alternative to tree search in 'planning by inference'[76,77]. This involves conditioning on start and goal states, then inferring a distribution over action/state sequences. For example, CSCG is a Bayesian model that naturally implements this procedure, thereby suggesting a hippocampal role in action inference. The models that learn grid codes can use vector-based planning[33,75,78]. This enables navigation and short-cutting behaviors reminiscent of animal behaviors, while also transferring policies from one environment to another because grid cells generalize.

The observation that grid cells resemble eigenvectors of the spatial transition matrix[37] (or of place cells[79]) has led to interesting suggestions about mechanisms for planning and exploration. This is because the eigenvectors for one-step, two-step and multi-step transitions are all the same. They are the same for the SR too. Only the relative weighting (eigenvalues) of eigenvectors change. Intuitively, this means the same eigenvectors can be used for exploration, planning, sampling in replay, or any other type of multi-step navigation. Indeed, with bespoke eigenvalue weightings, very different strategies emerge[80], such as turbulence or super-diffusion (Lévy flights), and can be seen in rodent hippocampal replay[81]. Conveniently, with the SR weighting of eigenvalues, all you need for planning is the start and goal grid codes[82].

So far we have considered diffusive transition matrices (matrices without actions). However, by making transition matrices dependent on action, we can play games like path integration. For instance, recasting individual actions as transition matrices, such that sequentially applying the `North`, `West`, `South` and `East` transition matrices returns you to the starting point. In space, at least, these transition matrices have the same eigenvectors, but different (complex) eigenvalues. Hence, path integration is reduced to successively adding the eigenvalues associated with each action[83]. This unifies path integration with SR-like planning. Interestingly, it also unifies models of path integration since the eigenvectors are plane waves (not grids as the transitions are unidirectional) just like those required for VCOs[84,85], and the transition matrix is just like the weight matrices required for CANNs[86].

Credit assignment through generalization and the interplay with striatal reinforcement learning. Credit assignment is the attribution of value to state. RL typically assumes the underlying state-space is fixed, and values are slowly assigned to these states. However, there is no requirement for state representations to be fixed; they can change to better represent value. For example, after encountering a goal, GVCs form[71] (cells that are active at certain distances and directions from goals; Fig. 3d). This can be interpreted as augmenting the state representation (with new cells). Importantly, since GVCs use path integration, once discovering a goal, all GVCs can be built as the animal navigates. Although GVCs have only been recorded in bats[71], we hypothesize they play a general role across species.

We propose building state-space, s, compositionally from reusable building blocks (for example, OVCs, GVCs, BVCs and grid cells that generalize within and across environments) dramatically helps behavior. Intuitively, this is because the compositional representations can come 'pre-credit assigned', and thus immediately provide a state-space that accurately predicts value (or policy) in novel situations. This is understood with a little formalism. RL says values (or actions) are a function of state—$f(\mathbf{s})$. Thus, to get new behaviors in new situations, you can either change the function, f, or change your state, $\mathbf{s}$. In general, there is no fast or efficient way to change the f (Bellman updates and gradient descent are slow, whereas explicit planning is highly computationally costly). However, the state, $\mathbf{s}$, can be adjusted rapidly, for example, by addition of vector cells (GVCs, BVCs and OVCs), and the function f can stay the same. f just needs to be general and work with many different state combinations; f 'pre-credit assigns' the compositional representations. This is easy to learn; just train over many different combinations of vector cells and goal locations!

For example, in an open-field task with changing goal locations, all that's required is learning a function f([GVCs]), and then appropriately placing the GVC representation in the right place (centered around the goal; Fig. 3d). Things are less intuitive in environments with multiple objects and boundaries, but the idea is the same, only now other local bases may be required, for example f([GVCs, OVCs, BVCs, ...]). The only online role of the cognitive map is inferring which pre-learned and pre-credit assigned representations to compose together. This is credit assignment through generalization, and is akin to meta-RL[87], as previous statistical knowledge (for example, GVCs, BVCs and OVCs) can be integrated on the fly to solve novel tasks.

Where do these representations come from in the first place? The cognitive map models suggest these representations are learned from statistics of behavior. Just as OVCs can be learned when behavior is biased toward objects[61], GVCs can be learned when behavior is biased toward goals. In general, cortex must learn from sequences of behavior. This suggests an interplay of learning between generalization and RL. In entirely novel tasks, when the agent has a naive state-space, behavior is learned via classic RL (perhaps in striatum). Initial actions will be bad, but eventually RL will learn policies toward goals. These behavioral sequences can be replayed to the cortico-hippocampal system, which extracts statistics and learns a better state-space (for example, compositional representations; GVCs, BVCs, OVCs) from these policies. This is a virtuous cycle; learned cortical representations can be provided back to striatum as an RL state-space, which can then learn better policies, and so on. When cortical representations are good, behavior can be generated entirely from generalization, with no need for new striatal RL. This relates to recent machine learning methods in offline RL, where sequence models learn the statistics of behavioral sequences from conventional RL algorithms, after which the sequence model can be used for planning[88,89]. In sum, this proposal offers a new role of cortical–basal ganglia interaction for constructing RL state-spaces and generalizing policies.

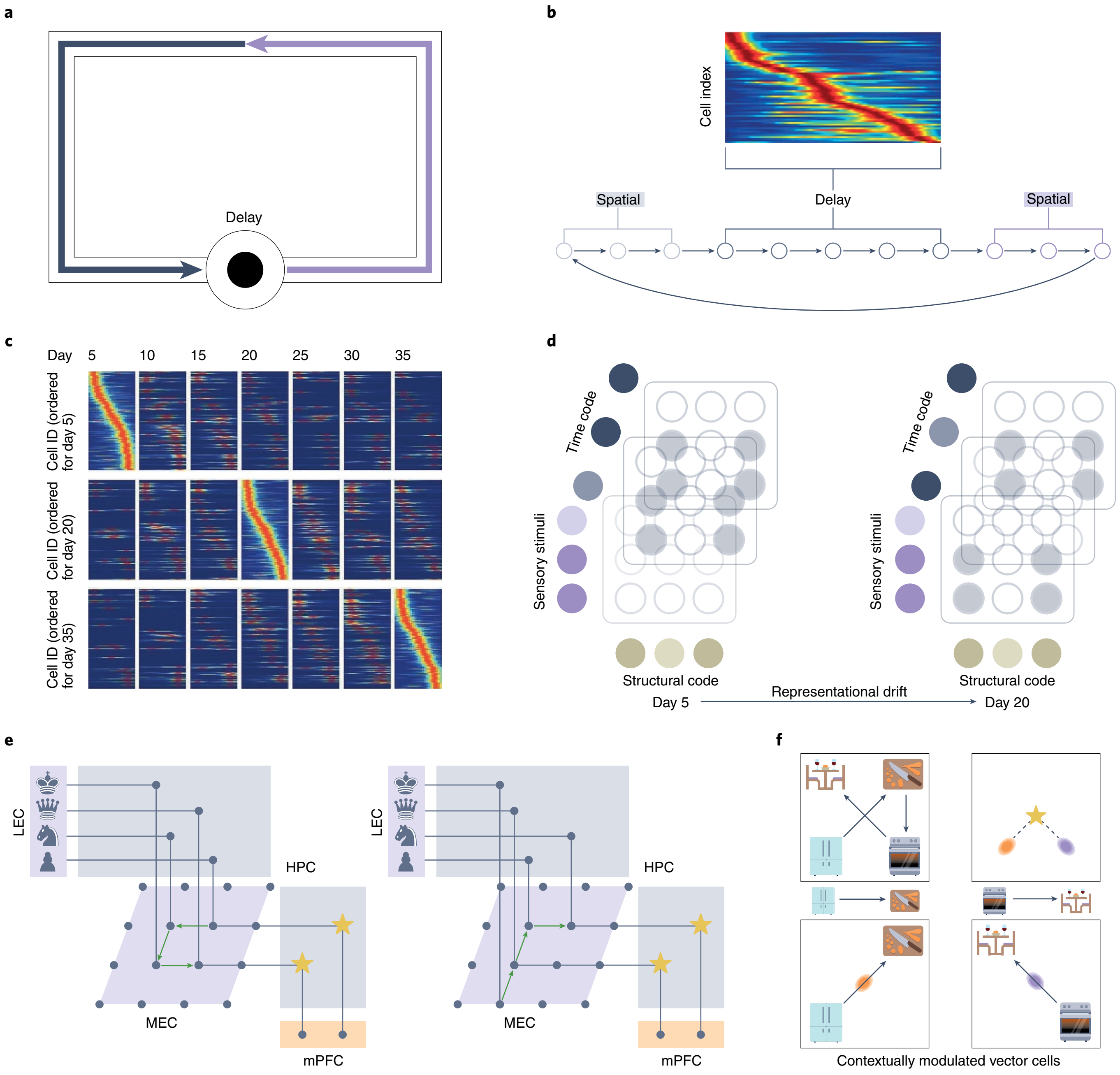


Fig. 4 | Representing time and hierarchies of abstraction in cognitive maps. a,**b**, Neuronal representations of time might structure tasks according to 'progress' through that task. **a**, In a task with a delay period, the full (**b**) latent state-space of the task includes the delay period, as 'progress' through the delay must be represented to predict when the delay period ends: in essence, latent states for predicting the future. Indeed, sequences of hippocampal cells fire during the delay period as if they were coding time[103,104]. Reproduced from ref. [128], Society of Neuroscience. **c**, Time additionally impacts representations via drift[99]. Here, hippocampal (and other) representations slowly change over many days, such that an entirely different representation encodes the same location. Reproduced with permission from ref. [99], Springer Nature America, Inc. **d**, Although the mechanism, or function, of drift is unknown, a tantalizing possibility, inspired by the reviewed models, is that representational drift is remapping in disguise. In particular, as in TEM, if hippocampal representations reflect a triple conjunction of space, sensory stimuli and time, then drifting hippocampal representations can parsimoniously be due to changing time representation while space and sensory representations remain the same. Rather than spatial remapping (Box 4, figure **b**), this is temporal remapping. **e**,**f**, Representing hierarchical tasks. **e**, Schematic of a hierarchical version of TEM, where an additional prefrontal module is included. Should this module represent location in task at an abstracted level (for example, 'just before the oven' in a recipe), this abstract and nonspatial representation can contextualize the hippocampal–entorhinal system, that is, set goals in space. We note that this schematic is a simplification of true neuroanatomy (for example, medial PFC (mPFC) and hippocampus connections may go via the reuniens). **f**, This predicts new cell types, such as route-dependent GVCs: representations that point toward goal locations but only at specific points in the task (for example, only before chopping the vegetables). This is analogous to splitter cells, although these representations can occur anywhere in space, not just at specific points on a T-maze. Icons are from https://www.flaticon.com.

Replay: offline state-space construction. If behavioral control in new worlds is reduced to state-space composition, it becomes important to construct state-spaces rapidly and accurately, and to store them in memory so they can inform future decisions. To build such memories requires path integration (for example, to ensure the correct GVC is tied (composed) to the correct location) but, to be efficient, as much of the compositions as possible should be done offline (that is, not using path integration for the animal's actual location).

An appealing substrate for this composition is replay[90]. For example, when an animal receives a reward, it is important that all other states are aware of their relative location to the reward. Replay can use path integration away from the reward, successively tying (composing) each new GVC to its respective hippocampal/cortical location (perhaps building landmark cells in the hippocampus[91]; this is a similar mechanism to the simultaneous grid and place cell replay from Evans and Burgess[92], but now used to instantiate rewarding policies, instead of ensuring consistency between place and grid representations). Now, should the animal return to a state, the state representation already 'knows' about its relation to reward (Fig. 3e). It is no longer necessary to hold all goal locations in mind, as the state-space composition is stored in memory. All heavy computations of building state-spaces take place offline, thus the computational burden is reduced for online behavior. This relates to ideas from RL that cast replay as a mechanism for optimal credit assignment to existing states[93], or a mechanism for building state-spaces from scratch[27,94]. However, in a generalization framework (outlined in the section above), these two computational processes are subsumed by the single process of composing state-spaces from pre-learnt bases. Notably, this framework makes predictions not only about optimal patterns of hippocampal replay, but also if and when these patterns will align with replay of more abstract representations in entorhinal and frontal cortices[95,96].

When neural representations factorize. Grid cells were once thought of as representations of space and space alone. By apparently ignoring sensory (or other nonspatial) details of an environment, grid cells were considered a factorized representation of space. Similarly, other spatial representations found in the entorhinal cortex, such as OVCs and BVCs, are seemingly factorized, since they compositionally augment the entorhinal grid representation to represent different environmental configurations; however, recent evidence has shown that grid cells warp toward consistently rewarded locations[97]. Factorized representations should not warp, since warping is an environment-specific phenomenon; warping around rewards does not transfer to different spatial configurations of rewards.

We suggest a tension between warping and not warping due to the pressure to generalize versus precisely representing a single task (Fig. 3f). With infrequent task switches (that is, repetitively solving a task), it is more efficient to learn and store a bespoke warped representation (warped since the animal performs stereotyped looping behaviors in space), as generalization is not necessary and storing one representation is more efficient than combining many. With regular task switches (for example, solving different goal configurations of the same task), the pressure for generalization is high, and so compositional bases are favorable. This idea can be stated succinctly: when the set of tasks that an animal faces is itself factorized, then cellular representations for that task will also be factorized (made up of compositional bases; Fig. 3f). This hypothesis makes simple and falsifiable predictions in spatial tasks with environmental rewards: when rewards and space regularly occur in any combination (factorized), both representations of space (grid cells) and reward (reward-vector cells) will exist. By contrast, when rewards and space always occur in the same combination, a bespoke, warped representation will suffice. We note early evidence that these task demands lead to factorized[98] and warped[97] representations.

Open questions

The role of time in memory and cognitive maps. The discussion of cognitive map models so far assumes that learned representations remain stable over time. This clearly cannot be the case, since we can remember events at the same place and same conditions but on different days. Empirical evidence indeed indicates that neural representations drift over time and experience (Fig. 4c), challenging traditional notions of engrams and receptive fields[99–102].

But how can the hippocampus maintain a stable representation of space if the cellular basis of this representation drifts? Generalization models offer a solution: hippocampal cells bind multiple factors together, thus only one factor needs to change for the entire representation to change (Fig. 4d). If entorhinal cortex learns abstracted representations of time as well as space, then, as the temporal code progresses, the hippocampal code will drift to new cells, but these new cells will only differ in their connections to the entorhinal cells that represent time, not those that represent space (Fig. 4d). In this view, representational drift is just hippocampal remapping, but where time has changed, not sensory observations or space. A prediction that follows is that the order of drifting cells is not random.

The hippocampus represents time through more than just drift. For example, pure 'time cells' emerge when rodents are required to stay still, or run on a wheel, for a particular duration (Fig. 4a)[103,104]. These cells can be understood as enabling prediction of when the delay period finishes (Fig. 4b). Crucially, this temporal representation is just one part of an overall map relating experiences to one another. More precisely, during the delay period, space is not changing, but position in task is changing. Cognitive map models suggest it is the overall task position that is being represented in 'time cells'.

Interacting levels of abstraction. We have shown how models can build abstract representations that generalize over different sensory realizations, but the real power of abstraction comes when this process happens repeatedly, so that abstractions can lead to further abstractions. When we are learning to cook a new recipe, we don't need to relearn the rules of space to find the oven, and when the recipe is learnt it can easily be transferred to kitchens with new spatial layouts.

The latent state tasks discussed earlier (Fig. 2), had both task ('go left then right') and space elements, but these came in a fixed configuration; the latent space would not have generalized if the T-maze became a W-maze. Cooking recipes in different kitchens is equivalent to a switch from T-maze to W-maze; we need something new in the models to account for this. One attractive option is for spatial and task representations to be separately represented, or 'factorized', so they can be arbitrarily combined (ovens being in different locations in different kitchens). Given enough experiences of different kitchens, this factorization could emerge from training. However, using the same tricks as before (the hippocampus as a mediator of factorized representations), the required number of recipes and kitchens for training can be dramatically reduced. One possibility is that the different representations observed in frontotemporal cortices[105–110] might reflect such a factorization, with entorhinal representations grounded in interactions with the physical environment, and neurons in the prefrontal cortex (PFC) representing abstract, task-related invariances, such as 'location in task'[30,96,105,110–112].

Although factorization allows representation of any space–task combination, these representations must interact to actually understand any given space–task combination. The go-to-oven medial PFC representation needs to be linked to the spatial location of the oven, or vector cells pointing toward the oven, to actually navigate to the oven. This linking could occur through hippocampal memories (Fig. 4e). Interestingly, the same vector cells can be reused to point toward the oven or the chopping board. This makes a prediction: contextually modulated[107] vector cells depending on 'location in task' (Fig. 4f).

Building models of interacting task and spatial representations, with principles of abstraction, generalization and path integration, allows neural representations from RL tasks to be understood in the same language as space. As emergent task-level (medial PFC) representations potentially reveal insights into how cells might represent task structure itself, they will be of interest whenever animals are shaped to perform tasks.

From sequences to other domains of cognition. The models we described translate the problem of building maps into the problem of understanding the structure of sequences. This raises two questions. First, there are many other sequence problems not traditionally related to cognitive maps; can these can be understood similarly to space and tasks[113,114]? Second, can our understanding from sequence problems extend to problems in other cognitive domains?

Regarding the first question, machine learning has shown that sequence learners (RNNs, long short-term memory units[115], Transformers[116]) can perform well on tasks including language processing, mathematical understanding and logic problems[117,118]. This makes sense as these are sequence problems in which generalization is key: each comprises content (words/numbers) combined within different structures (grammatical rules/mathematical operators) and vice versa. Although mathematics and language engage large (and different) cortical territories[119], the neuronal representations that support these functions might be understood with principles similar to state representation, factorization and path integration described above (indeed, a recent paper showed that hippocampal models of generalization are Transformers[120]). For example, mathematical operators, such as addition and subtraction, bear similarity to forwards and backwards actions on a line (similarly for integration and differentiation).

Regarding the second question, much of the neural processing underlying cognitive problems does not seemingly require sequence transitions. For example, understanding that a football and the Earth are both spheres does not require learning from sequences; thus, it is not clear whether organizing principles similar to space play a role in learning these abstractions. Nevertheless, there are analogies between path integration and understanding spheres: the data-generative factors of a ball—sizes, shapes and colors—are all examples of variables that can be projected onto a manifold where 'actions' such as `add-red`, `bigger`, `remove-red`, then `smaller` have a meaning. Indeed, machine learning methods learn such manifolds from images inputted in no particular sequence[121,122]. Some nonsequential problems can also be reformulated sequentially, for example, although an image is not sequential, it can be viewed sequentially. It is notable that grid-like cells have been observed in monkey[123] and human[124,125] entorhinal cortex during saccades on images, and when humans view silhouettes of stacked objects, component objects are replayed sequentially[126].

Note that we do not claim entorhinal cortex solves and represents all types of structures (for example, family trees). Other brain regions are likely involved in path-integrating structures more abstract than physical space, and we posit their interactions with the hippocampus obey similar computational principles. However, the entorhinal cortex is particularly well placed anatomically to index the hippocampus with a rich distributed representation, so even in these cases, entorhinal representations may relay the structural representations to their hippocampal targets.

Conclusion

The hippocampal formation is a poster child for cognitive neuroscience because of its beautifully organized neuronal responses and the profound effects of its damage. However, although these experimental findings seem self-explanatory when examined in simple situations like open-field foraging, they have been hard to relate to complex real-world behaviors. Moreover, it is not clear whether the discoveries gleaned from navigation studies have broader implications for understanding more general cognitive processes. By reimagining the problem, the ideas and models reviewed here offer concrete and formal methods for addressing this. By asking questions such as 'what really is space to the brain?', they have been able to make connections between how neurons behave in space and how they behave in nonspatial tasks. They provide new computational explanations for how these processes might support behavior, and for the link between space and memory. These contributions have relied on a genuine link between theory and experiment, and this cross-disciplinary collaboration will continue to increase our understanding of how brains make sense of the structure of experience, and use it to construct flexible behaviors.

Data availability

No data were generated in this Review.

Code availability

Python and TensorFlow code are available at https://github.com/djcrw/generalising-structural-knowledge.

Received: 12 August 2021; Accepted: 25 July 2022;
Published online: 26 September 2022

References

1. Scoville, W. B. & Milner, B. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* **20**, 11–21 (1957).
2. Cohen, N. J. & Squire, L. R. Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science* **210**, 207–210 (1980).
3. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, 1978).
4. Hafting, T., Fyhn, M., Molden, S., Moser, M. B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).
5. Hassabis, D., Kumaran, D., Vann, S. D. & Maguire, E. A. Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl Acad. Sci. USA* **104**, 1726–1731 (2007).
6. Tolman, E. C. Cognitive maps in rats and men. *Psychol. Rev.* **55**, 189–208 (1948).
7. Turner, C. H. The homing of ants: an experimental study of ant behavior. *J. Comp. Neurol. Psychol.* **17**, 367–434 (1907).
8. Zanforlin, M. & Poli, G. The burrowing rat: a new technique to study place learning and orientation. *Acti. Memorie* **82**, 653–670 (1970).
9. Behrens, T. E. J. et al. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
10. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: statistics, structure and abstraction. *Science* **331**, 1279–1285 (2011).
11. Bartlett, F. C. & Burt, C. Remembering: a study in experimental and social psychology. *Br. J. Educ. Psychol.* **3**, 187–192 (1932).
12. Harlow, H. F. The formation of learning sets. *Psychological Rev.* **56**, 51–65 (1949).
13. Moser, E. I., Moser, M.-B. & McNaughton, B. L. Spatial representation in the hippocampal formation: a history. *Nat. Neurosci.* **20**, 1448–1464 (2017).
14. Aronov, D., Nevers, R. & Tank, D. W. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature* **543**, 719–722 (2017).
15. Knudsen, E. B. & Wallis, J. D. Hippocampal neurons construct a map of an abstract value space. *Cell* **184**, 4640–4650 (2021).
16. Nieh, E. H. et al. Geometry of abstract learned knowledge in the hippocampus. *Nature* https://doi.org/10.1038/s41586-021-03652-7 (2021).
17. Doeller, C. F., Barry, C. & Burgess, N. Evidence for grid cells in a human memory network. *Nature* **463**, 657–661 (2010).
18. Constantinescu, A. O. et al. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
19. Bao, X. et al. Grid-like neural representations support olfactory navigation of a two-dimensional odor space. *Neuron* **102**, 1066–1075 (2019).
20. Park, S. A., Miller, D. S., Nili, H., Ranganath, C. & Boorman, E. D. Map making: constructing, combining and inferring on abstract cognitive maps. *Neuron* **107**, 1226–1238 (2020).
21. Bongioanni, A. et al. Activation and disruption of a neural mechanism for novel choice in monkeys. *Nature* **591**, 270–274 (2021).

22. Rueckemann, J. W., Sosa, M., Giocomo, L. M. & Buffalo, E. A. The grid code for ordered experience. *Nat. Rev. Neurosci.* **22**, 637–649 (2021).
23. Radulescu, A., Shin, Y. S. & Niv, Y. Human representation learning. *Annu. Rev. Neurosci.* **44**, 253–273 (2021).
24. Sanders, H., Wilson, M. A. & Gershman, S. J. Hippocampal remapping as hidden state inference. *eLife* **9**, e51140 (2020).
25. Stoianov, I., Maisto, D. & Pezzulo, G. The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning. *Prog. Neurobiol.* **217**, 102329 (2022).
26. Niv, Y. Learning task-state representations. *Nat. Neurosci.* **22**, 1544–1553 (2019).
27. Sutton, R. S. & Barto, A. G. Reinforcement learning: an introduction. in *IEEE Transactions on Neural Networks* https://doi.org/10.1109/TNN.1998.712192 (2017).
28. Bellman, R. A Markovian decision process. *J. Math. Mech.* 6, 679–684 (1957).
29. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* **20**, 251–256 (2010).
30. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–278 (2014).
31. Watkins, C. J. C. H. & Dayan, P. Technical note: Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
32. Tolman, E. C., Ritchie, B. F. & Kalish, D. Studies in spatial learning. I. Orientation and the short-cut. *J. Exp. Psychol.* **36**, 13–24 (1946).
33. Bush, D., Barry, C., Manson, D. & Burgess, N. Using grid cells for navigation. *Neuron* **87**, 507–520 (2015).
34. Stemmler, M., Mathis, A. & Herz, A. V. M. Connecting multiple spatial scales to decode the population activity of grid cells. *Sci. Adv.* **1**, e1500816 (2015).
35. Foster, D. J., Morris, R. G. M. & Dayan, P. A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* **10**, 1–16 (2000).
36. Gustafson, N. J. & Daw, N. D. Grid cells, place cells and geodesic generalization for spatial reinforcement learning. *PLoS Comput. Biol.* **7**, e1002235 (2011).
37. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
38. Piray, P. & Daw, N. D. A model for learning based on the joint estimation of stochasticity and volatility. *Nat. Commun.* **12**, 6587 (2021).
39. Dusek, J. A. & Eichenbaum, H. The hippocampus and memory for orderly stimulus relations. *Proc. Natl Acad. Sci. USA* **94**, 7109–7114 (1997).
40. Wood, E. R., Dudchenko, P. A., Robitsek, R. J. & Eichenbaum, H. Hippocampal neurons encode information about different types of memory episodes occurring in the same location. *Neuron* **27**, 623–633 (2000).
41. Frank, L. M., Brown, E. N. & Wilson, M. Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron* **27**, 169–178 (2000).
42. Komorowski, R. W., Manns, J. R. & Eichenbaum, H. Robust conjunctive item-place coding by hippocampal neurons parallels learning what happens where. *J. Neurosci.* **29**, 9918–9929 (2009).
43. Carpenter, F., Manson, D., Jeffery, K., Burgess, N. & Barry, C. Grid cells form a global representation of connected environments. *Curr. Biol.* **25**, 1176–1182 (2015).
44. Grieves, R. M., Wood, E. R. & Dudchenko, P. A. Place cells on a maze encode routes rather than destinations. *eLife* **5**, 1–24 (2016).
45. Sun, C., Yang, W., Martin, J. & Tonegawa, S. Hippocampal neurons represent events as transferable units of experience. *Nat. Neurosci.* **23**, 651–663 (2020).
46. Taube, J., Muller, R. & Ranck, J. Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J. Neurosci.* **10**, 420–435 (1990).
47. Darwin, C. Origin of certain instincts. *Nature* **7**, 417–418 (1873).
48. Mittelstaedt, M. L. & Mittelstaedt, H. Homing by path integration in a mammal. *Naturwissenschaften* **67**, 566–567 (1980).
49. Etienne, A. S. & Jeffery, K. J. Path integration in mammals. *Hippocampus* **14**, 180–192 (2004).
50. Loomis, J. M. et al. Nonvisual navigation by blind and sighted: assessment of path integration ability. *J. Exp. Psychol.* **122**, 73–91 (1993).
51. Maaswinkel, H., Jarrard, L. E. & Whishaw, I. Q. Hippocampectomized rats are impaired in homing by path integration. *Hippocampus* **9**, 553–561 (1999).
52. Sreenivasan, S. & Fiete, I. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nat. Neurosci.* **14**, 1330–1337 (2011).
53. Mathis, A., Herz, A. V. M. & Stemmler, M. Optimal population codes for space: grid cells outperform place cells. *Neural Comput.* **24**, 2280–2317 (2012).
54. Chen, G., Lu, Y., King, J. A., Cacucci, F. & Burgess, N. Differential influences of environment and self-motion on place and grid cell firing. *Nat. Commun.* **10**, 630 (2019).
55. Anderson, M. I. & Jeffery, K. J. Heterogeneous modulation of place cell firing by changes in context. *J. Neurosci.* **23**, 8827–8835 (2003).
56. Bostock, E., Muller, R. U. & Kubie, J. L. Experience-dependent modifications of hippocampal place cell firing. *Hippocampus* **1**, 193–205 (1991).
57. Muller, R. U. & Kubie, J. L. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *J. Neurosci.* **7**, 1951–1968 (1987).
58. Fyhn, M., Hafting, T., Treves, A., Moser, M. B. & Moser, E. I. Hippocampal remapping and grid realignment in entorhinal cortex. *Nature* **446**, 190–194 (2007).
59. Yoon, K. et al. Specific evidence of low-dimensional continuous attractor dynamics in grid cells. *Nat. Neurosci.* **16**, 1077–1084 (2013).
60. Manns, J. R. & Eichenbaum, H. Evolution of declarative memory. *Hippocampus* **16**, 795–808 (2006).
61. Whittington, J. C. R. et al. The Tolman–Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* **183**, 1249–1263 (2020).
62. Mark, S., Moran, R., Parr, T., Kennerley, S. W. & Behrens, T. E. J. Transferring structural knowledge across cognitive maps in humans and models. *Nat. Commun.* **11**, 4783 (2020).
63. Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proc. Natl Acad. Sci. USA* **105**, 10687–10692 (2008).
64. Høydal, Ø. A., Skytøen, E. R., Andersson, S. O., Moser, M. -B. & Moser, E. I. Object-vector coding in the medial entorhinal cortex. *Nature* **568**, 400–404 (2019).
65. Hartley, T., Burgess, N., Lever, C., Cacucci, F. & O'Keefe, J. Modeling place fields in terms of the cortical inputs to the hippocampus. *Hippocampus* **10**, 369–379 (2000).
66. Becker, S. & Burgess, N. Modelling spatial recall, mental imagery and neglect. *Adv. Neural Inf. Process. Syst.* **13**, 96–102 (2001).
67. Barry, C. et al. The boundary vector cell model of place cell firing and spatial memory. *Rev. Neurosci.* **17**, 71–97 (2006).
68. Solstad, T., Boccara, C. N., Kropff, E., Moser, M.-B. & Moser, E. I. Representation of geometric borders in the entorhinal cortex. *Science* **322**, 1865–1868 (2008).
69. Lever, C., Burton, S., Jeewajee, A., O'Keefe, J. & Burgess, N. Boundary vector cells in the subiculum of the hippocampal formation. *J. Neurosci.* **29**, 9771–9777 (2009).
70. Gauthier, J. L. & Tank, D. W. A dedicated population for reward coding in the hippocampus. *Neuron* **99**, 179–193 (2018).
71. Sarel, A., Finkelstein, A., Las, L. & Ulanovsky, N. Vectorial representation of spatial goals in the hippocampus of bats. *Science* **355**, 176–180 (2017).
72. Grieves, R. M. & Jeffery, K. J. The representation of space in the brain. *Behav. Processes* **135**, 113–131 (2017).
73. Eichenbaum, H. Time cells in the hippocampus: a new dimension for mapping memories. *Nat. Rev. Neurosci.* **15**, 732–744 (2014).
74. George, D. et al. Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. *Nat. Commun.* **12**, 2392 (2021).
75. Uria, B. et al. The spatial memory pipeline: a model of egocentric to allocentric understanding in mammalian brains. Preprint at *bioRxiv* https://doi.org/10.1101/2020.11.11.378141 (2020).
76. Botvinick, M. & Toussaint, M. Planning as inference. *Trends Cogn. Sci.* **16**, 485–488 (2012).
77. Friston, K. The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* **13**, 293–301 (2009).
78. Banino, A. et al. Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**, 429–433 (2018).
79. Dordek, Y., Soudry, D., Meir, R. & Derdikman, D. Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *eLife* **5**, 1–36 (2016).
80. McNamee, D. C., Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. Flexible modulation of sequence generation in the entorhinal–hippocampal system. *Nat. Neuro.* https://doi.org/10.1038/s41593-021-00831-7 (2021).
81. Pfeiffer, B. E. & Foster, D. J. Autoassociative dynamics in the generation of sequences of hippocampal place cells. *Science* **349**, 180–183 (2015).
82. Baram, A. B., Muller, T. H., Whittington, J. C. R. & Behrens, T. E. J. Intuitive planning: global navigation through cognitive maps based on grid-like codes. Preprint at bioRxiv https://doi.org/10.1101/421461 (2018).
83. Yu, C., Behrens, T. E. J. & Burgess, N. Prediction and generalisation over directed actions by grid cells. *International Conference on Learning Representations* (2021).
84. O'Keefe, J. & Recce, M. L. Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3**, 317–330 (1993).
85. Burgess, N., Barry, C. & O'Keefe, J. An oscillatory interference model of grid cell firing. *Hippocampus* **17**, 801–812 (2009).
86. Burak, Y. & Fiete, I. Do we understand the emergent dynamics of grid cell activity? *J. Neurosci.* **26**, 9352–9354 (2006).

87. Wang, J. X. et al. Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
88. Chen, L. et al. Decision transformer: reinforcement learning via sequence modeling. Preprint at https://arxiv.org/abs/2106.01345 (2021).
89. Janner, M., Li, Q. & Levine, S. Offline reinforcement learning as one big sequence modeling problem. Preprint at https://arxiv.org/abs/2106.02039 (2021).
90. Foster, D. J. & Wilson, M. A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680–683 (2006).
91. Deshmukh, S. S. & Knierim, J. J. Influence of local objects on hippocampal representations: landmark vectors and memory. *Hippocampus* **23**, 253–267 (2013).
92. Evans, T. & Burgess, N. Coordinated hippocampal-entorhinal replay as structural inference. *Adv. Neural Inf. Process. Syst.* **32**, 1731–1743 (2019).
93. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018).
94. Momennejad, I. et al. The successor representation in human reinforcement learning. *Nat. Hum. Behav.* **1**, 680–692 (2017).
95. Ólafsdóttir, H. F., Carpenter, F. & Barry, C. Coordinated grid and place cell replay during rest. *Nat. Neurosci.* **19**, 792–794 (2016).
96. Kaefer, K., Nardin, M., Blahna, K. & Csicsvari, J. Replay of behavioral sequences in the medial prefrontal cortex during rule switching. *Neuron* **106**, 154–165 (2020).
97. Boccara, C. N., Nardin, M., Stella, F., O'Neill, J. & Csicsvari, J. The entorhinal cognitive map is attracted to goals. *Science* **363**, 1443–1447 (2019).
98. Butler, W. N., Hardcastle, K. & Giocomo, L. M. Remembered reward locations restructure entorhinal spatial maps. *Science* **363**, 1447–1452 (2019).
99. Ziv, Y. et al. Long-term dynamics of CA1 hippocampal place codes. *Nat. Neurosci.* **16**, 264–266 (2013).
100. Driscoll, L. N., Pettit, N. L., Minderer, M., Chettih, S. N. & Harvey, C. D. Dynamic reorganization of neuronal activity patterns in parietal cortex. *Cell* **170**, 986–999 (2017).
101. Rule, M. E., O'Leary, T. & Harvey, C. D. Causes and consequences of representational drift. *Curr. Opin. Neurobiol.* **58**, 141–147 (2019).
102. Rubin, A., Geva, N., Sheintuch, L. & Ziv, Y. Hippocampal ensemble dynamics timestamp events in long-term memory. *eLife* **4**, e12247 (2015).
103. Pastalkova, E., Itskov, V., Amarasingham, A. & Buzsaki, G. Internally generated cell assembly sequences in the rat hippocampus. *Science* **321**, 1322–1327 (2008).
104. MacDonald, C. J., Lepage, K. Q., Eden, U. T. & Eichenbaum, H. Hippocampal 'time cells' bridge the gap in memory for discontiguous events. *Neuron* **71**, 737–749 (2011).
105. Zhou, J. et al. Evolving schema representations in orbitofrontal ensembles during learning. *Nature* **590**, 606–611 (2021).
106. Zhou, J. et al. Complementary task structure representations in hippocampus and orbitofrontal cortex during an odor sequence task. *Curr. Biol.* **29**, 3402–3409 (2019).
107. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).
108. Bernardi, S. et al. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954–967 (2020).
109. Morton, N. W., Schlichting, M. L. & Preston, A. R. Representations of common event structure in medial temporal lobe and frontoparietal cortex support efficient inference. *Proc. Natl Acad. Sci. USA* **117**, 29338–29345 (2020).
110. Samborska, V., Butler, J. L., Walton, M. E., Behrens, T. E. & Akam, T. Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nat. Neurosci.* (in the press).
111. Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human orbitofrontal cortex represents a cognitive map of state space. *Neuron* **91**, 1402–1412 (2016).
112. Yu, J. Y., Liu, D. F., Loback, A., Grossrubatscher, I. & Frank, L. M. Specific hippocampal representations are linked to generalized cortical representations in memory. *Nat. Commun.* **9**, 2209 (2018).
113. Hawkins, J., Lewis, M., Klukas, M., Purdy, S. & Ahmad, S. A framework for intelligence and cortical function based on grid cells in the neocortex. *Front. Neural Circuits* **12**, 121 (2019).
114. Lewis, M. Hippocampal spatial mapping as fast graph learning. Preprint at https://arxiv.org/abs/2107.00567 (2021).
115. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 17351780 (1997).
116. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **20**, 5999–6009 (2017).
117. Brown, T. B. et al. Language models are few-shot learners. Preprint at https://arxiv.org/abs/2005.14165 (2020).
118. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. Preprint at https://arxiv.org/abs/2010.11929 (2020).
119. Amalric, M. & Dehaene, S. Origins of the brain networks for advanced mathematics in expert mathematicians. *Proc. Natl Acad. Sci. USA* **113**, 4909–4917 (2016).
120. Whittington, J. C. R., Warren, J. & Behrens, T. E. J. Relating transformers to models and neural representations of the hippocampal formation. In *International Conference on Learning Representations* (2022).
121. Higgins, I. et al. β-VAE: learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations* (2017).
122. Higgins, I. et al. Towards a definition of disentangled representations. Preprint at https://arxiv.org/abs/1812.02230 (2018).
123. Killian, N. J. & Buffalo, E. A. Grid cells map the visual world. *Nat. Neurosci.* **21**, 161–162 (2018).
124. Nau, M., Navarro Schröder, T., Bellmund, J. L. S. & Doeller, C. F. Hexadirectional coding of visual space in human entorhinal cortex. *Nat. Neurosci.* **21**, 188–190 (2018).
125. Julian, J. B., Keinath, A. T., Frazzetta, G. & Epstein, R. A. Human entorhinal cortex represents visual space using a boundary-anchored grid. *Nat. Neurosci.* **21**, 191–194 (2018).
126. Schwartenbeck, P. et al. Generative replay for compositional visual understanding in the prefrontal-hippocampal circuit. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.06.447249 (2021).
127. Bellmund, J. L. S., Gärdenfors, P., Moser, E. I. & Doeller, C. F. Navigating cognition: spatial codes for human thinking. *Science* **362**, eaat6766 (2018).
128. Salz, D. M. et al. Time cells in hippocampal area CA3. *J. Neurosci.* **36**, 7476–7484 (2016).
129. Dayan, P. Improving generalization for temporal difference learning: the successor representation. *Neural Comput.* **5**, 613–624 (1993).
130. Mehta, M. R., Quirk, M. C. & Wilson, M. A. Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron* **25**, 707–715 (2000).
131. Derdikman, D. et al. Fragmentation of grid cell maps in a multicompartment environment. *Nat. Neurosci.* **12**, 1325–1332 (2009).
132. Krupic, J., Burgess, N. & O'Keefe, J. Neural representations of location composed of spatially periodic bands. *Science* **337**, 853–857 (2012).
133. Garvert, M. M., Dolan, R. J. & Behrens, T. E. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife* **6**, e17086 (2017).
134. Schapiro, A. C., Turk-browne, N. B., Botvinick, M. M., Norman, K. A. & Schapiro, A. C. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, 20160049 (2017).
135. Momennejad, I. Learning structures: predictive representations, replay, and generalization. *Curr. Opin. Behav. Sci.* **32**, 155–166 (2020).
136. Todorov, E. Linearly solvable Markov decision problems. In *Advances in Neural Information Processing Systems* 1369–1376 https://doi.org/10.7551/mitpress/7503.003.0176 (2007).
137. Cormack, G. V. & Horspool, R. N. S. Data compression using dynamic Markov modelling. *Comput. J.* **30**, 541–550 (1987).
138. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **39**, 1–22 (1977).
139. Zhang, K. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. *J. Neurosci.* **16**, 2112–2126 (1996).
140. Skaggs, W. E., Knierim, J. J., Kudrimoti, H. S. & McNaughton, B. L. A model of the neural basis of the rat's sense of direction. *Adv. neural Inf. Process. Syst.* **7**, 173–180 (1995).
141. Samsonovich, A. & McNaughton, B. L. Path integration and cognitive mapping in a continuous attractor neural network model. *J. Neurosci.* **17**, 5900–5920 (1997).
142. Tsodyks, M. Attractor neural network models of spatial maps in hippocampus. *Hippocampus* **9**, 481–489 (1999).
143. Burak, Y. & Fiete, I. R. Accurate path integration in continuous attractor network models of grid cells. *PLoS Comput. Biol.* **5**, e1000291 (2009).
144. Ben-Yishai, R., Bar-Or, R. L. & Sompolinsky, H. Theory of orientation tuning in visual cortex. *Proc. Natl Acad. Sci. USA* **92**, 3844–3848 (1995).
145. Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. Ring attractor dynamics in the *Drosophila* central brain. *Science* **356**, 849–853 (2017).
146. Gardner, R. J. et al. Toroidal topology of population activity in grid cells. *Nature* **602**, 123–128 (2022).
147. Cueva, C. J. & Wei, X. -X. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. Peeprint at https://arxiv.org/abs/1803.07770 (2018).

148. Sorscher, B., Mel, G. C., Ganguli, S. & Ocko, S. A. A unified theory for the origin of grid cells through the lens of pattern formation. *Adv. Neural Inf. Process. Syst.* **32**, 10003–10013 (2019).
149. Pritzel, A. et al. Neural episodic control. Preprint at https://arxiv.org/abs/1703.01988 (2017).
150. Hebb, D. O. *The Organization of Behavior; a Neuropsychological Theory* (Wiley, 1949).
151. Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities (associative memory/parallel processing/categorization/content-addressable memory/fail-soft devices). *Biophysics* **79**, 2554–2558 (1982).
152. McKenzie, S. et al. Hippocampal representation of related and opposing memories develop within distinct, hierarchically organized neural schemas. *Neuron* **83**, 202–215 (2014).
153. Bunsey, M. & Eichenbaum, H. Conservation of hippocampal memory function in rats and humans. *Nature* **379**, 255–257 (1996).

Acknowledgements

We thank N. Burgess and C. Sun for helpful comments on earlier drafts of the manuscript. We thank the following funding sources: Sir Henry Wellcome Post-doctoral Fellowship (222817/Z/21/Z) to J.C.R.W.; Wellcome Trust DPhil Scholarship to D.M.; and Wellcome Principal Research Fellowship (219525/Z/19/Z), Wellcome Collaborator award (214314/Z/18/Z), and JS McDonnell Foundation award (JSMF220020372) to T.E.J.B.. The Wellcome Centre for Integrative Neuroimaging and Wellcome Centre for Human Neuroimaging are each supported by core funding from the Wellcome Trust (203139/Z/16/Z, 203147/Z/16/Z).

Author contributions

J.C.R.W. and T.E.J.B. conceptualized the manuscript. J.C.R.W. and D.M. performed simulations. J.C.R.W. drafted the manuscript with input from D.M. J.C.R.W. and T.E.J.B. edited the manuscript with input from D.M. and J.J.W.B.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-022-01153-y.

Correspondence should be addressed to James C. R. Whittington.

Peer review information *Nature Neuroscience* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.