Internally generated sequences in learning and executing goal-directed behavior

Giovanni Pezzulo[1], Matthijs A.A. van der Meer[2], Carien S. Lansink[3,4], and Cyriel M.A. Pennartz[3,4]

[1] Institute of Cognitive Sciences and Technologies, National Research Council, Via San Martino della Battaglia 44, 00185 Roma, Italy
[2] Department of Biology and Centre for Theoretical Neuroscience, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada
[3] Swammerdam Institute for Life Sciences – Center for Neuroscience, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[4] Amsterdam Brain and Cognition, Research Priority Program Brain and Cognition, Nieuwe Achtergracht 129, 1018 WS Amsterdam, The Netherlands

A network of brain structures including hippocampus (HC), prefrontal cortex, and striatum controls goal-directed behavior and decision making. However, the neural mechanisms underlying these functions are unknown. Here, we review the role of 'internally generated sequences': structured, multi-neuron firing patterns in the network that are not confined to signaling the current state or location of an agent, but are generated on the basis of internal brain dynamics. Neurophysiological studies suggest that such sequences fulfill functions in memory consolidation, augmentation of representations, internal simulation, and recombination of acquired information. Using computational modeling, we propose that internally generated sequences may be productively considered a component of goal-directed decision systems, implementing a sampling-based inference engine that optimizes goal acquisition at multiple timescales of on-line choice, action control, and learning.

Goal-directed and habitual behavior

Goal-directed behavior is distinct from habitual or reflexive behavior due to its deliberate, informed nature. Hallmarks of goal-directed behavior include sensitivity to the current value of the outcome [1] and the ability to generate action sequences that accomplish a desired outcome – even if a novel sequence of actions is required [2]. These abilities generally require knowledge of the causal structure of the environment (a world model), such that the consequences of actions can be predicted and used to guide behavior.

These features of goal-directed behavior stand in contrast to habitual or reflexive behaviors, which are conducted automatically, without taking into account the current value of behavioral goals. Over the past few decades, evidence has accumulated to suggest that a network of brain structures involving the HC, prefrontal cortex, amygdala, and ventral–medial sectors of the striatum controls goal-directed behavior (Box 1, Figure IA), whereas habits have been argued to be mediated by the dorsolateral striatum, its afferent sensorimotor cortices, and possibly infralimbic prefrontal cortex [3–6]. Recent computational models capture this distinction as follows. Goal-directed behavior has been simulated in 'model-based' learning schemes, in which the agent builds an explicit internal model of its state-space, containing stimulus-outcome and action-outcome relations that include the specific identity of the outcome (Box 2). This scheme contrasts with 'model-free' methods of reinforcement learning, where an entity such as a neural network comes to control behavior by learning stimulus-response mappings guided by a scalar reinforcement signal [7–9].

Here, we primarily focus on goal-directed behavior, specifically addressing the question how brain mechanisms support behavioral decisions and prospective planning in real time, based on acquired information that is not represented by direct neuronal responses to acutely available sensory input. Often organisms have only a few opportunities to learn from previous choices and are forced to making decisions 'on the fly', that is, based on their general understanding of the structure of their environment (e.g. [7]). Using examples from rodent navigation studies, we review recent experimental and theoretical work, the convergence of which suggests that internally generated sequences (IGSs; see Glossary) of neural activity, manifested in multiple subtypes and occurring across multiple brain structures, are instrumental for acquiring and conducting goal-directed behaviors.

Neural coding of state parameters and internally generated sequences

Traditionally, the problem of how the brain mediates goal-directed behavior has been cast in rather static neural coding schemes. The paradigmatic example is spatial coding in the HC, whose 'place cells' [10] support the idea

Corresponding author: Pennartz, C.M.A. (c.m.a.pennartz@uva.nl).

Keywords: forward sweep; generative models; hippocampus; decision making; reinforcement learning; spatial navigation; replay; inference; prospection; theta rhythm; ventral striatum.

1364-6613/

http://dx.doi.org/10.1016/j.tics.2014.06.011

Glossary

Action value: The value or expected return of executing a given action (usually, in a given state or context). In the reinforcement learning literature, action value is usually expressed as a Q-value that depends on both state and action: Q (state, action). Here, action value is expressed as the probability of achieving reward by R executing a given action A in a given state S: P(R|A,S).

Bayes' theorem: A mathematical procedure for updating probabilities or beliefs (e.g., of obtaining a reward) by combining prior information and evidence.

Cached value: . A scalar value that captures how much reward or punishment may be expected in the future, based on the history of returns obtained in the past. A cached value does not specify the nature or features of the outcome.

Dynamic Bayesian networks: A class of computational models performing statistical inferences over variables that unfold in time (e.g., what is the probability of reaching a state S while knowing the starting state and the executed action) following Bayes' rule [92].

Generative model of plan values: An internal model that encodes the probabilistic relations between states, actions, and rewards. Such a model permits to generate observable data (in this case, expected reward observations) given some other hidden (non-directly observable) parameters, and ultimately permits to estimate the value of a plan. The functioning of the specific submodels (dynamic model, observation model, and reward model) is explained in Figure 1.

Internally generated sequence (IGS): A sequence of multi-neuron firing activity that does not reflect an ongoing behavioral sequence (e.g., of actions, positions visited) but is instead generated on the basis of internal brain dynamics. IGSs may arise spontaneously or can be triggered by external cues [93].

Instrumental controller: A behavioral controller (i.e., a device for action selection) that supports instrumental behavior (i.e., an action aimed at achieving a goal or obtaining a reward) and can learn arbitrary actions (e.g., press a lever) to this purpose. In the reinforcement learning literature, instrumental controllers are usually divided into model-based and model-free classes and contrasted with Pavlovian controllers, which are limited to an innate and reinforcer-specific repertoire of actions (e.g., approach, salivate).

Mixed instrumental controller: A controller that combines (mixes) the functionalities of model-based and model-free ones [44] rather than arbitrating between them [8,94].

Partially observable Markov decision process (POMDP): A POMDP models decision processes of an agent whose state transitions possess the Markov property (i.e., the probability that the process moves to a new state S_{t+1} is only influenced by the current state S_t and selected action A) but for which the agent cannot directly observe the underlying (hidden) state and thus has to infer it from observations.

Pavlovian value: The value of a given state, for example, a place at which an agent is located. In the reinforcement literature this is usually expressed as a state-dependent value V(S) and contrasted with the aforementioned Q value that also depends on action. In probabilistic terms Pavlovian value can be expressed as the probability that a given state provides reward: P(R|S).

Plan value: The cumulative value of executing an action plan (e.g., a spatial or behavioral trajectory). Here, this value is calculated using IGSs: by stochastically retrieving successor states from hippocampal memory, considering their values (coded in VS), and integrating these states and values. The integration can follow a diffusion-to-bound rule [63,95] or a Bayesian scheme that uses cached Q values as priors [44].

Sampling-based inference engine: A device or neural architecture for inferring future events and system states from previously acquired knowledge. Here, inference is synonymous to probabilistic inference as performed by the generative model shown in Figure 1. Sampling-based means that the inference is approximate and uses a sampling method [92].

Value of information: The amount an agent should pay (or otherwise provide) to obtain new information prior to making a decision. In the current context, it indicates how many (computational) resources the agent should spend to refine prior value estimates using IGSs and model-based evaluation. Thus, the value of information can be considered a part of the trade-off between the costs and benefits of producing IGSs.

that this structure codes a spatial map of the environment of an animal [2,11]. The hippocampal spatial coding scheme has been recently expanded to incorporate other dimensions of state space, such as the temporal position of an agent in a behavioral sequence, or information about neutral and reward-predictive objects in the environment [12–16].

Other nodes of the network for goal-directed behavior that interact with the HC (Box 1, Figure IA) are thought to code further aspects of the state space of an organism beyond spatial location. Prefrontal cortical neurons, for instance, are thought to code a 'task space' as they have been implicated in coding task rules [17–20], but also plans, behavioral policies, and goals or goal locations [21–23]. The amygdala has been broadly implied in forming stimulus-outcome associations, and regulates not only Pavlovian responses, but also instrumental behavior via projections from basolateral amygdala to striatum [24,25].

Box 1. Multiple types of internally generated sequences

IGSs in the HC represent spatial locations and trajectories other than the current position of an animal. Most IGSs are associated with SWR complexes in the HC local field potential [30,32,37,99,100], that is, area CA3 generated, ~100 ms bouts of high-frequency oscillations (140–200 Hz). SWRs induce a state of high excitability in CA3 and CA1 networks [101] and likely influence neural activity in downstream areas such as medial prefrontal cortex and VS, where replay has been identified as well [100,102]. IGSs are time compressed [43,102–104], which likely aids associative learning mediated by spike-timing dependent plasticity [105,106], fast memory retrieval, and option evaluation for memory-based planning and decision making [33].

During slow-wave sleep and in awake states upon trajectory initiation, sequential patterns of recently activated place cells can be replayed in a forward direction, that is, the order of the original trajectory is preserved (Figure IB,C) [30,43,102–104,107]. At trajectory completion and reward consumption, however, replay can occur in the reverse direction (Figure IC) [31,32]. It is currently unknown whether forward and backward IGSs make different contributions to learning and/or decision making. However, two possibilities are that: (i) backward replay is particularly well suited for updating cached values of recently traveled paths or recent events, akin to the idea of post-reward decay of 'eligibility traces' in reinforcement learning [32]; and (ii) backward replay enables inference of what actions need to be taken to reach a desired goal [40]. These ideas remain to be tested experimentally. Suppression of sleep or awake SWRs has been found to impair behavioral performance, indicating their causal contribution to memory formation and retrieval [47,108,109]. Consistent with consolidation theories, replay was shown to occur cross-structurally between the HC and targets such as parietal and visual cortices and VS [43,110,111]. These areas may replay specific content of the composite memory trace corresponding to the multifaceted experience they represented earlier.

IGSs at the choice point of a task representing paths ahead of the rat were found during theta (6–10 Hz) oscillations in area CA3 ('forward sweeps'; Figure ID; [34]). VS reward-related units participate in forward sweeps [35], thereby likely contributing motivational information at decision moments.

In awake states, IGSs do not necessarily match the most recent behavioral experience, because they have been shown to reflect spatially remote trajectories [112] and combinations of trajectories that were never experienced as such [48]. In addition, IGSs can represent novel trajectories spanning between the current location of the rat and a well-known goal ([36]; Figure IE). These predictive sequences based on earlier unrelated behaviors may support planning and may contribute to 'cognitive map' formation.

Lastly, IGSs representing future, novel trajectories were found to be preplayed (Figure IF), either in forward or reverse order, in sleep or awake periods before mice entered a novel track and physically travelled through these place fields for the first time [37,113]. These IGSs are not driven by experience but seem to be preconfigured in the HC network as a scaffold for coding new experiences.

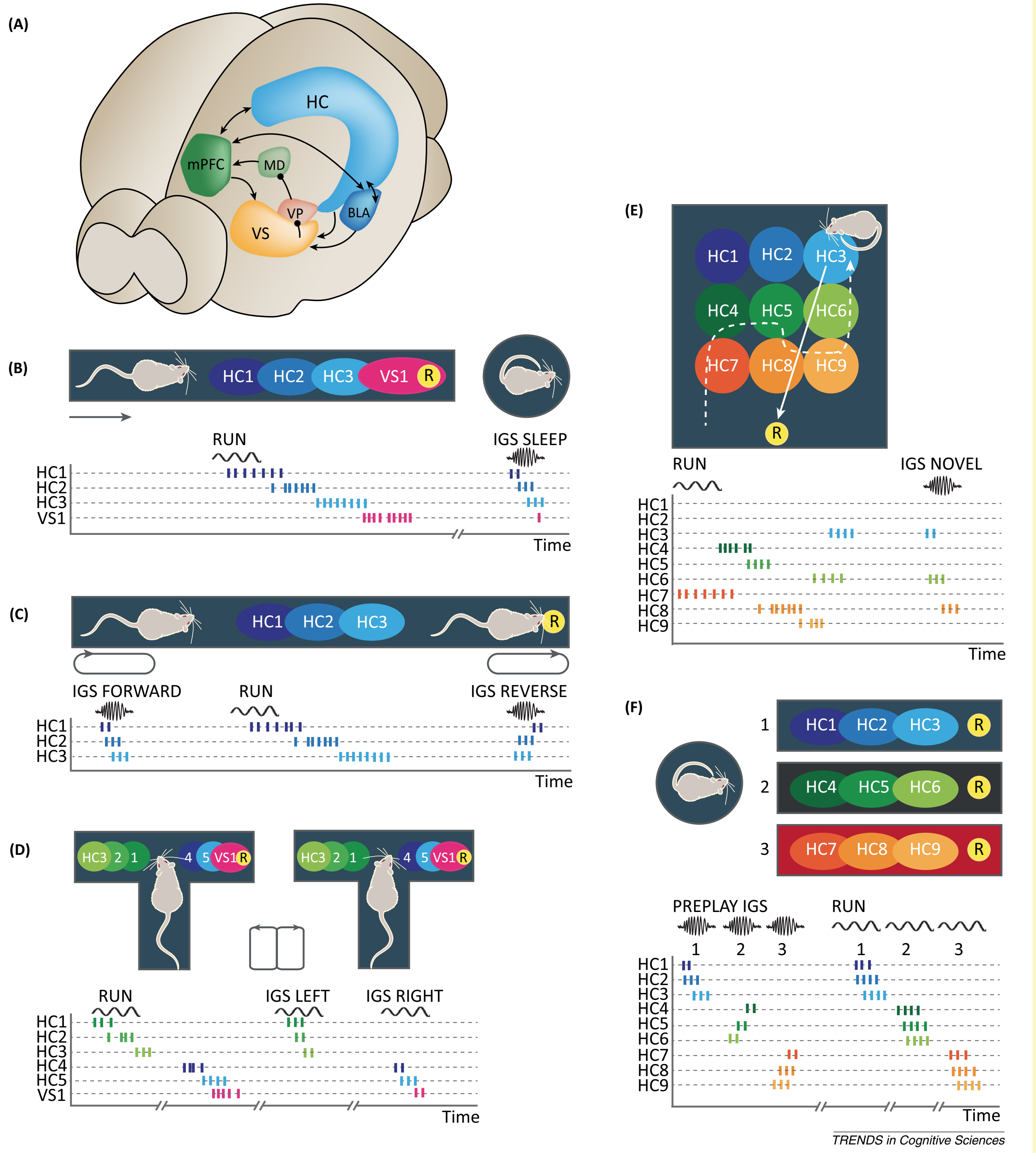

Figure I. Brain areas involved in goal-directed behavior and classes of IGSs. **(A)** Schematic of brain structures contributing to goal-directed behaviors as represented in the rodent brain. Projections between areas have been depicted with arrows (excitatory) and broken lines (inhibitory). (B–F) IGSs can be classified by their resemblance to the neuronal activation patterns during a behavioral experience. **(B)** As a rat runs along a track (RUN), hippocampal neurons (HC 1–3) are activated at specific locations (blue ellipses), whereas a ventral striatal neuron (VS1) is activated (magenta ellipse) around the reward location (R; yellow dot). Spike patterns during track running and sleep are plotted below the track schematic. During periods of rest and sleep after the running experience neuronal firing patterns are replayed in a forward direction (IGS SLEEP). The replayed IGSs are associated with SWR events in the HC local field potential, while theta oscillations (6–10 Hz) are the dominant rhythm during running. IGSs are compressed in time and may contain spikes from several brain areas including VS. **(C)** Awake replay. When a rat initiates a new journey on the track neuronal patterns can be replayed in a forward direction (IGS FORWARD). By contrast, when a rat pauses at the completion of a journey, IGSs can be replayed in a reversed direction (IGS REVERSE). **(D)** IGSs representing different available future options are generated at decision points on a track (IGS LEFT & RIGHT). These 'forward sweeps' are associated with theta oscillations. The same neurons were activated before when the rat explored both routes on the track (RUN). **(E)** Novel goal-oriented IGSs can be constructed from patterns of place cell activity generated during previous trajectories when a rat initiates a new route towards a goal site (yellow dot, R). Note that the order of the novel IGS does not correspond to the activation order of place cells during the journey of the rat into the arena (RUN; colored circles along dashed white line) but to the place cells that will be activated along the route leading to the goal (colored circles along solid white line). **(F)** Different possible place cell sequences [preplay IGS 1 (HC 1–3), 2 (HC 4–6) and 3 (HC 7–9) are preplayed during sleep or awake periods before a rat enters a novel track (either IGS 1, 2, or 3)]. Sequences preplay in forward or reversed order in equal probabilities. The sequences generated before entry on the novel track were not activated during preceding exploratory behavior of the animal. Abbreviations: BLA, basolateral amygdala; HC, hippocampus; IGS, internally generated sequence; MD, mediodorsal thalamus; mPFC, medial prefrontal cortex; SWR, sharp wave-ripple; VP, ventral pallidum; VS, ventral striatum.

Box 2. Model-free, model-based, and the space in between

In the canonical distinction between model-free and model-based control, the model-free controller chooses the action that has been most successful in the past. Popular algorithms for model-free learning and control, such as Q learning and temporal-difference reinforcement learning, demonstrate that the requisite running total of 'cached values' can be learned without knowledge of task structure (i.e., a world model). A model-based controller, in contrast, can generate online predictions based on specific relations between actions and outcomes, enabling flexible responding when cached values are uninformative or unavailable [8,9]. Thus, a model-free controller is tied to the repetition of past successful actions, whereas a model-based controller can improvise (for instance, plan a novel alternate route avoiding a roadblock).

Recent conceptualizations of model-free and model-based control are more subtle than the above dichotomy suggests. For instance, model-free controllers need to segment tasks into appropriate situations or 'states' before values can be assigned to them. Even for simple tasks, this requires determining which environmental features are relevant. More complex tasks can be tackled by constructing augmented state representations that contain estimates of hidden, past or expected states, which may rely on a world model [96]. Furthermore, model-based algorithms can 'train' the value estimates of a model-free controller off-line (e.g., DYNA [78]). Thus, model-free controllers can rely on model-based state representations and/or value functions, blurring the distinction between the two approaches. Computational studies have suggested that IGSs are capable of facilitating these interactions [97,98].

Online model-based controllers attempt to select the best action based on explicit prediction of the consequences of that action using a world model. Such a world model often takes the form of a matrix of transition probabilities between states, learned from experience. For instance, if you go left out of a novel elevator and end up in the departmental lounge, you update P (lounge|elevator, go left). This knowledge can then drive internal simulation thought to occur through IGSs. However, knowledge for model-based control need not be limited to direct experience. Some of the most useful applications of model-based control require making inferences based on limited data. For instance, 'shortcut' and 'preplay' IGSs suggest that the HC is capable of spatial inferences not constrained by direct experience [37,48].

The ventral striatum (VS) has been proposed to integrate hippocampal, amygdalar, prefrontal, and thalamic information to construct outcome predictions that invigorate motivated behaviors and support their selection [26–29]. Similarly to the HC, neurons in these connected structures are classically thought to fire primarily when the dimension to which they are tuned (e.g., rule, action, cue, and predictive value) is directly experienced by the animal.

This scheme leaves unresolved several questions the agent is confronted with at environmental choice points. In addition to 'where am I?', it must address, for instance: 'where did I come from?', 'what do I want?', and 'how should I get it?' For rapid, on-line decision making, these types of inference need to be available quickly and in a well- structured form, allowing arbitration between choice alternatives.

A breakthrough was made with the discovery that ensemble firing patterns in the HC are not confined simply to coding current location, but could include sequences corresponding to trajectories in space not currently experienced. Early interpretations of such 'replay', originally reported during post-task sleep [30], emphasized a role in memory consolidation, because a rehearsal of behavioral information initiated in the HC may drive memory strengthening (Box 1; Figure IB). More recently such sequences have also been found in the awake state, either during hippocampal sharp wave-ripple (SWR) activity (characterized by 140–200 Hz oscillations) during pauses, reward consumption, or rest (Box 1, Figure IC) [31–33], or as an ongoing phenomenon during the HC theta rhythm (6–10 Hz oscillations) associated with active behavior (manifested in part as 'forward sweeps'; Box 1, Figure ID; [34,35]). We refer to both types as internally generated sequences (IGSs). From the viewpoint of goal-directed behavior, IGSs are interesting not only as potential neural correlates of memory consolidation, but also as events that represent a retrieval (or sampling) of a memory trace that can be directly utilized to drive prospective decision making. The repertoire of 'out-of-field' activity patterns has been recently expanded to include sequences predictive of novel paths in a well-known environment (Box 1, Figure IE; [36]) and 'preplay' events (Box 1, Figure IF; [37]; here, all of these events are labeled IGSs).

Computational functions of internally generated sequences

How may IGSs be of use in the execution and acquisition of goal-directed behaviors? Taking the example of a spatial maze harboring hidden rewards at fixed locations, computational models have simulated neural networks mediating rodent navigation behavior, incorporating sensorimotor cortex, HC–entorhinal cortex, striatum, and connected areas such as prefrontal cortex and amygdala (e.g. [38–40]). Using path integration, the locomotion of rodents throughout the maze enables the HC to establish a place-cell map [41,42]. The HC emits spatial information, among others, to VS neurons, which come to represent the (reward-predictive) value of maze positions as a result of synaptic strengthening induced by paired neural activity coding for place (HC) and reward (VS) co-occurrences [38,43]. Here, we use VS as a focal point to illustrate a wider circuit important for valuation and motivation, which also includes the orbitomedial prefrontal cortices, ventral pallidum, and amygdala. Following a model-based scheme, the resultant predictive value coded by VS neurons is then used to evaluate states in terms of their predicted motivational consequences and to drive or invigorate behaviors such as approach to goal sites [14,26,44]. In this concept, medial prefrontal input transmits information about task constraints, defining under which conditions places or cues are valid predictors of reward [26]. Other striatal sectors are proposed to code outcome-predictive values based on different input dimensions, such as action information in the case of dorsomedial striatum, which has been implied in action-outcome learning [26,45,46].

A first function of IGSs highlighted by these models is the storage and consolidation of recently experienced associations, such as place–reward associations. During sleep, the HC and VS have been shown to replay spatial and reward information jointly and in a time-compressed fashion that is compatible with physiological time windows for

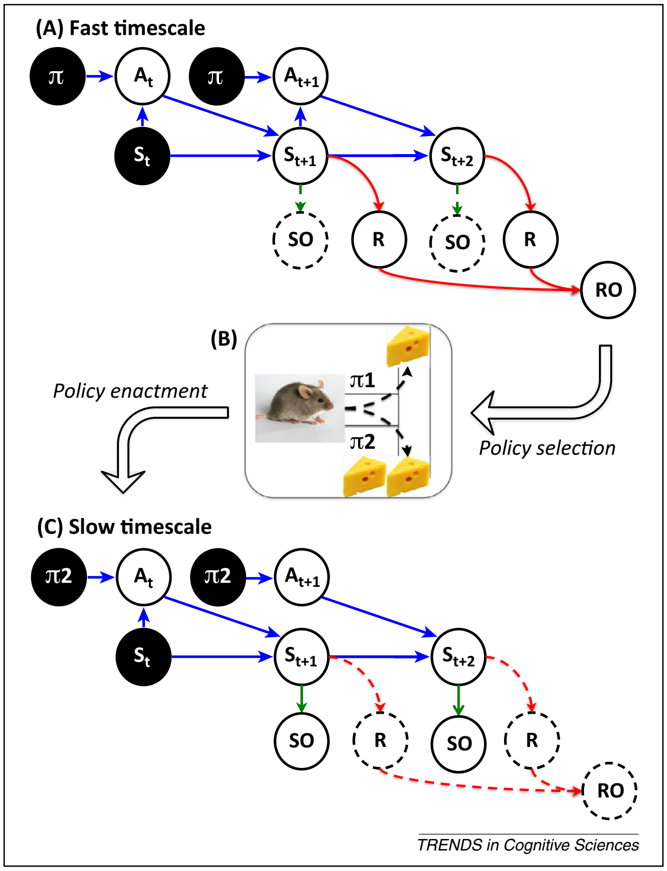

Figure 1. A generative model of plan values supporting vicarious trial-and-error and goal-directed behavior. The generative model, implemented by a dynamic Bayesian network [92], has the dual role of planning fictive outcomes for one or more policies at the two relatively fast timescales laid out within the action–perception cycle (see main text) and making inferences about partially observed states of the world at a third, slower timescale, laid out between cycles of action and perception. **(A)** Nodes indicate probability distributions over variables of interest: π is a policy or – with some approximation – action plan (e.g., a spatial trajectory), *A* an action (e.g., going right or left), *S* a state (e.g., a place visited), *SO* a sensory observation, *R* a reward, *RO* the sum of reward observations along a plan or path. In this framework, reward and value functions can be absorbed in the inferential process by treating them as prior beliefs [49,75,114]. Furthermore, the scheme permits to design reward functions which are more complex than the RO node shown here, and may also incorporate computational costs in the inference process [115]. State and action variables are time indexed (subscripts) and in the current example, time runs from *t* to *t+2*. Black nodes indicate variables that are known (clamped) before the inference process starts. Edges represent probabilistic relations between variables: blue arrows represent the dynamic model of the agent. This model encodes contingencies between actions and states and affords forward modeling: inferring the value of a given state using knowledge of the previous state and action: $P(S_{t+1}|S_t,A_t)$. Green arrows denote the observation model. In general, the system cannot directly observe its current state S_t, but it can infer it indirectly from sensory observation *SO* as well as from an estimation of its previous state S_{t-1}: $P(S|S_{t-1},SO)$. In this scheme, the HC formation might encode the observation model and parts of the dynamic model: the probabilistic relations between present and future states $P(S_{t+1}|S_t)$ that permit (re)generating state sequences (S_t, S_{t+1}, S_{t+2}, . . .), forming paths. During inference, the estimation of state sequences can be conditioned on other signals provided by other brain areas. For example, areas coding motor commands or proprioceptive information such as on head direction may provide action-related information and this may endow forward sweeps with directionality. Red arrows encode the reward model: $P(R|S)$. Similar to how states are estimated based on sensory observations, the probability and/or amount of reward available in a given state can be inferred from multiple observations of reward in that state. The VS is a candidate structure for encoding the reward model. **(B)** The proposed generative model can implement a form of vicarious trial-and-error to solve choice problems such as illustrated here. Here, an animal facing a decision point in a T-maze [34] compares the value of two different action plans (the competing policies π1 and π2 in A) by estimating probabilistically the rewards they may yield if executed. This problem is cast as estimating the probability of obtaining reward observations *RO* by starting from the current state S_t (which is assumed to be known, e.g., 'being here') and executing one of the two competing policies (π1 or π2; in case of π1 the inference would be $P(RO|S_t, \pi1)$. This estimation is accomplished by 'internally simulating' the execution of the policy (which requires clamping the π node to π1 and the S node to the current location S_t of the animal) to predict possible reward observations RO. To do this, both the dynamic model and reward model are needed: the former permits to foresee which states will be visited if one executes policy π1. The latter permits to retrieve the reward values associated with each state along the path, which need to be summed up to estimate the total value of the path. The probabilistic inference performed here is approximate (sampling-based) [92]: successor states (in the HC internal model) are retrieved stochastically and their values (in the VS internal model) integrated to form an estimate of plan values, like in sequential sampling choice models [65]. IGSs represent stochastic samples across the dynamic model and generative model of reward. As a next step, the other policy π2 can be tested in a similar way, again given the current state S_t: $P(RO|S_t, \pi2)$. The values of the two policies can then be compared for policy selection, which corresponds here to the selection of an action plan or spatial trajectory. For this fast type of inference, sensory observations are not used, hence the corresponding nodes and edges are represented by broken lines. **(C)** The selected policy (π2) is subsequently enacted during overt behavioral performance. Here, the same generative model can be used to sample sensory outcomes (SO) and inferring hidden world states (S). The policies and actions need not be inferred (because they were already selected at step B) but are treated as prior beliefs, and updated based on sensory observations. Now, the generative model of reward is not being used, hence its nodes and edges are represented by broken lines. This type of inference occurs more slowly as compared to (A), that is, at the third timescale already mentioned. Inferences occurring at fast timescales can be nested within slower types to optimize planning and overt action control [40,51].

synaptic plasticity [43], but during maze exploration forms of backward or forward replay also occur (Box 1), raising the possibility that these support primary acquisition of place preferences.

Second, IGSs may augment the neural coding of current input to create a richer state representation, which may for instance include information about the previous route of the animal to its current position (to exclude visiting a site where all reward had been consumed previously). Here, awake replay of the path sequence leading up to the current position of the agent may support working memory [47].

A third function of IGSs is derived from Tolman's idea [2] of 'vicarious trial-and-error' and holds that IGSs subserve internal simulation of behavioral sequences and their motivational consequences. When the model agent (Figure 1) reaches a decision point, it performs a series of internal simulations, which in turn activate specific place cells and VS neurons according to previously acquired associations [44]. Replay events and theta-sequenced activity (including 'forward sweeps') in HC and VS have been advanced as plausible neural correlates of this forward-modeling process [34,36,48]. It is not intuitively clear what a mere rumination of 'old' information in an internal model may contribute to decision making in addition to memory storage and retrieval. In the model of Pezzulo *et al*. [44], an important determinant of choosing a particular path at a choice point is the 'total value' of the action plan (e.g., a path or spatial trajectory), computed by sequentially accumulating all activity values of striatal neurons associated with the goal linked to that plan, or even across multiple candidate plans [49]. Because real-world pairings of place–reward experiences are scarce and the estimates of the 'cached' values of actions are thus inaccurate, IGSs are proposed to work as a sampling-based inference engine and allow the system to stochastically resample the distribution of place–place and place–reward contingencies represented in the model. Thereby, they reduce the variance in the total-value estimate, enabling an improvement in informed choice. This ability is particularly important in situations where previously cached values are no longer relevant, such as occurs during a motivational shift.

A fourth function we propose for IGSs is the recombination of previously acquired elements to generate novel constructs, for example, new place–place sequences. These recombinations enable systems to plan novel paths to known goals ([36]; Box 1, Figure IE), to improve self and goal localization, and to build an integrated map of their environment or task space, potentially promoting novel insight [40,48,50]. As for internal simulation, this function can be subsumed under the concept of a sampling-based inference engine, but now emphasizing that stochastic resampling from memory results in combinations not experienced by the agent before. In summary, we propose that IGSs constitute a multi-purpose resource that uses a common sampling-based mechanism to support learning and several executive brain functions.

These four functions of IGSs require neural processes operating at multiple timescales. If planning is based on some form of approximate Bayesian inference, the fastest timescale can be associated with the stochastic sampling of probability distributions that forms the basis of action selection. Specifically, we can conceive of the brain as sequentially sampling fictive outcomes of a given plan or policy using IGSs [44]. The second timescale pertains to serial sampling over different policies, which permits to perform an evaluation of alternative plans. These two timescales are compatible with the action–perception cycle of the animal, therefore, IGSs can be used for vicarious trial and error, and rapid prospective planning at decision points. Planning in partially observable environments (or, more technically, partially observable Markov decision processes; POMDPs) requires inferring hidden (nonobservable) states of the world (e.g., 'where will I be next if I follow a given plan?'). In probabilistic inference schemes this can be accomplished by sampling the sensory consequences of the selected action plan at a third, slower timescale, viz., that of the action–perception cycle [40,51]. Figure 1 considers these different, nested timescales: the two faster timescales of sampling fictive outcomes within the action–perception cycle (Figure 1A) and the slower timescale of sampling sensory consequences and state inference between action–perception cycles (Figure 1B). The latter, slower state inference process also comes into play during action execution to ensure that the behavioral plan can fulfill the expectations and to adapt it to unpredicted circumstances. Critically, we propose that the same generative model is used for both planning and action execution, and that this entity produces IGSs which model predicted or actual transitions between states. Finally, we distinguish a fourth, slow timescale, laid out across many instances of the action–perception cycle, which enables the learning of time-invariant parameters of the generative model, needed to optimize future choices. At this slowest timescale (e.g., during sleep), IGSs afford self-training by repeatedly resampling past or fictive experience (Figure 2A).

We focus on a sampling-based inference scheme for two reasons. First, stochastic sampling permits to address large planning problems, where the finite size of a theta cycle and SWR events poses challenges. We address this issue in the next section. Second, samples of simulated experiences and 'fake reward observations' (e.g., sample trajectories produced by replay or preplay in the HC–VS circuit) can be used to update values in a neurally plausible fashion, which parallels the way temporal-difference reinforcement learning methods update the estimate of action values [52]. Common alternatives to sampling-based approaches include predictive coding and variational Bayesian schemes. These schemes also solve planning and choice problems using forward and backward sweeps over a fictive future [51].

Hippocampal–striatal system as an inference engine for goal-directed choice and vicarious trial and error

To explain the hypothesis of a sampling-based inference engine further, we consider the HC and VS in more detail. Focusing the model on spatial reward search, these structures jointly encode a generative model of plan value: the HC encodes place–place pairs, or longer sequences thereof, and regenerates these representations in IGSs. By contrast, the VS encodes place–reward pairs, resulting in firing activity patterns reflecting reward expectation (or value) coupled to specific environmental locations (cf. [53]). Both structures can generate place–place or place–reward sequences not experienced before. The combined HC–VS information is thus cast as a type of 'plan value' inference that can be used for decision making and vicarious trial and error (Figure 1). HC IGS activity simulates possible paths and, simultaneously, the VS generates covert reward expectations associated with specific path locations [54]. This mechanism enables an improvement in informed choice over a purely habitual mechanism when experience is too scarce to estimate the total value of a path robustly, and has the flexibility of a model-based system that is sensitive to devaluation and motivational shifts (Box 3). A similar model-based approach may support other functions, including self-localization and route planning by sampling place–place sequences or place-sensory observation contingencies, which may be encoded jointly by the HC and entorhinal cortex [40,49,55].

So far, experimental tests of the content and behavioral relevance of IGSs have been limited to relatively simple environments, such as T-mazes and open fields. In these settings, IGS content seems to have a modest but significant bias towards the intended goal [36,56]. This suggests that IGS content may direct subsequent decision making, particularly given the plausibility of IGSs in the HC being associated with evaluative components downstream. However, it is unclear how this proposal scales to more complex (deep) environments. Possible scaling solutions include the fact that: (i) IGSs can span multiple SWRs, which allows representation of prolonged paths [57]; (ii) IGSs may include representations at multiple (larger) scales, such as those that reflecting wide place fields in ventral HC or coarse time scales in medial prefrontal cortex [58,59]; (iii) IGS mechanisms may exploit the structure of the environment (e.g., the presence of *en route* bottlenecks or landmarks) to reduce the search space; and (iv) IGSs may be interleaved or otherwise interact with model-free cached values and only explore the search space in a piecewise manner (e.g., only up to states whose (intermediate)

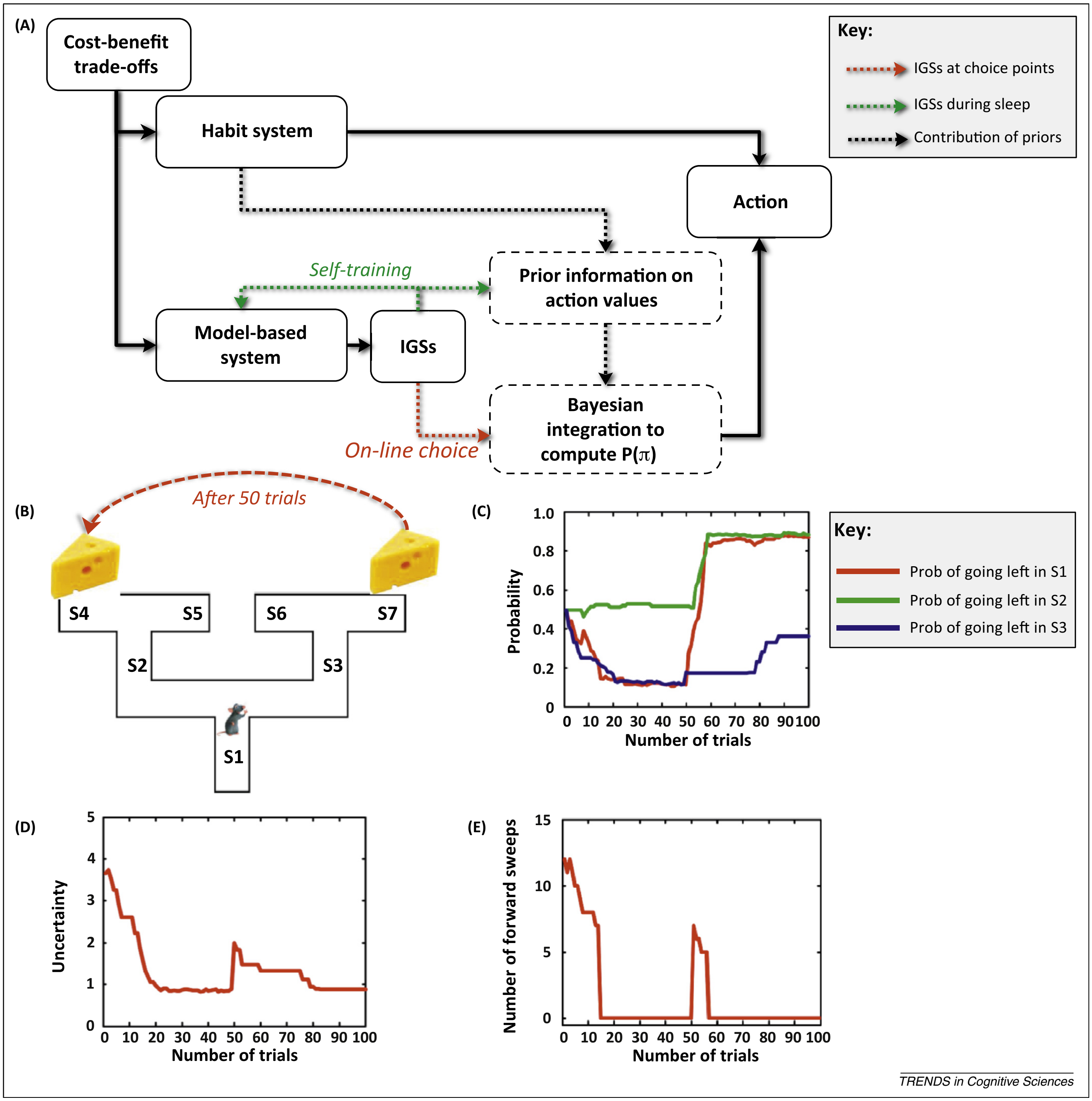

Figure 2. Computational scheme with internally generated sequences (IGSs) in a mixed instrumental controller. **(A)** In this scheme [44], cached action values provided by model-free systems are represented as Bayesian priors ('prior information on action values'). Cost–benefit considerations suggest that, under certain conditions such as after sufficient training or low uncertainty, prior (or cached value) information is sufficient to make an online decision. If not, the model-based system of Figure 1 is recruited and IGSs are executed to reduce the uncertainty in the estimation of action plan values. In this case, the stochastic sampling of reward observations (*RO*; Figure 1) is combined with the priors using Bayes' rule, hence producing a more robust plan value estimate that combines model-free and model-based elements. P(π) is the probability of selecting a policy π. Besides for online choice, the same IGS-based mechanism can be used offline (e.g., during sleep) to self-train the generative model and to refine priors (i.e., train the model-free system) so that subsequent choices can be made without recruiting the model-based system [78]. We propose that the neural architecture supporting the larger instrumental controller includes at minimum the HC–VS nexus, dorsomedial and dorsolateral striatum, orbitomedial prefrontal cortex, and amygdala. The HC–VS nexus may jointly encode the generative model shown in Figure 1. The prefrontal cortex is proposed to play a dual role. On the one hand, it may encode policies and rules that contextualize place–outcome, action–outcome contingencies and other task-relevant information, and maintain this information in working memory. On the other hand, the prefrontal cortex may support cost–benefit analysis and finding the optimal balance of information-seeking versus exploitation using multiple sources of information, including the uncertainty of action choices and outcomes [116–118], task volatility [1,94] and the so-called value of information, which reports the expected benefit for the decision that might be provided by the model-based system [44,94]. These two aspects may be particularly associated, in the primate brain, with (dorso)lateral and medial portions of prefrontal cortex, respectively. The dorsolateral striatum may harbor model-free, habitual control mechanisms and provide priors on cached values. Finally, the amygdala, ventral pallidum, and hypothalamus might contextualize values by considering the current motivational needs of the animal. **(B)** Simulation results from a model of the mixed instrumental controller described in [44]. A simulated rat performs multiple runs in a double T-maze where the reward location is first kept constant at S7 but subsequently switches to S4 after 50 trials. **(C)** Initially, the rat learns an optimal policy fast, with a decreased probability of turning left at S1 (red line) favoring a turn to the right at this choice point, and a decreased probability of turning left at S3 (blue line), favoring a turn to the right here. After 50 trials the model is quickly adjusted to favor turning to the left at S1 and to the left at S2. **(D, E)** As uncertainty of action outcome decreases during learning, also the prevalence of forward sweeps decreases, marking a transition from goal-directed to habitual control. When the reward location is switched from S7 to S4 after 50 iterations, the uncertainty rises again and forward sweeps are reinstated until the rat has learned the new contingencies. Parts (B)–(E) were originally published in [44]. Abbreviations: HC, hippocampus; VS, ventral striatum.

Box 3. Are internally generated sequences part of a goal-directed decision-making mechanism?

According to an influential theory of instrumental behavior, goal-directed action requires the representation of an outcome at the moment an action is selected [1]. A covert reward expectation in VS elicited during forward sweeps [54] may meet this criterion. Here, the system is thought to use an outcome expectation for choice behavior, not a value accumulated over the history of past reinforcements. For a behavior to qualify as 'goal-directed' it is also necessary that a contingency exists between an instrumental action and the outcome [1]. Lesion studies indicate that VS activity is not required *per se* to meet this more stringent criterion. However, we propose that, during IGSs, joint HC and VS representations do establish a contingency between potential paths, conceived as behavioral sequences or plans in the HC [34,36,119] and a rewarding outcome, coded in the VS. The specific association of an IGS-coded action sequence and reward expectation may meet this second criterion for goal-directed behavior.

The idea that IGSs in the HC–VS subsystem are part of a goal-directed mechanism is not antithetic to the currently predominant hypothesis that the VS encodes Pavlovian values [120,121]. These views can be reconciled by virtue of Pavlovian values contributing to instrumental responses in multiple ways, as evident in Pavlovian-to-instrumental transfer [122]. Neurophysiologically, IGSs may covertly elicit stimulus- or location-driven reward expectations that can be used to calculate plan values and to invigorate or select actions. More broadly, the HC–VS projection provides one out of several mechanisms supporting action invigoration or selection in the basal ganglia, with different striatal territories encoding different contingencies (e.g., action–outcome) jointly contributing to selection [26].

At first glance, the role we attribute to the HC–VS subsystem may also seem to contradict the idea that dorsomedial (but not ventral) sectors of the striatum support goal-directed choice. Indeed, the dorsomedial (not ventral) striatum has been implicated in goal-directed learning and the formation of action–outcome contingencies [45,46,123]. We propose that the HC–VS subsystem supports goal-directed choice via a different route than action–outcome contingency learning: a mechanism that forms and evaluates plan representations 'on the fly' using IGSs and the generative model encoded by this subsystem (see Figure 1 in main text). Plans and plan values shaped in this manner can be used to invigorate action, either by direct VS output to effector systems or by interactions with dorsal striatal territories [26].

value was previously cached). State of the art planning algorithms (e.g., Monte Carlo planning [60,61]; Box 2) are able to address large search problems and provide testable predictions on which strategies and heuristics might be used in deep environments where extensive search is not feasible. Such Monte Carlo type simulations have also been proven useful in modeling semantic memory formation using offline replay of individual episodic memory traces [62].

The next important question is how this simulated activity can be used online to drive action selection. Similar to sampling-based theories of choice in economic and perceptual domains [63–65], actions can be selected by integrating multiple samples of the value of different paths or action plans serially, and then comparing the alternatives. Although it is unknown how the plan carrying maximal value is selected, various mechanisms for competition in the striatum, or downstream basal ganglia structures, have been proposed [66–69]. VS output resulting from this competitive selection may drive basic behaviors such as consumption behavior or Pavlovian approach directly [70,71], or may alternatively affect action-value and action-selection mechanisms in the dorsal striatal system via (in)direct connections, such as via VS–ventral tegmental area–dorsal striatum feedback loops [72–74].

The role of internally generated sequences in wider brain networks for instrumental control

The sampling-based inference scheme leaves open the issue of when, during learning and decision making, IGSs are useful and when they are not. We recently proposed [44] that an IGS-based sampling mechanism is part of a wider brain network that combines the functionalities of model-based and model-free controllers for decision making and learning using cost–benefit considerations (Figure 2). From a normative perspective, a key computational imperative is to reduce uncertainty (or optimize precision) on current and future states, and their values, before a choice is made [51,75]. When a given action plan has reliably yielded reward in the past, this prior information has a high precision. Under this condition, inference may be avoided to save energetic costs and enable rapid action. The agent can directly drive behavior using the cached action values of the model-free system (represented as Bayesian priors in Figure 2). This fallback on the model-free system corresponds to acting by habit [76]. When high certainty is lacking, new evidence should be integrated prior to making an informed decision. One way to gather evidence online is to execute an overt information-seeking action, for example, a saccade [51] or a locomotor exploration of unknown locations. We suggest IGSs play a similar, but internal information-seeking and explorative role both during online decision making and offline consolidation.

When used during choice or pre-choice periods, IGSs thus play a covert, information-seeking role and produce 'fake reward observations'. In the HC–VS model (Figure 2), these are used within a Bayesian scheme, where cached action values (provided by the model-free subsystem) act as priors that are updated with the sampled evidence (provided by IGSs) to produce better estimates of these values. This procedure can be repeated to compare all possible action sequences and is consistent with the characteristic serial nature of forward IGSs [44] (Box 1).

In novel environments IGSs may train the model and construct place–place and place–reward predictions from episodic events. The same computational scheme can support both model-building and model-free learning offline, when the agent is far from choice points. For example, during sleep or behavioral pauses IGSs can be used to train the model-free system (i.e., to modify the priors on cached action values; Figure 2) by replaying and simulating experience, as previously exemplified by the DYNA architecture [77,78].

Concluding remarks

IGSs in the HC and connected structures such as the VS have been previously interpreted as a vehicle for the offline consolidation of one-shot, episodic-like memories, and as the online recall of recent memory. Here, we reviewed evidence that IGSs may be productively considered a component of decision and action systems – a sampling based inference engine – which attempts to optimize reward acquisition at the multiple timescales of online

Box 4. Outstanding questions

- What are the roles of IGSs associated with forward SWRs, backward SWRs, and forward sweeping in theta sequences in online decision making? Do these three forms make the same, or dissociable, contributions? More specifically, can the demonstrated role of SWR IGSs in choosing maze arms [47] be understood as working-memory or state-space augmentation, as search, or otherwise? What is the causal connection, if any, between the content of theta sequences and subsequent choice?
- Which factors determine the content of IGSs? Can we identify optimality principles and accompanying meta-control problems that govern the balance between exploitation (of cached values) and exploration (of internal models coded in memory structures and inferences expressed through IGSs)?
- How do IGS-based mechanisms scale to more complex (deep) environments where extensive search is not feasible?
- Do IGSs occur in humans? It is not clear how IGSs similar to those found in rodents may relate to the subjective, conscious experience in humans accompanying HC-mediated cognitive processes such as scene imagination and episodic future thinking. At the physiological level, place cell recordings in humans are not yet of sufficient density to determine sequences, although individual cells do participate in recall [124].
- Specific predictions arising from the IGS-as-planning hypothesis need to be tested. These include: (i) IGS content should be sensitive to motivational state, as subjected to Pavlovian stimuli and context; (ii) IGSs should preferentially occur at times of uncertainty, and (iii) IGSs should preferentially sample the most informative trajectories in larger, complex environments.

choice, action execution and learning. In addition to fulfilling functions in the consolidation and online augmentation of representations, computational studies highlight how IGSs may serve internal simulation and recombination of previously acquired information in a sampling-based inference scheme. These functions of IGSs support both online decision making during flexible behaviors and model-based learning, and are proposed to interact with brain structures involving the dorsolateral striatum to contribute to model-free (habitual) behavior.

By the same token, this proposal offers a perspective on integrating and reconciling different theories of HC function: on the one hand, the use of IGSs in HC target areas involved in outcome expectation and motivation may explain how the HC contributes to goal-directed spatial navigation, learning of complex motor sequences and the linkage between episodic and procedural memory systems. On the other hand, IGS-based mechanisms are fully compatible with the role of the human HC in episodic memory, ‘mental time travel’ and the production of novel, imagined constructs from memory [79–81]. The HC is closely associated with the ‘default mode network’ which includes posterior cingulate, lateral posterior parietal, and medial prefrontal regions, and has been implicated in internally oriented cognitive processes such as thinking about the past and future, imagery, and mind wandering [82–88]. Because this network has been identified mainly on the basis of functional magnetic resonance imaging (fMRI) activity in humans, it remains to be investigated whether it can sustain fast internally generated sequences supporting internal modeling and simulation (Box 4).

Although we reviewed data indicating IGS-based forms of prospection in rodent spatial navigation, the proposed mechanism may play a more general role in action and perception systems. Computational models have used IGSs for action execution (e.g., handwriting) and observation in the context of the mirror neuron system [89], in skilled robot action [90], and as a mechanism shedding light on information flow in the brain and its instabilities [91]. Because the content of IGSs can be experimentally accessed at fine timescales, predictions from these models can be directly tested.

Acknowledgments

We thank Guido T. Meijer for constructing Box 1, figure IA, and Paul F.M.J. Verschure and two anonymous referees for their helpful comments. The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. FP7-ICT-270108 (Goal-Leaders) to G.P. and C.M.A.P., and from The Netherlands Organization for Scientific Research (to C.S.L.: VENI 863.11.0101; to C.M.A.P.: ALW 820.02.020; to M.A.A.v.d.M.: VENI 863.10.013).

References

1 Balleine, B.W. and Dickinson, A. (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* 37, 407–419

2 Tolman, E.C. (1948) Cognitive maps in rats and men. *Psychol. Rev.* 55, 189–208

3 Bornstein, A.M. and Daw, N.D. (2011) Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Curr. Opin. Neurobiol.* 21, 374–380

4 Gruber, A.J. and McDonald, R.J. (2012) Context, emotion, and the strategic pursuit of goals: interactions among multiple brain systems controlling motivated behavior. *Front. Behav. Neurosci* 6, 50

5 Smith, K.S. *et al.* (2012) Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18932–18937

6 van der Meer, M. *et al.* (2012) Information processing in decision-making systems. *Neuroscientist* 18, 342–359

7 Niv, Y. *et al.* (2006) Choice values. *Nat. Neurosci.* 9, 987–988

8 Daw, N.D. *et al.* (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704–1711

9 Sutton, R.S. and Barto, A.G. (1998) *Reinforcement learning: an introduction*, MIT Press

10 O’Keefe, J. and Dostrovsky, J. (1971) The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* 34, 171–175

11 O’Keefe, J. and Conway, D.H. (1978) Hippocampal place units in the freely moving rat: why they fire where they fire. *Exp. Brain Res.* 31, 573–590

12 Wood, E.R. *et al.* (1999) The global record of memory in hippocampal neuronal activity. *Nature* 397, 613–616

13 Leutgeb, S. *et al.* (2005) Independent codes for spatial and episodic memory in hippocampal neuronal ensembles. *Science* 309, 619–623

14 Lansink, C.S. *et al.* (2012) Reward cues in space: commonalities and differences in neural coding by hippocampal and ventral striatal ensembles. *J. Neurosci.* 32, 12444–12459

15 Pastalkova, E. *et al.* (2008) Internally generated cell assembly sequences in the rat hippocampus. *Science* 321, 1322–1327

16 Kraus, B.J. *et al.* (2013) Hippocampal “time cells”: time versus path integration. *Neuron* 78, 1090–1101

17 Wallis, J.D. *et al.* (2001) Single neurons in prefrontal cortex encode abstract rules. *Nature* 411, 953–956

18 Durstewitz, D. *et al.* (2010) Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. *Neuron* 66, 438–448

19 Mante, V. *et al.* (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* 503, 78–84

20 Kehagia, A.A. *et al.* (2010) Learning and cognitive flexibility: frontostriatal function and monoaminergic modulation. *Curr. Opin. Neurobiol.* 20, 199–204

21 Mulder, A.B. *et al.* (2003) Learning-related changes in response patterns of prefrontal neurons during instrumental conditioning. *Behav. Brain Res.* 146, 77–88

22 Hok, V. *et al.* (2005) Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* 102, 4602–4607

23 Genovesio, A. *et al.* (2012) Encoding goals but not abstract magnitude in the primate prefrontal cortex. *Neuron* 74, 656–662

24 Parkinson, J.A. *et al.* (1999) Dissociation in effects of lesions of the nucleus accumbens core and shell on appetitive pavlovian approach behavior and the potentiation of conditioned reinforcement and locomotor activity by D-amphetamine. *J. Neurosci.* 19, 2401–2411

25 Di Ciano, P. *et al.* (2001) Differential involvement of NMDA, AMPA/kainate, and dopamine receptors in the nucleus accumbens core in the acquisition and performance of pavlovian approach behavior. *J. Neurosci.* 21, 9471–9477

26 Pennartz, C.M.A. *et al.* (2011) The hippocampal-striatal axis in learning, prediction and goal-directed behavior. *Trends Neurosci.* 34, 548–559

27 Voorn, P. *et al.* (2004) Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci.* 27, 468–474

28 Mendelsohn, A. *et al.* (2014) Between thoughts and actions: motivationally salient cues invigorate mental action in the human brain. *Neuron* 81, 207–217

29 Roesch, M.R. *et al.* (2009) Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. *J. Neurosci.* 29, 13365–13376

30 Wilson, M.A. and McNaughton, B.L. (1994) Reactivation of hippocampal ensemble memories during sleep. *Science* 265, 676–679

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

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

33 Carr, M.F. *et al.* (2011) Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nat. Neurosci.* 14, 147–153

34 Johnson, A. and Redish, A.D. (2007) Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J. Neurosci.* 27, 12176–12189

35 van der Meer, M.A.A. *et al.* (2010) Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron* 67, 25–32

36 Pfeiffer, B.E. and Foster, D.J. (2013) Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* 497, 74–79

37 Dragoi, G. and Tonegawa, S. (2011) Preplay of future place cell sequences by hippocampal cellular assemblies. *Nature* 469, 397–401

38 Chersi, F. *et al.* (2013) Mental imagery in the navigation domain: a computational model of sensory-motor simulation mechanisms. *Adaptive Behav.* 21, 251–262

39 Erdem, U.M. and Hasselmo, M. (2012) A goal-directed spatial navigation model using forward trajectory planning based on grid cells. *Eur. J. Neurosci.* 35, 916–931

40 Penny, W.D. *et al.* (2013) Forward and backward inference in spatial cognition. *PLoS Comput. Biol.* 9, e1003383

41 Mittelstaedt, M.L. and Mittelstaedt, H. (1980) Homing by path integration in a mammal. *Naturwissenschaften* 67, 566–567

42 McNaughton, B.L. *et al.* (2006) Path integration and the neural basis of the 'cognitive map'. *Nat. Rev. Neurosci.* 7, 663–678

43 Lansink, C.S. *et al.* (2009) Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biol.* 7, e1000173

44 Pezzulo, G. *et al.* (2013) The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Front. Psychol.* 4, 92

45 Yin, H.H. *et al.* (2008) Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *Eur. J. Neurosci.* 28, 1437–1448

46 Balleine, B.W. *et al.* (2009) The integrative function of the basal ganglia in instrumental conditioning. *Behav. Brain Res.* 199, 43–52

47 Jadhav, S.P. *et al.* (2012) Awake hippocampal sharp-wave ripples support spatial memory. *Science* 336, 1454–1458

48 Gupta, A.S. *et al.* (2010) Hippocampal replay is not a simple function of experience. *Neuron* 65, 695–705

49 Botvinick, M. and Toussaint, M. (2012) Planning as inference. *Trends Cogn. Sci.* 16, 485–488

50 Wagner, U. *et al.* (2004) Sleep inspires insight. *Nature* 427, 352–355

51 Friston, K. *et al.* (2012) Active inference and agency: optimal control without cost functions. *Biol. Cybern.* 106, 523–541

52 Kearns, M. *et al.* (2002) A sparse sampling algoritm for near-optimal planning in a large Markov decision process. *Machine Learn.* 49, 193–208

53 Lansink, C.S. *et al.* (2008) Preferential reactivation of motivationally relevant information in the ventral striatum. *J. Neurosci.* 28, 6372–6382

54 van der Meer, M.A.A. and Redish, A.D. (2009) Covert expectation-of-reward in rat ventral striatum at decision points. *Front. Integr. Neurosci.* 3, 1

55 Solway, A. and Botvinick, M.M. (2012) Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychol. Rev.* 119, 120–154

56 Singer, A.C. *et al.* (2013) Hippocampal SWR activity predicts correct decisions during the initial learning of an alternation task. *Neuron* 77, 1163–1173

57 Davidson, T.J. *et al.* (2009) Hippocampal replay of extended experience. *Neuron* 63, 497–507

58 Kjelstrup, K.B. *et al.* (2008) Finite scale of spatial representation in the hippocampus. *Science* 321, 140–143

59 Hyman, J.M. *et al.* (2012) Contextual encoding by ensembles of medial prefrontal cortex neurons. *Proc. Natl. Acad. Sci. U.S.A.* 109, 5086–5091

60 Silver, D. and Veness, J. (2010) Monte-Carlo Planning in Large POMDPs. In *NIPS* (Lafferty, J.D. *et al.*, eds), pp. 2164–2172, Curran Associates

61 Geffner, H. and Bonet, B. (2013) A concise introduction to models and methods for automated planning. *Synthesis Lect. Artif. Intell. Machine Learn.* 8, 1–141

62 Battaglia, F.P. and Pennartz, C.M. (2011) The construction of semantic memory: grammar-based representations learned from relational episodic information. *Front. Comput. Neurosci.* 5, 36

63 Ratcliff, R. (1978) A theory of memory retrieval. *Psychol. Rev.* 85, 59–108

64 Stewart, N. *et al.* (2006) Decision by sampling. *Cognit. Psychol.* 53, 1–26

65 Shadlen, M.N. and Kiani, R. (2013) Decision making as a window on cognition. *Neuron* 80, 791–806

66 van Dongen, Y.C. *et al.* (2005) Anatomical evidence for direct connections between the shell and core subregions of the rat nucleus accumbens. *Neuroscience* 136, 1049–1071

67 Leblois, A. *et al.* (2006) Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *J. Neurosci.* 26, 3567–3583

68 Redgrave, P. *et al.* (2011) Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. *Neuroscience* 198, 138–151

69 Zhang, J. *et al.* (2009) A neural computational model of incentive salience. *PLoS Comput. Biol.* 5, e1000437

70 Dalley, J.W. *et al.* (2005) Time-limited modulation of appetitive Pavlovian memory by D1 and NMDA receptors in the nucleus accumbens. *Proc. Natl. Acad. Sci. U.S.A.* 102, 6189–6194

71 Baldo, B.A. and Kelley, A.E. (2007) Discrete neurochemical coding of distinguishable motivational processes: insights from nucleus accumbens control of feeding. *Psychopharmacology (Berl.)* 191, 439–459

72 Pennartz, C.M. *et al.* (2009) Corticostriatal Interactions during learning, memory processing, and decision making. *J. Neurosci.* 29, 12831–12838

73 Haber, S.N. *et al.* (2000) Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J. Neurosci.* 20, 2369–2382

74 Belin, D. and Everitt, B.J. (2008) Cocaine seeking habits depend upon dopamine-dependent serial connectivity linking the ventral with the dorsal striatum. *Neuron* 57, 432–441

75 Friston, K. (2010) The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138

76 Dolan, R.J. and Dayan, P. (2013) Goals and habits in the brain. *Neuron* 80, 312–325

77 Gershman, S.J. *et al.* (2014) Retrospective revaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.* 143, 182–194

78 Sutton, R.S. (1990) Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*. pp. 216–224 Bellevue

79 Tulving, E. (1983) *Elements of episodic memory*, Clarendon Press

80 Huijbers, W. *et al.* (2011) Imagery and retrieval of auditory and visual information: Neural correlates of successful and unsuccessful performance. *Neuropsychologia* 49, 1730–1740

81 Hassabis, D. *et al.* (2007) Patients with hippocampal amnesia cannot imagine new experiences. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1726–1731

82 Raichle, M.E. *et al.* (2001) A default mode of brain function. *Proc. Natl. Acad. Sci. U.S.A.* 98, 676–682

83 Daselaar, S.M. *et al.* (2010) Modality-specific and modality-independent components of the human imagery system. *Neuroimage* 52, 677–685

84 Andrews-Hanna, J.R. *et al.* (2010) Functional-anatomic fractionation of the brain's default network. *Neuron* 65, 550–562

85 Szpunar, K.K. *et al.* (2007) Neural substrates of envisioning the future. *Proc. Natl. Acad. Sci. U.S.A.* 104, 642–647

86 Addis, D.R. *et al.* (2011) Hippocampal contributions to the episodic simulation of specific and general future events. *Hippocampus* 21, 1045–1052

87 Mason, M.F. *et al.* (2007) Wandering minds: the default network and stimulus-independent thought. *Science* 315, 393–395

88 Ostby, Y. *et al.* (2012) Mental time travel and default-mode network functional connectivity in the developing brain. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16800–16804

89 Friston, K. *et al.* (2011) Action understanding and active inference. *Biol. Cybern.* 104, 137–160

90 Yamashita, Y. and Tani, J. (2008) Emergence of functional hierarchy in a multiple timescale neural network model: a humanoid robot experiment. *PLoS Comput. Biol.* 4, e1000220

91 Rabinovich, M.I. *et al.* (2008) Transient cognitive dynamics, metastability, and decision making. *PLoS Comput. Biol.* 4, e1000072

92 Murphy, K. (2012) *Machine learning: a probabilistic perspective*, MIT Press

93 Bendor, D. and Wilson, M.A. (2012) Biasing the content of hippocampal replay during sleep. *Nat. Neurosci.* 15, 1439–1444

94 Keramati, M. *et al.* (2011) Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol.* 7, e1002055

95 Gold, J.I. and Shadlen, M.N. (2007) The neural basis of decision making. *Annu. Rev. Neurosci.* 30, 535–574

96 Nakahara, H. and Hikosaka, O. (2012) Learning to represent reward structure: a key to adapting to complex environments. *Neurosci. Res.* 74, 177–183

97 Johnson, A. and Redish, A.D. (2005) Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Netw.* 18, 1163–1171

98 Connor, P.C. *et al.* (2013) An elemental model of retrospective revaluation without within-compound associations. *Learn. Behav.* 42, 22–38

99 Kudrimoti, H.S. *et al.* (1999) Reactivation of hippocampal cell assemblies: effects of behavioral state, experience, and EEG dynamics. *J. Neurosci.* 19, 4090–4101

100 Pennartz, C.M.A. *et al.* (2004) The ventral striatum in off-line processing: ensemble reactivation during sleep and modulation by hippocampal ripples. *J. Neurosci.* 24, 6446–6456

101 Buzsáki, G. (1986) Hippocampal sharp waves: their origin and significance. *Brain Res.* 398, 242–252

102 Euston, D.R. *et al.* (2007) Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science* 318, 1147–1150

103 Skaggs, W.E. and McNaughton, B.L. (1996) Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* 271, 1870–1873

104 Nádasdy, Z. *et al.* (1999) Replay and time compression of recurring spike sequences in the hippocampus. *J. Neurosci.* 19, 9497–9507

105 Abbott, L.F. and Nelson, S.B. (2000) Synaptic plasticity: taming the beast. *Nat. Neurosci.* 3 (Suppl), 1178–1183

106 Markram, H. *et al.* (1997) Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* 275, 213–215

107 Lee, A.K. and Wilson, M.A. (2002) Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron* 36, 1183–1194

108 Girardeau, G. *et al.* (2009) Selective suppression of hippocampal ripples impairs spatial memory. *Nat. Neurosci.* 12, 1222–1223

109 Ego-Stengel, V. and Wilson, M.A. (2010) Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus* 20, 1–10

110 Qin, Y.L. *et al.* (1997) Memory reprocessing in corticocortical and hippocampocortical neuronal ensembles. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 352, 1525–1533

111 Ji, D. and Wilson, M.A. (2007) Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nat. Neurosci.* 10, 100–107

112 Karlsson, M.P. and Frank, L.M. (2009) Awake replay of remote experiences in the hippocampus. *Nat. Neurosci.* 12, 913–918

113 Dragoi, G. and Tonegawa, S. (2013) Distinct preplay of multiple novel spatial experiences in the rat. *Proc. Natl. Acad. Sci. U.S.A.* 110, 9100–9105

114 Kappen, H.J. *et al.* (2009) Optimal control as a graphical model inference problem. *Machine Learn.* 87, 159–182

115 Ortega, P.A. and Braun, D.A. (2013) Thermodynamics as a theory of decision-making with information-processing costs. *Proc. R. Soc.* 469, 20120683

116 Kepecs, A. *et al.* (2008) Neural correlates, computation and behavioural impact of decision confidence. *Nature* 455, 227–231

117 Dayan, P. (2012) Twenty-five lessons from computational neuromodulation. *Neuron* 76, 240–256

118 Schultz, W. *et al.* (2008) Explicit neural signals reflecting reward uncertainty. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 363, 3801–3811

119 Cabral, H.O. *et al.* (2014) Oscillatory dynamics and place field maps reflect hippocampal ensemble processing of sequence and place memory under NMDA receptor control. *Neuron* 81, 402–415

120 Day, J.J. *et al.* (2007) Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat. Neurosci.* 10, 1020–1028

121 Liljeholm, M. and O'Doherty, J.P. (2012) contributions of the striatum to learning, motivation, and performance: An associative account. *Trends Cogn. Sci.* 16, 1364–6613

122 Everitt, B.J. and Robbins, T.W. (2005) Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat. Neurosci.* 8, 1481–1489

123 de Borchgrave, R. *et al.* (2002) Effects of cytotoxic nucleus accumbens lesions on instrumental conditioning in rats. *Exp. Brain Res.* 144, 50–68

124 Miller, J.F. *et al.* (2013) Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. *Science* 342, 1111–1114