Formalising planning and information search in naturalistic decision-making

Hunt Lt, N. Daw, P. Kaanders, M A Maclaver, U. Mugan, Emmanuel Procyk, A. Redish, E. Russo, Jacqueline Scholl, K. Stachenfeld, et al.

Formalising planning and information search in naturalistic decision-making

Hunt LT[1*], Daw ND[2], Kaanders P[3], MacIver MA[4], Mugan U[4], Procyk E[5], Redish AD[6], Russo E[7], Scholl J[3], Stachenfeld K[8], Wilson CRE[5], Kolling N[1*]

1. Wellcome Centre for Integrative Neuroimaging, Department of Psychiatry, University of Oxford, UK
2. Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton NJ, USA
3. Wellcome Centre for Integrative Neuroimaging, Department of Experimental Psychology, University of Oxford, UK
4. Center for Robotics and Biosystems, Department of Neurobiology, Department of Biomedical Engineering, Department of Mechanical Engineering, Northwestern University, Evanston IL, USA
5. Univ Lyon, Université Claude Bernard Lyon 1, Inserm, Stem Cell and Brain Research Institute U1208, 69500 Bron, France
6. Department of Neuroscience, University of Minnesota, Minneapolis MN USA
7. Dept. of Theoretical Neuroscience, Central Institute of Mental Health, 68159 Mannheim, Germany; Dept. of Psychiatry and Psychotherapy, University Medical Center, Johannes Gutenberg University, 55131 Mainz
8. Google DeepMind, London, UK

* correspondence to: laurence.hunt@psych.ox.ac.uk, nils.kolling@psych.ox.ac.uk

Competing interests: The authors declare no completing interests.

Decisions made by mammals and birds are often temporally extended. They require planning and sampling of decision-relevant information. Our understanding of such decision making remains in its infancy compared to simpler, forced choice paradigms. However, recent advances in algorithms supporting planning and information search provide a lens through which we can explain neural and behavioural data in these tasks. We review these advances to obtain a clearer understanding for why planning and curiosity originated in certain species but not others; how activity in the medial temporal lobe, prefrontal and cingulate cortices may support these behaviours; and how planning and information search may complement each other as means to improve future action selection.

Decisions in natural environments are temporally extended and sequential. In many species, they involve planning, information search, and choice between many alternatives. They may require action selection to unfold over long timescales. They can be characterised by periods of deliberation and information sampling, where the agent simulates the future consequences of its actions before committing to a final choice.

This contrasts with much decision-making research in neuroscience to date. Many decision-making paradigms focus around repeated choices between a limited number of options simultaneously presented to the agent. Adopting this reductive viewpoint has been highly fruitful – it has meant that formal algorithms borrowed from other fields can be applied when interpreting behavioural and neural data. For example, algorithms borrowed from signal detection theory are applied to interpret sensory detection tasks, such as 2-alternative forced choice paradigms[1]. Algorithms from model-free reinforcement learning[2], or economics[3], are applied to interpret reward-guided decision tasks. Algorithms from foraging theory[4] are used to interpret decisions about whether to stay or depart from a currently favoured patch location.

In this Perspective, we argue that the recent development of novel algorithms and frameworks allows us to move beyond reductive paradigms, and progress towards studying decision making in naturalistic, temporally extended environments. This progress creates challenges for the field. Which model organisms can be used to study naturalistic choices, and how might their cognitive abilities be compared to humans? How do we design paradigms that are more naturalistic but remain experimentally tractable? What is the behavioural and neurophysiological evidence that animals are planning or making use of sampled information?

We seek to emphasise an important relationship between planning and information search during naturalistic decision making. Both are about not pursuing immediate reward, but instead improving selection of future actions. While physically searching or sampling information is an overt action, planning relies upon mental simulation and is typically covert. Planning is thus a form of *internal* information search, over past experiences. Cognitive processes leading to overt actions are easier to measure experimentally. We argue that by understanding the neural basis of tasks requiring overt information search, we may gain insight into neural mechanisms supporting covert planning.

Why do (certain) animals plan?

We first need to ask: why plan at all? Current understanding of plan-based control regards such action choices as depending upon the explicit consideration of possible prospective future courses of actions and consequent outcomes. Conversely, there is no explicit consideration of action outcome under habit-based control[5-7]. Planning, therefore, can create new information because it is compositional. It concatenates bits of knowledge about actions' short-term consequences to work out their long-term values. By contrast, habit-based action choices are sculpted by prior experience alone without such inference. Whereas habit-based action selection is automatic, fast, and inflexible, plan-based action selection requires deliberation, which allows actions to adapt to changing environmental contingencies.

Evolutionary conditions selecting for planning. Habit-based action selection appears to be universal amongst vertebrates, both terrestrial and aquatic. In contrast, behavioural and neural evidence for plan-based action selection seems to only exist for mammals and birds[8-10] and appears either absent or ambiguous for reptiles, amphibians[11,12], and fish[13].

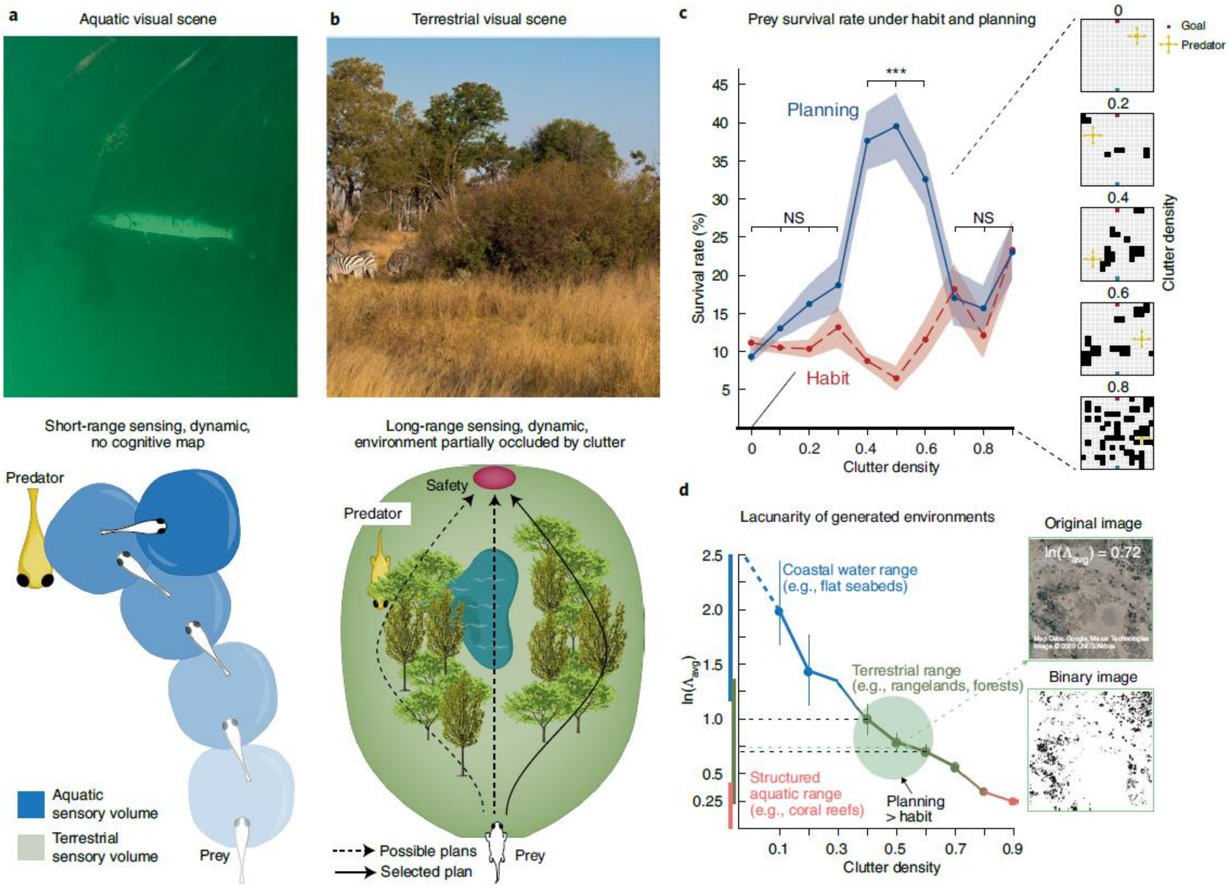

Figure 1. Aquatic versus aerial visual scenes, and how the corresponding habitats affect the utility of habit- and plan-based action selection during dynamic visually-guided behaviour. (a) *Example of an aquatic visual scene[150], from Current Biology, Vol. 27 Issue 14, Dan E. Nilsson, Evolution: An Irresistibly Clear View of Land, R716, Copyright (2017), with permission from Elsevier.. In such situations, typical of aquatic environments, visual range is limited and so predator-prey interactions occur at close quarters, requiring rapid and simple responses facilitated by a habit-based system.* ***(b)*** *Example of a terrestrial visual scene ("Zebra and giraffe" by Caty T, used under CC BY 2.0 / Cropped from original). Computational work[16] suggests that these scenarios confer a selective benefit (not present in aquatic habitats) to planning long action sequences, by imagining multiple possible futures (solid/dashed black arrows) and selection of the option with higher expected return (solid black arrow).* ***(c)*** *The computational work idealized predator-prey interactions as occurring within a 'grid world' environment (column on right; prey blue, predator yellow) where the density of occlusions was varied. Prey had to either use habit- or plan-based action selection to get to the safety (red square) while being pursued by the predator. The plot shows survival rate versus clutter density across random predator locations, under plan-based (blue solid) and habit-based action selection (red dashed). Line indicates mean ±s.e.m. across randomly generated environments (n.s. = not significant, p > 0.05, *** p < 0.001. Data from [16].* ***(d)*** *To relate clutter densities in the artificial worlds to those found in the real world, Mugan and MacIver[16] used lacunarity, a measure commonly used by ecologists to quantify spatial heterogeneity of gaps that arise from (for example) spatially discontinuous biogenic structure. The line plot shows the mean natural log of average lacunarity and the interquartile range of environments with a predetermined clutter level. Coastal, terrestrial, and structured aquatic environments can be partitioned based on previously published lacunarity value (for a full range of lacunarities across different environments, see [16]). The green circle highlights a zone of lacunarity where planning outstrips habit (based on* ***(c)****). Insert shows an example image from the Okavongo Delta in Botswana (≈800 m x 800 m, from Google Earth), considered a modern analogue of the habitats that early hominins lived within after branching from chimpanzees[24]. Its average lacunarity ($\ln(\Lambda_{avg})$) is 0.72. Images in* ***(a)/(b)*** *from ref. [150]; all other panels adapted from ref. [16]*

Recent computational work suggests that the increase in visual range[14] and environmental complexity[15] that accompanied the shift from life in water to life on land may have been a critical step in the evolution of planning[16] (**Fig. 1**). In particular, plan-based action selection may be advantaged in complex dynamic tasks when the animal has enough time and sufficiently precise updates—such as through long range vision—to forward simulate. Therefore, long-range imaging systems (i.e. terrestrial vision, but also mammalian aquatic echolocation) may be crucial in advantaging plan-based control in complex environments, due to their ability to detect the structure of a complex, cluttered environment with high temporal and spatial resolution. In such cases, the simultaneous apprehension of distal landmark information and other dynamic agents, be they prey or predator, allows planning to take place over the changing sensorium. When visual range is reduced, such as in nocturnal vision, plan-based control may only exist for stable environments over a previously established cognitive map. Thus, near-field detection of landmarks may be used to calibrate an allocentric map and planning used only initially to devise new paths through this stable environment.

The scenarios of short- and long-range dynamic environments shown in **Fig. 1a/b** drive the following hypothesis: plan-based action selection is evolutionarily selected for when the number of action selection possibilities with differing outcome values is so large, dynamic, and uncertain that habit-based action selection fails to be adaptive (**Fig. 1c**). Evolutionarily this scenario greeted the first vertebrates to live on land over 300 million years ago. The increase in both visual range[14] and environmental complexity[15] due to the change in viewing medium and habitat facilitated the observance of the large variety of uncertain action-outcome values over an extended period of time in predator-prey encounters, thus advantaging planning.

Variation in planning across terrestrial species. Within terrestrial species, there is also marked variation in planning complexity. Many mammalian species learn the latent structure of their environment and deploy this flexibly to select new behaviours. Original support for the idea that rodents learn a cognitive map of their environment came from studies by Tolman[17], in which rats immediately deployed the previously learnt structure of the environment in order to travel to reward-baited locations. Modern-day tests of similar behaviours show that such cognitive maps underlie hippocampal-dependent single-trial learning of new associations[18]. There is also evidence for planning in certain birds, exemplified by food caching behaviours in scrubjays[19] and tool use in New Caledonian Crows[20].

However, these tests of planning remain simplified compared to the flexible higher-order sequential planned behaviours observable in humans and other primates[21]. Between-species variation in primate brain size may partly be explained by the complexity of foraging environments over which different behaviours must be planned[16,22]. It remains unclear whether there are good analogues even in non-human primates of the hierarchically organised plan-based action selection[23] that underlies much of human behaviour. Work on the type of habitats which maximize the advantage of planning shows that a patchy mix of open grassland and closed forested zones confers the greatest advantage[16] (**Fig. 1d**). This appears to be the type of habitat that hominins invaded after splitting from forest-dwelling chimpanzees[24], and could, in combination with long range vision, be a contributing factor to hominid exceptionalism in planning[16]. In addition, the development of large social groups in primates (particularly hominins) demand sophisticated planning of multi-agent interactions[25]; social interactions not only require updating the likely behaviour of other agents, but also demand iterative inferences[26]. The near quadrupling in brain volume of early hominins compared to chimpanzees may relate to the high computational burden of planning due to both their foraging and social environment.

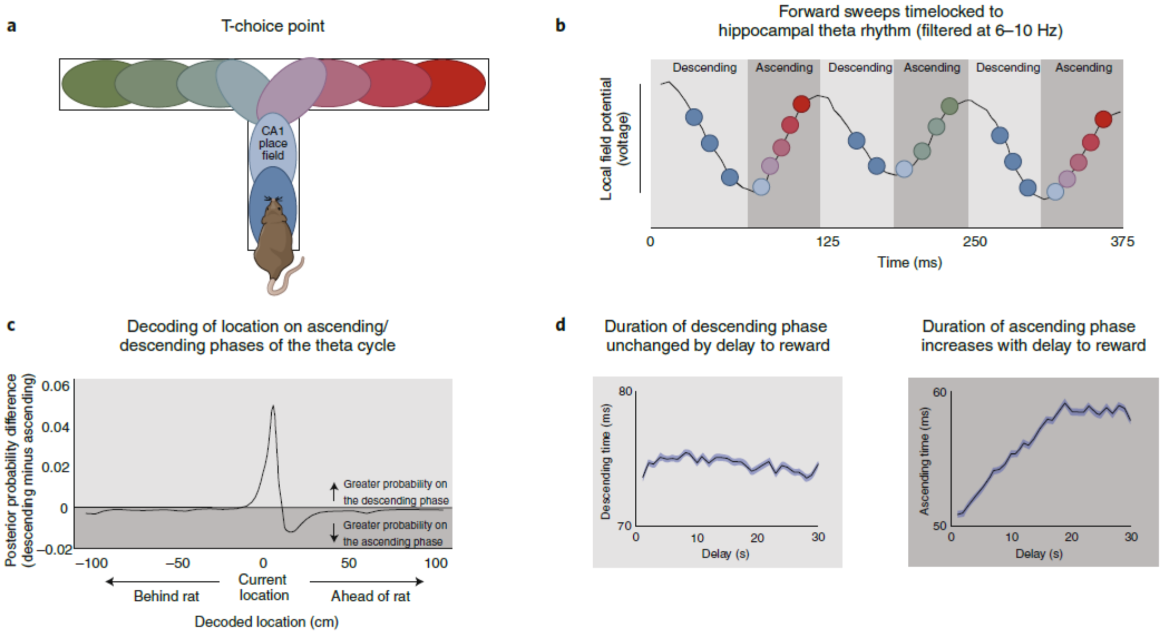

Figure 2. As rats approach a choice point, a theta-locked hippocampal representation sweeps ahead of the rat towards potential goals. ***(a)*** *A rat approaches a T-choice point. Each oval indicates the place field of a place cell in CA1 of the hippocampus.* ***(b)*** *Place cells fire at specific phases of the hippocampal theta rhythm, allowing different spatial locations to be decoded from neural activity (coloured circles) leading to a sweep forward ahead of the rat. The descending phase of the oscillation is dominated by cells with place fields centered at the rat's current location, where the ascending phase is dominated by cells with place fields ahead of the rat, sweeping towards different potential goals on individual theta cycles[36,37].* ***(c)*** *Bayesian decoding applied separately to the descending and ascending phases of the theta cycle finds more decoding of current location during the descending phase, but more decoding of locations ahead of the rat during the descending phase.* ***(d)*** *On a task in which the goal is delayed in time, the duration of the descending phase of the theta cycle is unchanged by the distance to the goal, but the ascending phase increases proportionally. Data for panels* ***(c)/(d)*** *adapted from ref. [10].*

Parallels between planning and information search. Intriguingly, between-species variation in planning sophistication can be related to between-species variation in curiosity. Curiosity can be defined as the natural intrinsic motivation and tendency to proactively explore the environment and gather information about its structure[27]. Primates in particular, and carnivores in general, have a biased tendency for curiosity and exploration compared to other species like reptiles that might be unmoved by new objects or neophobic[28]. Humans and non-human primates have an extended juvenile period, and playful curiosity during this period gives rise to increased brain growth and behavioural flexibility[29]. Curiosity-driven information search can also take advantage of existing cognitive capabilities. New Caledonian Crows, for example, use tools when exploring novel objects, suggesting they can generalise tool use from food retrieval to non-foraging activities[30].

This parallel between planning and curiosity reinforces the viewpoint that the primary goal of information sampling is to build up knowledge of the structure of the environment. Structural knowledge acquired during information sampling can then be flexibly deployed when planning actions online in new environments, or when reward locations or motivations change. Recent studies in cognitive science have made this link explicit, using information sampling behaviour to arbitrate between which planning strategies participants are using in a multistep decision task[31].

Plan-based action selection and curiosity may have given rise to evolutionary advantages. To study the algorithmic implementation of these behaviours, however, it becomes necessary to develop a formal framework against which they can be quantified, and their neural representations measured.

Formalising planning

Formally, value-based planning (e.g. tree search by a chess computer to find the best move) corresponds to computing the long-run utility of different candidate courses of action, in expectation over the possible resulting series of future situations and moves. In Reinforcement Learning (RL) algorithms, this type of evaluation is known as "model-based" planning.

Model-based planning relies on an "internal model" or representation of the task contingencies to forecast utility. Such a model can be used, in effect, to perform mental simulation to forecast the states and values likely to follow candidate action trajectories. This is contrasted with "model-free" trial and error, which is used to describe habit-based action selection[6]. This formalism has provided a foundation for reasoning about planning in psychology and neuroscience[5]: inspiring new tasks and predicting whether and when organisms are planning in classic tasks[17,32], and grounding the search for neural mechanisms that implement specific forms of planning[33]. It has also offered a formal perspective on how the brain decides when to plan, versus acting without further deliberation, by defining under what circumstances additional planning is likely to be particularly effective in improving one's choices[5,34].

Mental simulation in the hippocampal formation. There are many different variants of model-based planning, which share the central feature of using a cognitive map of the environment to simulate future trajectories, but differ in the pattern by which this occurs. Perhaps the most straightforward case searches through possible future paths from the current situation, using these sweeps to evaluate different courses of action. Neurophysiologically, the hippocampal formation is a likely candidate for the encoding of such a cognitive map[35], and this has guided the search for neural correlates of 'trajectory sweeping' during planning.

In spatial navigation in rodents, for example, place cell activity recorded during active exploration of the environment reflects the animal's current location. However, it also transiently represents other locations distal from the animal, including – suggestively – sequentially traversing paths in front of the animal. These nonlocal "sweeps" have been hypothesized to reflect episodes of explicit mental simulation through potential trajectories[7,33,36,37]. Notably, these events represent individual paths rather than a wavefront of future locations in parallel. Furthermore, consideration of each path takes time, and often occurs when the animal's locomotion is stopped. Thus deliberation, much like information search in the physical world, has an opportunity cost.

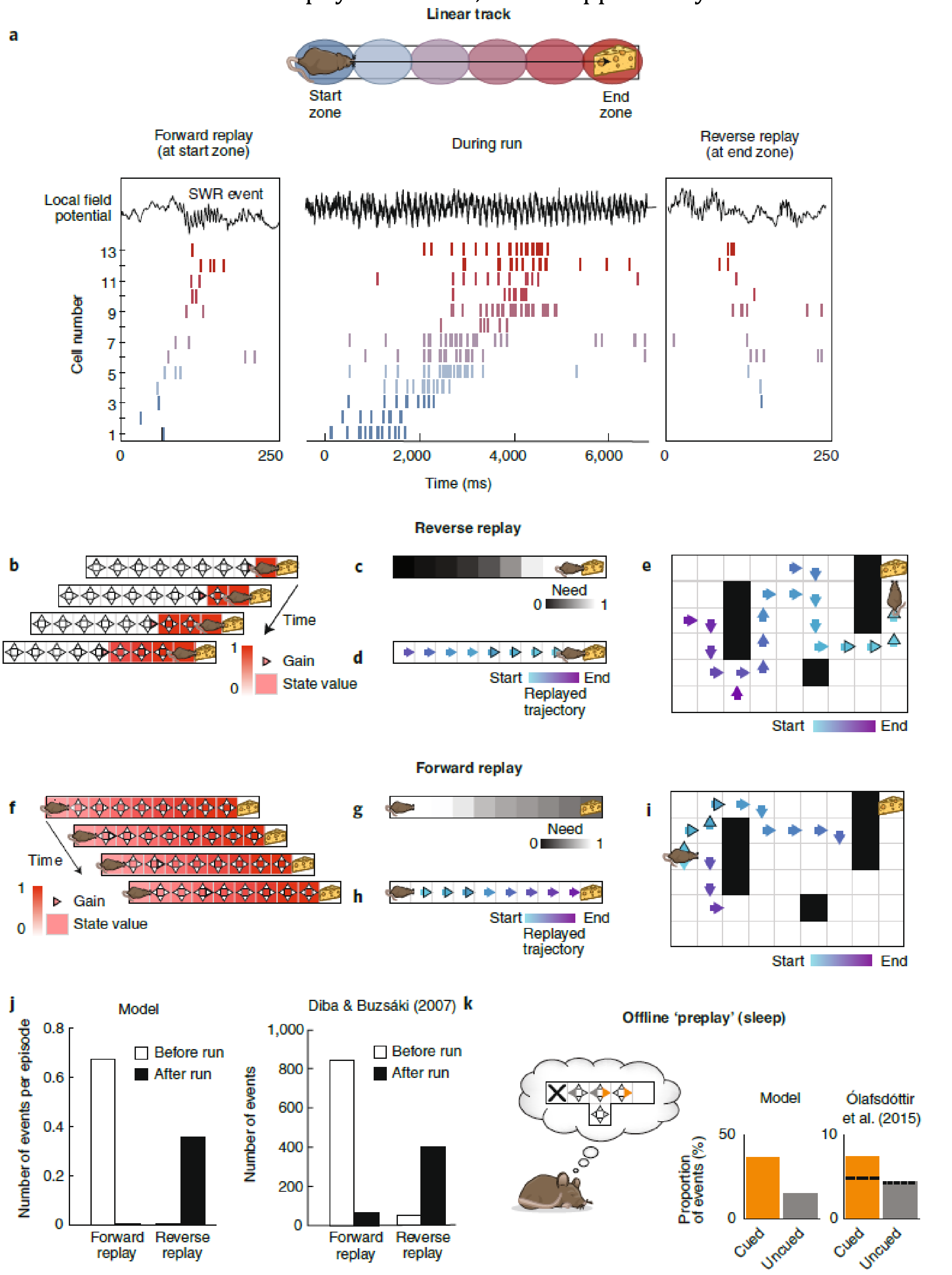

*Figure 3. **A normative model-based planning account of replay events, observed in hippocampal place cells and in simulations of spatial navigation tasks. (a)** Spike trains of rat hippocampal place cells before, during, and after running down a linear track to obtain a reward. Forward and reverse replay are observed before and after the lap, respectively, during sharp-wave ripple (SWR) events[38]. **(b-k)** Simulations of spatial navigation tasks, in which the agent evaluates memories of locations, called 'backups', preferentially by considering 'need' (how soon the location is likely to be encountered again) and 'gain' (how much behaviour can be improved from propagating new information to preceding locations). Simulated replay produces extended trajectories in forward and reverse directions[33]. **(b-d)** Gain term, need term and resulting trajectory for reverse*

replay on a linear track. There is a separate gain term ***(b)*** *for each action in a state (small triangles). If a state-action pair leads to an unexpectedly valuable next state, performing a backup of this state-action pair has high gain, as it will change the animal's behaviour in that state. Once this backup is performed, the preceding action (highlighted triangle) will now have high gain, and is likely to be backed up next. Multiple iterations of this process can lead to reverse replay.* ***(e)*** *Reverse replay can also be simulated in more naturalistic 2D open fields, tracking all the way from the goal to the starting location.* ***(f-h)*** *Gain term, need term and resulting trajectory for forward replay on a linear track. The need term* ***(g)****, derived from the successor representation of the agent (see* ***Fig. 4****), reflects locations likely to be visited in the future. If need term differences are larger than gain term differences, this term dominates in driving the replayed trajectory. Here, this tends to lead to forward replay.* ***(i)*** *Simulated forward replay events also arise in 2D open fields, sometimes exploring novel paths towards a goal.* ***(j)*** *The model predicts the balance between forward/reverse replay events observed before/after running down a linear track[33,38].* ***(k)*** *When an agent is simulated in an offline setting after exploring a T-maze and observing that rewards have been placed in the right arm, more backups of actions leading to the right arm are performed[33]. The same has been observed in rodent recordings during sleep[60]. Data for panels* ***(a)/(j)*** *adapted from ref. [38]; data for panel* ***(k)*** *adapted from ref. [60]; all other panels adapted from ref. [33].*

Two distinct types of nonlocal sweeps have captured attention: one involving isolated trajectories linked to a high-frequency event in the local field potential known as a sharp wave ripple[38,39], and the other linked to theta oscillations, involving repeated cycles of forward excursions that sometimes alternate between multiple potential paths[36,37] (see **Figs. 2** and **3**). Both types of events have been argued to be candidates for model-based evaluation by mental simulation, though these hypotheses are not mutually exclusive.

Theta cycling and mental simulation. Beyond the fact that non-local sweeps traverse relevant candidate paths, a number of additional observations surrounding theta cycling suggest their involvement in planning. First, these sequences sweep serially to the goals ahead of the animal during the ascending phase of the theta cycle[10,40], and coincide with prefrontal representations of goals[41] (**Fig. 2**). Second, journeys on which these non-local representations sweep forward to goals often include an overt external behaviour, known as 'vicarious trial and error' (VTE), which is also suggestive of deliberation[7]. During VTE, rats or mice pause at a choice point and orient back and forth along potential paths[7,17]. Advances in experimental task design have helped to isolate these behaviours linked to planning, and capture the degree to which subjects use plan-based versus habitual controllers when selecting between courses of action.

Taking VTE to indicate planning processes, VTE occurs when animals know the structure of the world (have a cognitive map), but don't know what to do on that map. VTE disappears as animals automate behaviors within a stable world[42,43] and reappears when reward contingencies change[44,45]. On tasks in which animals show phases of decision strategies, VTE occurs when agents need to use flexible decision strategies and disappears as behaviour automates (see [7] for review). This indicates that the presence or absence of VTE matches with the conditions that normatively favour model-based or model-free RL respectively[5]. During VTE, neural signals consistent with evaluation are found in the nucleus accumbens core[46].

Interestingly, disruption of hippocampus or medial temporal lobe can (in certain circumstances) increase rather than decrease VTE behaviour[47-49], suggesting that VTE may be initiated elsewhere. One candidate is prelimbic cortex, where temporary inactivation diminishes VTE in rats and also impairs hippocampal theta sequences[10]. This finding provides an intriguing link to studies of the role of monkey dorsal anterior cingulate cortex (dACC) in information sampling. Neural activity in dACC shifts between exploration and choice repetition occurring ahead of reward delivery, triggered after the accumulation of sufficient information to predict and plan the correct future solution to a problem[50]. This region also contains neural ensembles that are engaged whenever the animal explicitly decides to check on the current likelihood of receiving a large bonus reward[51] (see below).

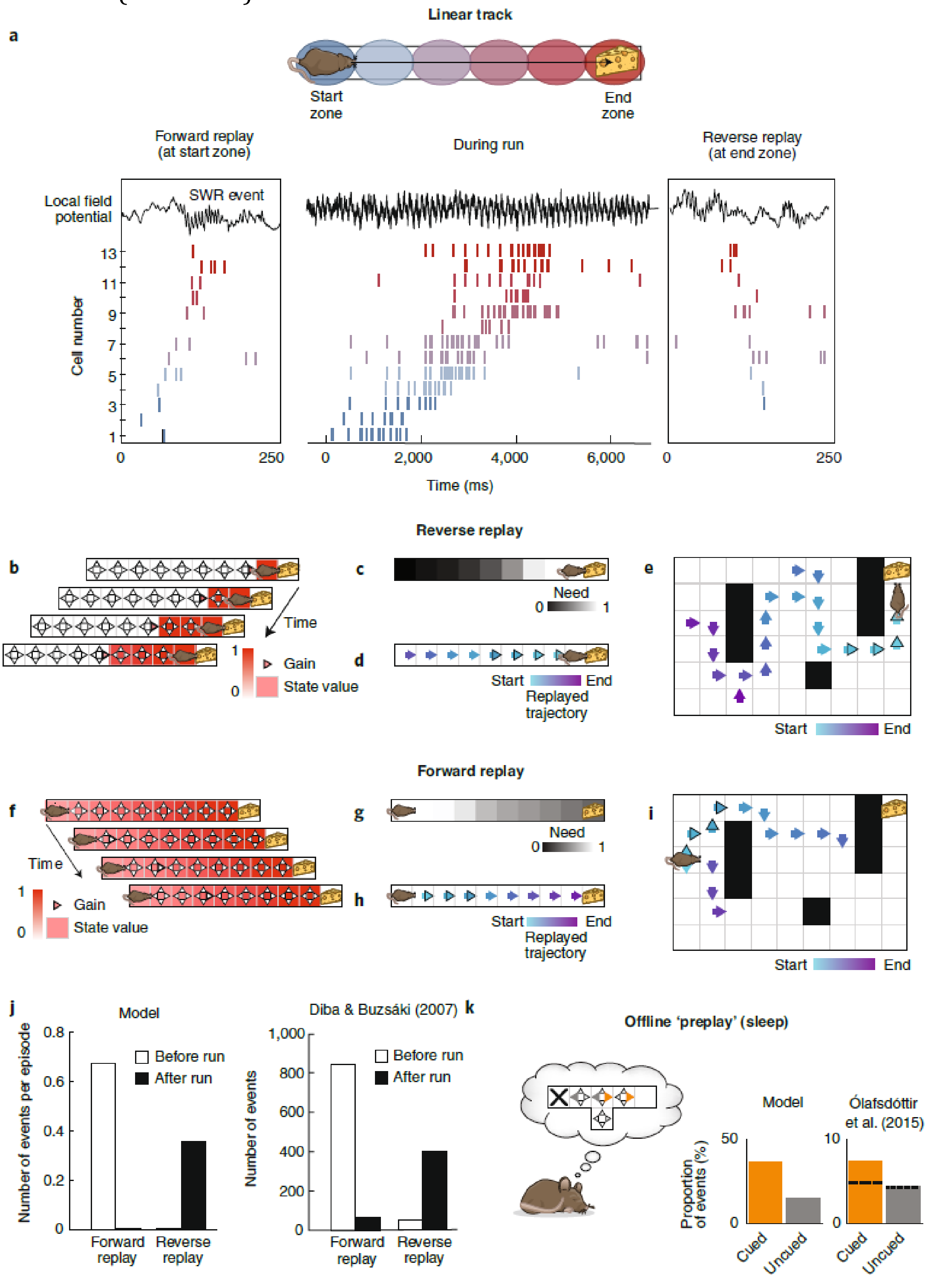

Figure 4. A normative model-based planning account of replay events, observed in hippocampal place cells and in simulations of spatial navigation tasks. (a) *Spike trains of rat hippocampal place cells before, during, and after running down a linear track to obtain a reward. Forward and reverse replay are observed*

before and after the lap, respectively, during sharp-wave ripple (SWR) events[38]. ***(b-k)*** *Simulations of spatial navigation tasks, in which the agent evaluates memories of locations, called 'backups', preferentially by considering 'need' (how soon the location is likely to be encountered again) and 'gain' (how much behaviour can be improved from propagating new information to preceding locations). Simulated replay produces extended trajectories in forward and reverse directions[33].* ***(b-d)*** *Gain term, need term and resulting trajectory for reverse replay on a linear track. There is a separate gain term* ***(b)*** *for each action in a state (small triangles). If a state-action pair leads to an unexpectedly valuable next state, performing a backup of this state-action pair has high gain, as it will change the animal's behaviour in that state. Once this backup is performed, the preceding action (highlighted triangle) will now have high gain, and is likely to be backed up next. Multiple iterations of this process can lead to reverse replay.* ***(e)*** *Reverse replay can also be simulated in more naturalistic 2D open fields, tracking all the way from the goal to the starting location.* ***(f-h)*** *Gain term, need term and resulting trajectory for forward replay on a linear track. The need term* ***(g)****, derived from the successor representation of the agent (see* ***Fig. 4****), reflects locations likely to be visited in the future. If need term differences are larger than gain term differences, this term dominates in driving the replayed trajectory. Here, this tends to lead to forward replay.* ***(i)*** *Simulated forward replay events also arise in 2D open fields, sometimes exploring novel paths towards a goal.* ***(j)*** *The model predicts the balance between forward/reverse replay events observed before/after running down a linear track[33,38].* ***(k)*** *When an agent is simulated in an offline setting after exploring a T-maze and observing that rewards have been placed in the right arm, more backups of actions leading to the right arm are performed[33]. The same has been observed in rodent recordings during sleep[60]. Data for panels* ***(a)/(j)*** *adapted from ref. [38]; data for panel* ***(k)*** *adapted from ref. [60]; all other panels adapted from ref. [33].*

Sharp-wave ripples and mental simulation. Nonlocal trajectory events during high frequency sharp-wave ripples (**Fig. 3a**) also have a number of characteristics consistent with planning. These events also occur when animals pause during ongoing task behaviour (particularly at reward sites[52,53]); they can produce novel paths[54]; they tend to originate at the animal's current location and predict its future path[55]; their characteristics change with time in a fashion consistent with changing need for model-based evaluation; and disrupting them causally affects trial-and-error task acquisition[56]. Interestingly, disrupting sharp-waves increases VTE, suggesting that sharp-wave-based and theta-based planning processes may be counterbalanced[53].

A key additional feature of these events is that the most obviously planning-relevant events – paths in front of the animal during task behaviour – are only one special case of a broader set of nonlocal trajectories, which occur in different circumstances and include paths that rewind behind the animal often following reward[57,58]; and wholly nonlocal events during quiet rest or sleep[54,59,60].

Recent computational modeling work[33] (**Fig. 3**) has aimed to explain these observations in terms of a normative analysis of model-based planning, considering not just when it is advantageous to plan, but which trajectory is most useful to consider next. Formally, this means prioritizing locations that will cause a substantial change in the agent's future behaviour (how much the agent stands to *gain* from performing the simulation). One should also prioritise locations that the animal is particularly likely to visit in the future (how much *need* there is to perform such a simulation). The expected value of a particular trajectory is then calculated as the product of these two terms (e.g **Fig. 3b-d, f-h**). Importantly, while this analysis captures the characteristics of forward sweeps during task behaviour (**Fig. 3f-i**), it also explains backward replay behind the animal when a reward is received (**Fig. 3b-e**), and trajectories that tend to occur during sleep (**Fig. 3k**), as a form of offline 'pre-planning' for when these situations are next encountered[61].

Human neuroimaging experiments also suggest that putative behavioural signatures of model-based planning are associated with forward or backward neural reinstatement at various time points[9,62-64]. Human replay appears to occur in the sequence to be used in future behaviour rather than the experienced sequence[65], and is particularly pronounced for experiences that will be of greater future benefit[66] as predicted by the prioritization framework[33].

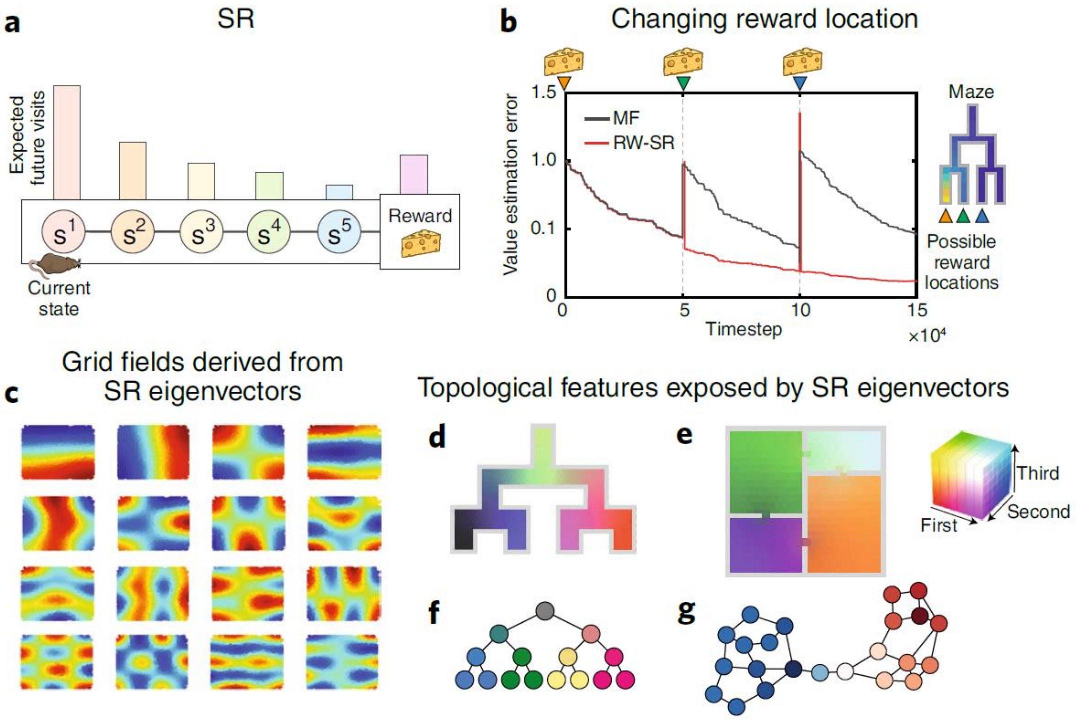

Figure 5. The successor representation (SR) allows for rapid revaluation, and extraction of components that identify key components of state space structure. ***(a)*** *Successor Representation (SR) at state s^1 for a policy that moves an agent toward the reward box (from ref.* [82]*). The SR encodes the (discounted) expected future visits to all states.* ***(b)*** *Comparison of model-free (MF) learning and Rescorla Wagner (RW) SR-based learning of a value function under changing reward locations (given a random walk policy). Following a change in the reward location, SR learning is only temporarily set back while the agent learns the new reward location, whereas MF learning must resume from scratch. The error is reported as the summed absolute difference between estimated and ground truth value at each state divided by the maximum ground truth value to normalize*[82]*.* ***(c)*** *First 16 eigenvectors for a rectangular graph consisting of 1600 nodes randomly placed in a rectangle, with edges weighted according to the diffusion distance between states*[82]*, are reminiscent of grid fields recorded in entorhinal cortex.* ***(d-g)*** *Examples of how topological features of an environment are exposed by SR eigenvectors. In* ***(d-f)****, each state is colored such that the first 3 eigenvectors set the RGB (see colour cube). This shows how states are differentiated by the first few eigenvectors, and how they expose bottlenecks and decision points. In* ***(g)****, the first eigenvector is shown, revealing clusters in the graph structure. Panels* ***(a)*** *to* ***(c)*** *adapted from ref.* [82]*.*

Efficiently representing large state spaces. No matter how simulation is implemented, model-based planning suffers from a potentially exponential growth in computation time as planning becomes deeper, except in small-scale toy problems with a limited range of possible future outcomes or state space[67]. This is because of how the decision tree branches. If, for example, at every planning step there are 2 new possibilities, the total number of possible paths to consider grows at 2^n. We therefore need formalisms that account for tractable planning at scale.

Representation learning is a framework for improving the scalability of reinforcement learning. Essentially, representation learning involves learning to represent your current state so as to reduce the burden on the downstream RL algorithm, usually by representing its position relative to task structure[68-70]. By making state representations more efficient, model-free agents become more sensitive to task structure and therefore more flexible to changes in reward contingencies. Alternatively, the learned representation may feed into a model-based planner, in which case the representation implicitly organizes the search or planning occurring over it.

Recent studies in human cognitive science have shown that humans can exploit environmental structure in order to learn efficient representations in multi-armed bandit tasks[71,72] and guide exploration in large decision spaces[73]. This structure typically depends upon learning that certain options are correlated with one another. For example, if many options are presented, but options that are close in space tend to be similar to one another, then humans exploit this spatial relationship in their choices and searches[73]. More broadly, structure learning links to the older idea of a 'learning set', in which experience on a task allows faster learning of new problems on the same task[35,74]. In machine learning, a similar phenomenon has been termed meta-learning[75].

The neural basis of structure learning remains relatively underexplored. Disconnection lesions between frontal and temporal cortex impair use of a learning set, demonstrating the importance of interactions between these brain regions[76], as also shown by transection of the fornix (a white matter structure linking hippocampus and frontal cortex)[77]. More recently, human imaging studies have used representational similarity analysis between different RL states to identify entorhinal cortex[71] and orbitofrontal cortex[71,78] as key nodes for learning task structures.

Compressing information about future state occupancy. Neural representations of the animal's current state must not only be rich enough to support sophisticated planning behaviours, but also to render planning computationally tractable. One solution is to learn a "predictive representation" of states expected to occur over multiple steps into the future, meaning that states that predict similar futures are constrained to have similar representations[79,80]. If two states lead to similar outcomes, it is safe to assume that anything learnt about one state (such as its value) should apply to the other as well. This can simplify planning, since predictive representations incorporate statistics about multiple steps of future events directly into the current representation. This allows anticipation of future states without the need to iteratively construct them via mental simulation.

One example is the successor representation[79,81]. The successor representation of one's current state is a vector encoding the expected number of visits to each possible future (or successor) state (**Fig. 4a**). In addition to simplifying planning, this accelerates value learning following changes (**Fig. 4b**). In neuroscience, the idea of predictive representation has been applied to explain some features of hippocampal place fields[82], such as asymmetric growth in fields with traversals[83], although it does not explain the sweeps and sequences discussed earlier. It can also account for human and animal revaluation behaviour[84,85] and properties of dopaminergic learning signals[86]. We also suggest that it might be worth asking whether other neural systems, such as striatum (which develops representations with experience[87,88]) or prefrontal cortex (which shows hierarchical abstraction[89,90]) show these successor representation properties.

A related idea is that the state transition map of a task can be represented in a compressed form by summing periodic components of different frequencies, in particular low-spatial and low-temporal frequency ones that coarsely predict state occupancy far into the future. These components can be constructed by taking principal components of the transition matrix[91], or equivalently the successor representation matrix[82]. The lower frequency components produce compressed representations that can support faster learning[91] and improved exploration[92]. By capturing smoothed, coarse-grained trends of how states predict each other, they pull out key structural elements such as clusters, bottlenecks, and decision points (**Fig. 4d-g**). These periodic functions share some features of grid cells[82] (**Fig. 4c**), thereby falling into a family of models that suggest entorhinal cortex provides a mechanism for incorporating the spatiotemporal statistics of task structure into hippocampal learning and planning[93,94]. Recent work has explored the use of this type of representation to permit efficient linear approximations to full model-based planning[95].

Taken together, prediction and compression comprise two key learning principles. Prediction motivates encoding relevant information about the structure of the environment, and compression causes this information to be represented compactly to make learning about reward more efficient.

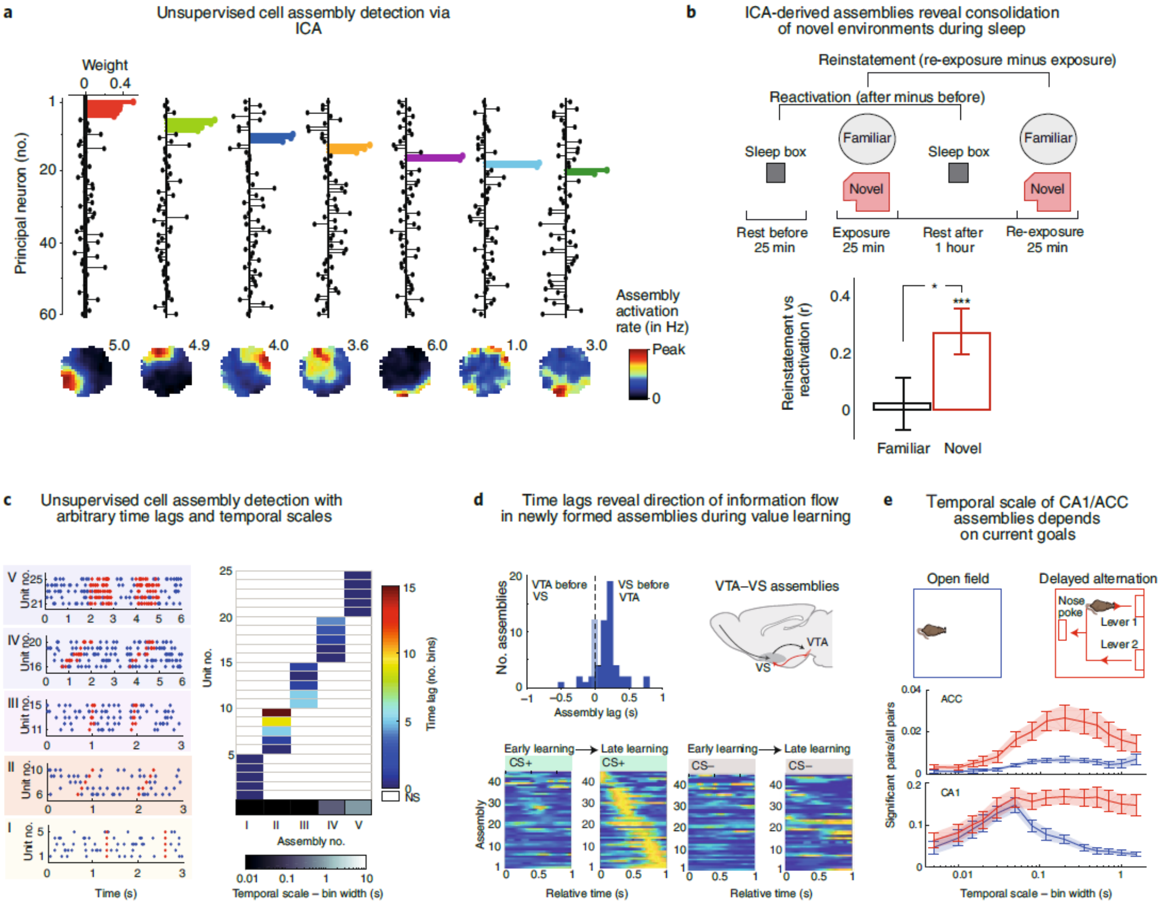

Figure 6. ***Unsupervised cell assembly detection to identify neural substrates of cognitive tasks. (a)*** *One approach to cell assembly detection identifies coincidently active populations of cells, via independent component analysis (ICA) of firing rate in 25ms bins[102]. Here, 7 cell assemblies are derived from 60 hippocampal CA1 principal neurons during exploration of a spatial arena. The derived cell assemblies show spatial tuning (bottom row).* ***(b)*** *After exposure to a novel spatial environment, greater 'reactivation' of the cell assemblies derived in* ***(a)*** *during sleep is correlated with greater 'reinstatement' of the same cell assembly pattern during subsequent re-exposure to the environment.* ***(c)*** *Another approach to cell assembly detection allows for detection of assemblies at arbitrary temporal scales (bin width of firing rate used), and arbitrary time lag in activation between different neurons.[96]* ***(d)*** *Top panels: distribution of timelags within detected cell assemblies between simultaneously recorded spiny projection neurons in ventral striatum (VS) and dopamine neurons in ventral*

tegmental area (VTA) during associative learning of value with a conditioned stimulus (CS+). VS neurons lead VTA neurons in recovered cell assemblies. Bottom panel: assemblies emerge with learning for the rewarded (CS+) but not unrewarded (CS-) stimulus. ***(e)*** *Cell assemblies in rat CA1 and anterior cingulate cortex (ACC) during open field exploration versus delayed alternation. In ACC, more significant assembly unit-pairs were found in the delayed alternation task across all temporal scales. In CA1, significantly more long-timescale cell assemblies were found during delayed alternation than during open field exploration (n.b. task differed slightly for CA1, requiring navigation through a figure-of-eight maze). Data for panels* ***(a)/(b)*** *from ref. [102]; panels* ***(c)/(e)*** *adapted from ref. [96]; panel* ***(d)*** *from ref. [104] (all panels used under CC-BY license).*

*Obstacles, and potential solutions, for measuring neural substrates of planning***.** The same reasons that make understanding planning so interesting also make it difficult to study. By definition, planning is internally generated and often covert. Place cell activity recorded during navigation allows decoding of planning events in spatial tasks (e.g. **Figs. 2/3)**, but it is less clear how to generalise this approach to non-spatial tasks, or to processes that occur over longer temporal scales.

Instead of anchoring the investigation to overt behavioural markers, a possible solution is to use unsupervised data mining to identify neural events of interest directly from spike train data. Techniques like cell assembly detection[96] and state space model estimation[97] uncover structures directly from spike train statistics without the need for any behavioural parametrization. Cell assembly detection is based on the assumption that assemblies relevant for a cognitive function generate recurring, albeit potentially noisy, stereotypical activity patterns. State space model estimation instead aims to capture the dynamics governing neural processes by fitting a set of differential equations on the experimental data.

Due to the combinatorial explosion of potential patterns to test, many existing cell assembly detection methods restrict their search to stereotypical activity profiles characterized by a specific lag configuration (synchronous[98,99] or sequential[100] unit activations) or temporal scale (single spike[98,100] or firing rate[99,101] coordination; see **Fig. 5a** for example). Such approaches have identified reactivation of cell assemblies during sleep, supporting the consolidation of learning novel spatial arenas[99,102] (**Fig. 5b**). Assembly-specific optogenetic silencing of these reactivation events impairs performance in approaching goal locations in a spatial navigation task[103], consistent with the role outlined above for replay during sleep as a substrate for planning future actions (**Fig. 3k**).

More recent techniques are now expanding the search to a wider set of testable pattern configurations[96,100,101] and timescales[96], treated as parameters to be inferred from the data (**Fig. 5c**). This approach has, for example, recently isolated the formation of interregional cell assemblies between dopaminergic midbrain and ventral striatum during value-based associative learning (**Fig. 5d**)[104]. In naturalistic planning tasks, a similar approach might identify events linking dopaminergic activity to hippocampal cell assembly activity subserving planning[105], although this remains to be tested. It is also possible to identify how the timescale of cell assemblies changes during goal-directed behaviour. For example, hippocampus and anterior cingulate cortex assembly temporal properties differ during passive exploration versus a delayed alternation task (**Fig. 5e**)[96].

Cognitive models of planning. So far, we have focused on different formal models of planning through well-defined state spaces or navigation through known structures such as physical mazes. However, human participants can also incorporate knowledge about their own future behavioural tendencies into their planning. There is evidence that humans might approximate the effects of increasing horizons[106] and use pre-emptive strategies to take into account their own future behavioural tendencies[107].

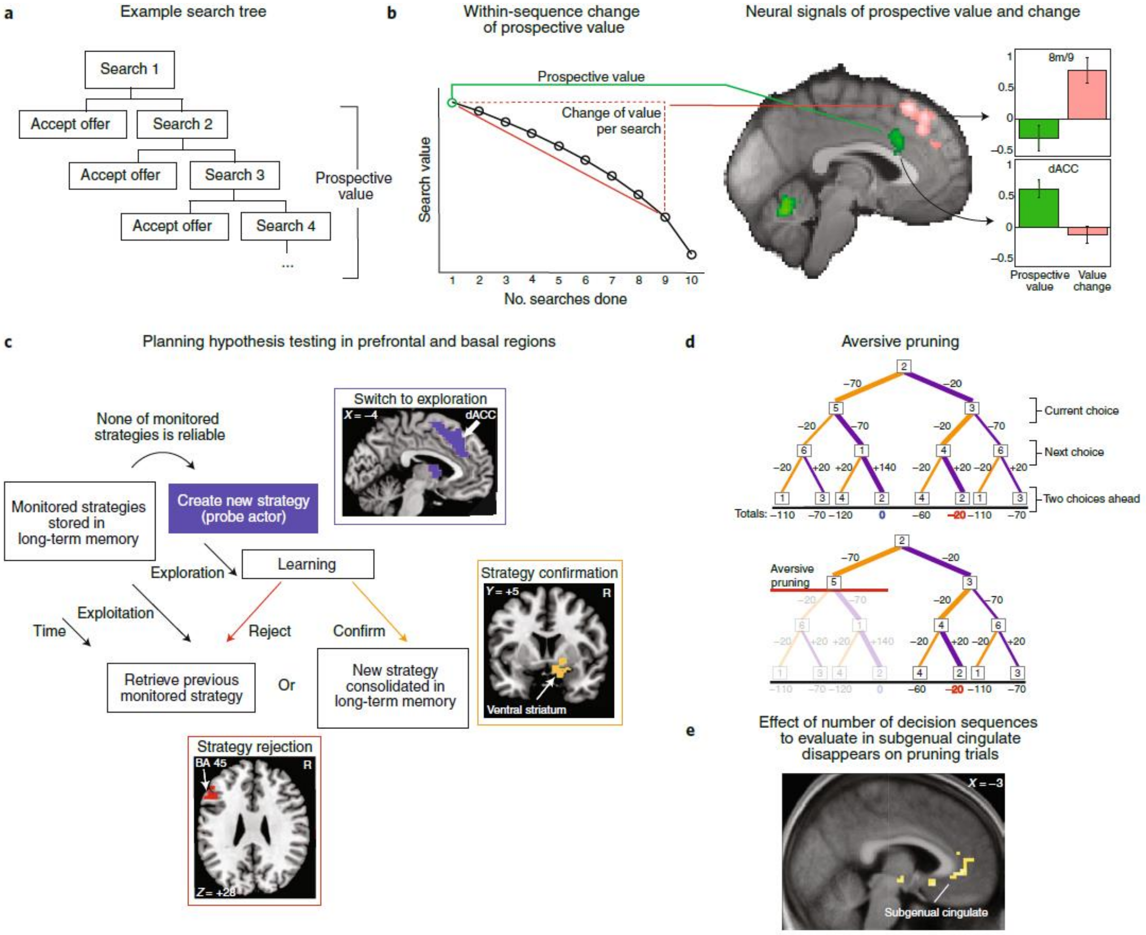
a
Example search tree
Search 1
Accept offer
Search 2
Accept offer
Search 3
Accept offer
Search 4
...
Prospective value
b
Within-sequence change of prospective value
Prospective value
Change of value per search
Search value
1 2 3 4 5 6 7 8 9 10
No. searches done
Neural signals of prospective value and change
8m/9
dACC
1
0.5
0
–0.5
Prospective value
Value change
c
Planning hypothesis testing in prefrontal and basal regions
None of monitored strategies is reliable
Monitored strategies stored in long-term memory
Create new strategy (probe actor)
Switch to exploration
X = –4
dACC
Exploration
Learning
Time
Exploitation
Reject
Confirm
Retrieve previous monitored strategy
Or
New strategy consolidated in long-term memory
Strategy confirmation
Y = +5
R
Ventral striatum
Strategy rejection
BA 45
R
Z = +28
d
Aversive pruning
–70
–20
Current choice
Next choice
Two choices ahead
Totals: –110 –70 –120 0 –60 –20 –110 –70
Aversive pruning
e
Effect of number of decision sequences to evaluate in subgenual cingulate disappears on pruning trials
X = –3
Subgenual cingulate

Figure 7. Cognitive planning behaviours can be functionally dissociated in several human fMRI studies. ***(a)*** *Planning is advantageous in a scenario where people can search a limited number of times and need to decide each time to accept the drawn offer or continue searching for a better one. The optimal solution to this problem is a search tree of all possible actions and outcomes for each potential search strategy. This allows computing prospective value - the value of continuing to search.* ***(b)*** *As people move through a sequence of searches and thus the opportunities to encounter good offers become fewer, prospective value decreases. Dorsal anterior cingulate cortex was sensitive to the initial prospective value, while activation in nearby dorsomedial frontal cortex (area 8m/9) correlated with how much the prospective value might change when going through the sequence. Thus it is linked to the potential required online adjustments in behavioural strategy[107]* ***(c)*** *In a model of reasoning fit to human responses in a task in which participants had to learn digit combinations through trial-and-error, different behavioral events were functionally dissociated in prefrontal and basal regions. Exploratory behaviour was associated with dACC activity, rejection of a new strategy was associated with dorsolateral prefrontal activity (BA 45), and confirmation of a new strategy was associated with ventral striatal activity. From [113]. Reprinted with permission from AAAS.* ***(d)*** *Aversive pruning is a non-optimal heuristic planning strategy in which the computational complexity is reduced by not computing the remained of a branch of a decision tree whenever a large loss is encountered[115].* ***(e)*** *While non-pruning trials had a clear value signal in subgenual cingulate cortex this was not present during trials where participants displayed aversive pruning.[116]*

Neurally, such considerations appear to involve an interplay between different dorsomedial and lateral prefrontal brain regions[107], which are regions uniquely specialised in primates. Human neuropsychology has established a fundamental role for dorsolateral prefrontal cortex (DLPFC) in lab-based planning tests[108] and in real-life strategic planning[109]. A neural basis for these functions is well established in monkey neurophysiological responses in DLPFC[21], whereas monitoring of constituent elements within extended sequential behaviours appears to depend upon dorsal anterior cingulate cortex (dACC) and pre-SMA regions[110].

Such responses contribute to a view of the frontal lobes as a rostro-caudal hierarchy, with more abstracted planning and control functions found more rostrally within this hierarchy[89]. The structures of representations that contribute to the elaboration of complex sequential plans can be seen to evolve as the task or environment is learned[111]. While dACC and its interactions with DLPFC appear particularly relevant for initial plan formation and prospective value generation, the nearby area 8m/9 considers how the initial plan will be prospectively adjusted following changes in the environment[107] (**Fig 6a/b**). One approach to formalise this process is to derive RL algorithms that learn mixtures of new plans across time, and appropriately decide whether a previously learnt plan should be reused or a new one depolyed[112]. Such models reveal functional dissociations when applied to fMRI data during strategy learning[113] (**Fig. 6c**).

However, even in more sophisticated cognitive behaviours, much of planning still boils down to sampling internal representations or simulating specific sequences of actions, outcomes and environmental dynamics. A major challenge, as in studies of navigation, remains knowing what the underlying representations or states are – over which actions are selected, outcomes are associated and environmental dynamics are predicted.

In behavioural tasks that involve mental simulation over multiple steps, several possible heuristics have been proposed for how humans might efficiently search through the large resulting state space. Each has had some supporting evidence. One option would be to only plan to a certain depth of a decision tree. In humans there is evidence for this[114]: people do not plan maximally deep, even when doing so would lead to greater reward. A related strategy is to stop sampling a specific branch if it appears to not be valuable (**Fig. 6d**). People indeed stop planning along branches that go through large losses, even when they are overall the best[115]. When this 'pruning' behaviour occurs, then subgenual cingulate activity no longer reflects the difficulty of the decision, defined in terms of the number of steps planned (**Fig. 6e**)[116]. An alternative strategy is to use 'hierarchical fragmentation'[117]: first plan a few steps, and from the best possible state there plan further. Finally, mixtures of explicit tree search and model free systems are also possible[118]. While the exact strategy used may be task-dependent, it is possible that newly developed methods for decoding sequences of representations in human MEG and fMRI data[64,65] could arbitrate between these heuristic planning strategies in multi-step cognitive tasks.

Information sampling as planning via exploration

Parallels between planning and information sampling. There are deep and as yet still relatively unexamined parallels between information creation, as in planning, and gathering new information, as in exploration. More particularly, they are parallel at the level of control – the decision about what (or whether) to explore, and what (or whether) to plan.

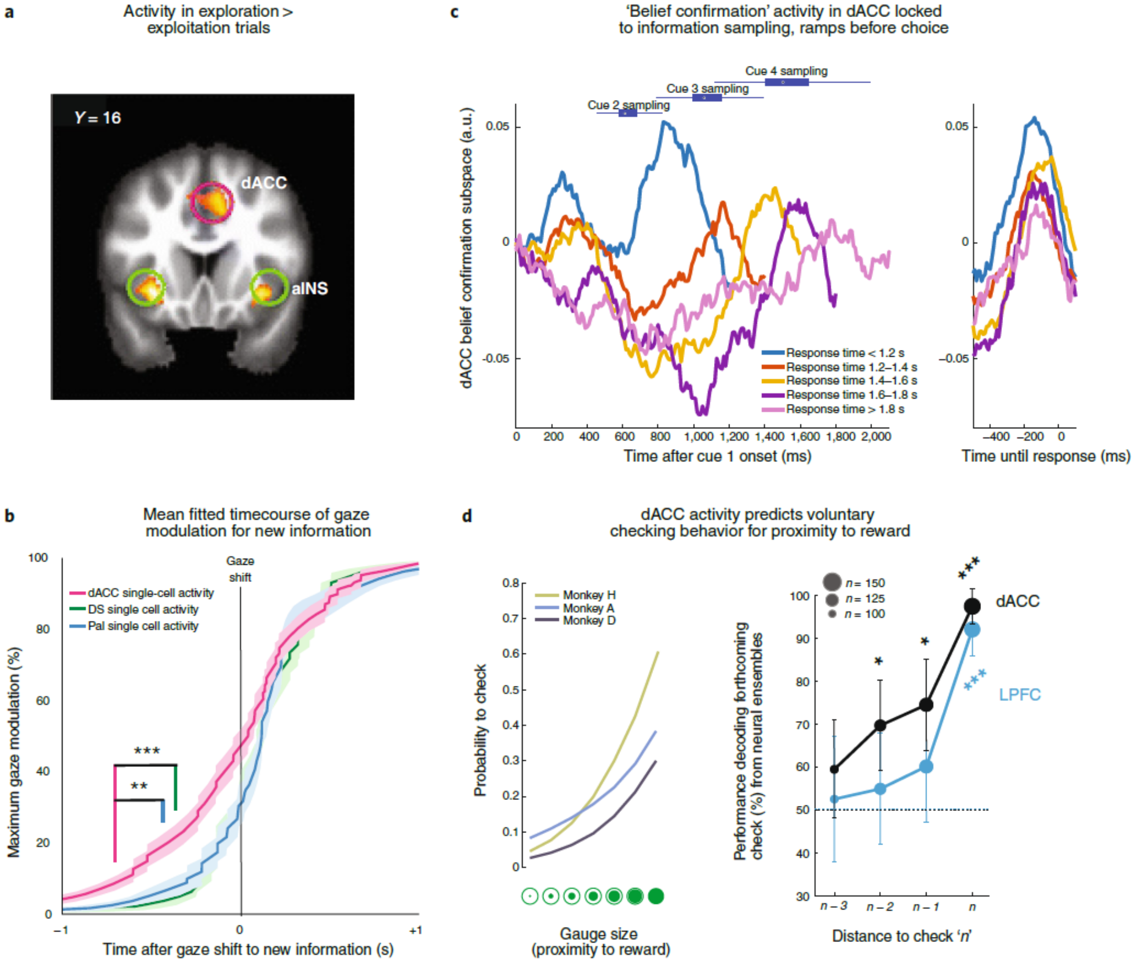

Figure 8. Activity in dorsal anterior cingulate cortex (dACC) associated with information sampling across multiple decision-making studies. (a) *Insula (aINS) and dorsal anterior cingulate cortex (dACC) show larger activity on exploration trials compared to exploitation trials in a human 'observe or bet' fMRI study.*[127] ***(b)*** *Activity in dorsal and ventral banks of ACC predicts gaze shifts to sample new information significantly earlier than interconnected portions of dorsal striatum (DS) and anterior palilidum (Pal) in monkey single-cell recordings.*[133] ***(c)*** *dACC population activity reflected whether new information confirmed or disconfirmed a belief about which option to choose in an economic choice task. This population also ramps prior to commitment to a final decision.*[134] ***(d)*** *Monkeys check a cue predictive of reward more when they are close to receiving a reward, and dACC single-cell activity predicts when a monkey will check the cue up to two trials beforehand.*[51]

In the RL framework, formal theories of optimal directed exploration[119,120] and deliberation[33,34] share essentially the same mathematical core. Whether accomplished "externally" through seeking new information in the world, or "internally" through model-based simulation, exploration is valuable to the extent that it changes your future choices. Indeed, the expected value of exploration can in principle be quantified as the increase in earnings expected to result from making better choices. This means, for instance, that both planning and exploration eventually have diminishing returns, after which they are unlikely to produce new actionable information (at which point one should act habitually or exploit, respectively). Also, even while they both can produce value, they must both be weighed against their opportunity cost, since planning comes at the expense of acting, and exploring comes at the expense of both exploiting and energy[121,122]. This ties them to yet a third closely related area of theory, optimal foraging[4] – i.e., optimizing search and foraging when the organism can only do one thing at a time. In such decisions, a choice is rarely a single motor impulse but instead a series of extended interactions with a particular goal in mind. Information sampling may not only benefit the initial choice, but also the planning of the series of future actions taken after a choice has been made.

So far, we have presented planning as a process of sampling and simulating the future. However, if an agent's knowledge about the world is wrong or incomplete, sampling the actual world, rather than a simulated one from memory, is essential. Importantly, an agent can direct their exploration towards parts of the environment that are known unknowns, either because they have an explicit model of the uncertainty of their estimates[122], or because they know how the environment will change over time[123]. This can be used to quantify the value of reducing uncertainty for different states[34] and to quantify the gain of information against the energetic cost of gaining that information[121,122].

Value of information as narrowing planning and improving predictions. While existing models do not predict information sampling and planning in a unified manner, empirical observations suggest that information sampling can be highly strategic. For example, humans explore more when the information is more valuable because it can be used in the future. Such exploration is not random, but directed toward options with more uncertainty[124]. Early fMRI studies of exploratory behaviour identified a network of regions including dACC (see also **Fig. 6c**), frontopolar cortex and intraparietal sulcus that governed switches away from a currently favoured option towards exploring an alternative[125,126]. Subsequent studies have to some extent dissociated these regions, into those that reflect a simple decision to sample information, which activates dACC (**Fig. 7a**)[127], versus frontopolar cortex that tracks estimates of option uncertainty across time[128]. Disrupting frontopolar cortex using transcranial magnetic stimulation selectively affects directed but not undirected exploration[129]. The converse is true of pharmacological interventions targeting the noradrenergic system[130], whose inputs to dACC have been shown to modulate switching into exploratory behaviour[131].

Interestingly, animals also value information when it is of no apparent reward value. Several species have been shown to gamble energy of movement proportionate to the expected information gain[122]. Given the advancement of planning, sampling and simulation models, it should be possible to predict what kind of information an agent would be willing to pay for ("simulation pruning") even if it does not directly link to reward, as it might nevertheless significantly benefit planning. For example, macaques will pay a cost to resolve uncertainty about a future outcome earlier[132]. This makes sense if the brain continuously predicts potential future outcomes through simulation and sampling but tries to avoid unnecessarily anticipating potential outcomes that could be ruled out.

A recent study showed that neurons in several interconnected subregions of primate dACC and basal ganglia are active around eye-gaze movements that resolve uncertainty, with dACC being first to predict saccades that resolve uncertainty[133] (**Fig. 7b**). In a task where multiple saccades must be made to sample information about two choice options, activity in dACC reports whether newly revealed evidence confirms or disconfirms a prior belief about which option should be chosen[134]. Activity in this dACC 'belief confirmation subspace' ramps immediately prior to commitment to a final decision (**Fig. 7c**), suggesting a role for dACC in transforming newly sampled information into future choice behaviour.

While the exploration-exploitation dilemma is often considered in terms of improving estimates of a static value function, another strong motivation for exploration in real-world behaviour is to sample when the world has changed. Indeed, macaques can adapt their search behaviour to specific features of environments[123]. Importantly, animals can even monitor internal representations of unobservable dynamic changes in the environment to optimize their checking behaviours and update those representations. Activity in dACC ramps across time prior to these checking behaviours, meaning that checks can be decoded on preceding trials[51] (**Fig 7d**).

Linking successor representations to information sampling in foraging problems. Ethological observations have shown that the exploratory patterns in many species follow statistical rules known as Lévy walks, with travel paths that follow scale-free power laws[135,136]. In conditions where prey are sparse, such patterns are more efficient than pure random movements to capture these prey. It is argued that this advantage will have acted as a selection pressure on adaptations that would give rise to Levy flight foraging[137].

Above, we highlighted the eigendecomposition of the successor representation as a model for grid cell activity in the entorhinal cortex during navigation and planning[82]; intriguingly, this may also provide a basis for generating Lévy walks. Different eigenvectors of this representation will occur at different spatial scales, meaning that they may be suitable for planning over different horizons. Indeed, recent evidence from a navigational planning task using human fMRI revealed a posterior-to-anterior spatial gradient in both hippocampus and prefrontal cortex, reflecting pattern similarity to successor states of increasing spatial scales[90].

When generating future actions, upweighting eigenvectors which represent low-frequency spatial information naturally leads the agent to adopt Lévy-like exploration of the environment. This exploration proves to be more efficient than random exploration when searching over environments with hierarchical structure, such as connected rooms[138]. By contrast, the sequences of samples generated by random exploration will better capture the true structure of the environment. This may explain why offline replay events in the hippocampus appear to follow a random diffusive pattern, even following behavioural exploration that has a Lévy-like superdiffusive structure[139] – at least in the absence of goals that shape replay events towards locations useful for planning[33]. One potential issue here is that Lévy-like exploration is only predictive in information-scarce and low resource density contexts[140]. In information-rich contexts in which search proceeds in range of sensory organs, energy-constrained proportional betting on the expected information distribution is showing promise for predicting trajectories across multiple species[122].

Linking theta oscillations to external sampling. It is also clear that some of the neural implementations of online planning discussed earlier are also relevant for information sampling behaviours. Exploration signals have been shown to exist in conditions of high uncertainty in form of nonlocal representation of space along each theta cycle at high-cost decision points (VTE)[36,141]. The very same theta cycles are also seen during internally generated sub-second patterns that govern sensory perception[142] and sensorimotor actions[143]. Thus, these patterns, currently thought to reflect adaptive mechanisms for sampling information from the external world, may be coordinated with the sub-second patterns of generative activity described here, which can in turn be likened to sampling from internal representations.

In biological agents as in artificial ones, a major purpose of external information sampling is to improve one's confidence in pursuing the most valuable course of action. Converging evidence from information sampling studies in humans[144-146] and non-human primates[134] indicates a bias towards sampling evidence from a goal that is currently most favoured, rather than the option that will maximally reduce uncertainty. This fits well with foraging models of choice, which argue that even simple binary decisions may be made as a sequence of accept-reject decisions rather than as a direct comparison between two alternatives[147]. Once animals commit to accepting an option, they pursue this goal even when it becomes costly to do so[148]; sampling information may benefit planning of future actions needed to pursue their goal. Formalising this account of choice may require us to reformulate the RL problem as being one of minimising distance to goals, rather than maximising discounted future reward[149].

Summary

In this review we have described some formal approaches, ideas and theories that have begun to breach into the territory of internal planning and information sampling in complex environments. Some of these have previously often been thought of as being too difficult, idiosyncratic or unstructured to be investigated directly. A couple of concepts have crystalized as being essential for this advance. Firstly, we conceive of planning as problem of internal sampling of a simulated environment, while trying to optimize such sampling toward the most valuable and most likely aspects of the future. Second, this progress is paired with a need to understand how states and knowledge are efficiently and conceptually organized to allow for planning in the first place. Knowing how to plan by sampling, and what to plan over, allows the assessment of the evolutionary as well as individual benefits of planning as well as predictions of specific behaviour and neural mechanisms linked to overall planning and memory retrieval, consolidation and decision making specifically.

References

1 Gold, J. I. & Shadlen, M. N. The neural basis of decision making. *Annu Rev Neurosci* **30**, 535-574, doi:10.1146/annurev.neuro.29.051605.113038 (2007).

2 Niv, Y. Reinforcement learning in the brain. *Journal of Mathematical Psychology* **53**, 139-154, doi:10.1016/j.jmp.2008.12.005 (2009).

3 Glimcher, P. W. & Fehr, E. *Neuroeconomics : decision making and the brain*. Second edition. edn, (Elsevier/AP, Academic Press is an imprint of Elsevier, 2014).

4 Mobbs, D., Trimmer, P. C., Blumstein, D. T. & Dayan, P. Foraging for foundations in decision neuroscience: insights from ethology. *Nat Rev Neurosci* **19**, 419-427, doi:10.1038/s41583-018-0010-7 (2018).

5 Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* **8**, 1704-1711, doi:10.1038/nn1560 (2005).

6 Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312-325, doi:10.1016/j.neuron.2013.09.007 (2013).

7 Redish, A. D. Vicarious trial and error. *Nat Rev Neurosci* **17**, 147-159, doi:10.1038/nrn.2015.30 (2016).

8 Jones, J. L. *et al.* Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science* **338**, 953-956, doi:10.1126/science.1227489 (2012).

9 Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D. & Daw, N. D. Model-based choices involve prospective neural activity. *Nat Neurosci* **18**, 767-772, doi:10.1038/nn.3981 (2015).

10 Schmidt, B., Duin, A. A. & Redish, A. D. Disrupting the medial prefrontal cortex alters hippocampal sequences during deliberative decision making. *J Neurophysiol* **121**, 1981-2000, doi:10.1152/jn.00793.2018 (2019).

11 Wilkinson, A. & Huber, L. in *The Oxford handbook of comparative evolutionary psychology* (eds J. Vonk & T. K. Shackelford) 129–143 (Oxford University Press, 2012).

12 Burghardt, G. M. Environmental enrichment and cognitive complexity in reptiles and amphibians: Concepts, review, and implications for captive populations. *Applied Animal Behaviour Science* **147**, 286-298, doi:10.1016/j.applanim.2013.04.013 (2013).

13 Broglio, C. *et al.* Hippocampal Pallium and Map-Like Memories through Vertebrate Evolution. *Journal of Behavioral and Brain Science* **05**, 109-120, doi:10.4236/jbbs.2015.53011 (2015).

14 MacIver, M. A., Schmitz, L., Mugan, U., Murphey, T. D. & Mobley, C. D. Massive increase in visual range preceded the origin of terrestrial vertebrates. *Proc Natl Acad Sci U S A* **114**, E2375-E2384, doi:10.1073/pnas.1615563114 (2017).

15 Stein, W. E., Berry, C. M., Hernick, L. V. & Mannolini, F. Surprisingly complex community discovered in the mid-Devonian fossil forest at Gilboa. *Nature* **483**, 78-81, doi:10.1038/nature10819 (2012).

16 Mugan, U. & MacIver, M. A. Spatial planning with long visual range benefits escape from visual predators in complex naturalistic environments. *Nat Commun* **11**, 3057, doi:10.1038/s41467-020-16102-1 (2020).

17 Tolman, E. C. Cognitive maps in rats and men. *Psychol Rev* **55**, 189-208, doi:10.1037/h0061626 (1948).

18 Tse, D. *et al.* Schemas and memory consolidation. *Science* **316**, 76-82, doi:10.1126/science.1135935 (2007).

19 Raby, C. R., Alexis, D. M., Dickinson, A. & Clayton, N. S. Planning for the future by western scrub-jays. *Nature* **445**, 919-921, doi:10.1038/nature05575 (2007).

20 Wimpenny, J. H., Weir, A. A., Clayton, L., Rutz, C. & Kacelnik, A. Cognitive processes associated with sequential tool use in New Caledonian crows. *PLoS One* **4**, e6471, doi:10.1371/journal.pone.0006471 (2009).

21 Tanji, J., Shima, K. & Mushiake, H. Concept-based behavioral planning and the lateral prefrontal cortex. *Trends Cogn Sci* **11**, 528-534, doi:10.1016/j.tics.2007.09.007 (2007).

22 Clutton-Brock, T. H. & Harvey, P. H. Primates, brains and ecology. *Journal of Zoology* **190**, 309-323, doi:10.1111/j.1469-7998.1980.tb01430.x (1980).

23 Conway, C. M. & Christiansen, M. H. Sequential learning in non-human primates. *Trends Cogn Sci* **5**, 539-546, doi:10.1016/s1364-6613(00)01800-3 (2001).

24 Le Fur, S., Fara, E., Mackaye, H. T., Vignaud, P. & Brunet, M. The mammal assemblage of the hominid site TM266 (Late Miocene, Chad Basin): ecological structure and paleoenvironmental implications. *Naturwissenschaften* **96**, 565-574, doi:10.1007/s00114-008-0504-7 (2009).

25 Dunbar, R. I. M. & Shultz, S. Why are there so many explanations for primate brain evolution? *Philos Trans R Soc Lond B Biol Sci* **372**, doi:10.1098/rstb.2016.0244 (2017).

26 Lee, D. & Seo, H. Neural Basis of Strategic Decision Making. *Trends Neurosci* **39**, 40-48, doi:10.1016/j.tins.2015.11.002 (2016).

27 Gottlieb, J. & Oudeyer, P. Y. Towards a neuroscience of active sampling and curiosity. *Nat Rev Neurosci* **19**, 758-770, doi:10.1038/s41583-018-0078-0 (2018).

28 Glickman, S. E. & Sroges, R. W. Curiosity in zoo animals. *Behaviour* **26**, 151-188, doi:10.1163/156853966x00074 (1966).

29 Montgomery, S. H. The relationship between play, brain growth and behavioural flexibility in primates. *Animal Behaviour* **90**, 281-286, doi:10.1016/j.anbehav.2014.02.004 (2014).

30 Wimpenny, J. H., Weir, A. A. & Kacelnik, A. New Caledonian crows use tools for non-foraging activities. *Anim Cogn* **14**, 459-464, doi:10.1007/s10071-010-0366-1 (2011).

31 Callaway, F. *et al.* Human planning as optimal information seeking. *PsyArXiv preprints*, doi:10.31234/osf.io/byaqd (2021).

32 Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204-1215, doi:10.1016/j.neuron.2011.02.027 (2011).

33 Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat Neurosci* **21**, 1609-1617, doi:10.1038/s41593-018-0232-z (2018).

34 Keramati, M., Dezfouli, A. & Piray, P. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput Biol* **7**, e1002055, doi:10.1371/journal.pcbi.1002055 (2011).

35 Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490-509, doi:10.1016/j.neuron.2018.10.002 (2018).

36 Johnson, A. & Redish, A. D. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *J Neurosci* **27**, 12176-12189, doi:10.1523/JNEUROSCI.3761-07.2007 (2007).

37 Kay, K. *et al.* Constant Sub-second Cycling between Representations of Possible Futures in the Hippocampus. *Cell* **180**, 552-567 e525, doi:10.1016/j.cell.2020.01.014 (2020).

38 Diba, K. & Buzsaki, G. Forward and reverse hippocampal place-cell sequences during ripples. *Nat Neurosci* **10**, 1241-1242, doi:10.1038/nn1961 (2007).

39 Buzsaki, G. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus* **25**, 1073-1188, doi:10.1002/hipo.22488 (2015).

40 Gupta, A. S., van der Meer, M. A., Touretzky, D. S. & Redish, A. D. Segmentation of spatial experience by hippocampal theta sequences. *Nat Neurosci* **15**, 1032-1039, doi:10.1038/nn.3138 (2012).

41 Zielinski, M. C., Shin, J. D. & Jadhav, S. P. Coherent Coding of Spatial Position Mediated by Theta Oscillations in the Hippocampus and Prefrontal Cortex. *J Neurosci* **39**, 4550-4565, doi:10.1523/JNEUROSCI.0106-19.2019 (2019).

42 van der Meer, M. A. & Redish, A. D. Expectancies in decision making, reinforcement learning, and ventral striatum. *Front Neurosci* **4**, 6, doi:10.3389/neuro.01.006.2010 (2010).

43 Gardner, R. S. *et al.* A secondary working memory challenge preserves primary place strategies despite overtraining. *Learn Mem* **20**, 648-656, doi:10.1101/lm.031336.113 (2013).

44 Steiner, A. P. & Redish, A. D. The road not taken: neural correlates of decision making in orbitofrontal cortex. *Front Neurosci* **6**, 131, doi:10.3389/fnins.2012.00131 (2012).

45 Powell, N. J. & Redish, A. D. Complex neural codes in rat prelimbic cortex are stable across days on a spatial decision task. *Front Behav Neurosci* **8**, 120, doi:10.3389/fnbeh.2014.00120 (2014).

46 Stott, J. J. & Redish, A. D. A functional difference in information processing between orbitofrontal cortex and ventral striatum during decision-making behaviour. *Philos Trans R Soc Lond B Biol Sci* **369**, doi:10.1098/rstb.2013.0472 (2014).

47 Hu, D. & Amsel, A. A simple test of the vicarious trial-and-error hypothesis of hippocampal function. *Proc Natl Acad Sci U S A* **92**, 5506-5509, doi:10.1073/pnas.92.12.5506 (1995).

48 Meyer-Mueller, C. *et al.* Dorsal, but not ventral, hippocampal inactivation alters deliberation in rats. *Behav Brain Res* **390**, 112622, doi:10.1016/j.bbr.2020.112622 (2020).

49 Kreher, M. A. *et al.* The perirhinal cortex supports spatial intertemporal choice stability. *Neurobiol Learn Mem* **162**, 36-46, doi:10.1016/j.nlm.2019.05.002 (2019).

50 Procyk, E., Tanaka, Y. L. & Joseph, J. P. Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. *Nat Neurosci* **3**, 502-508, doi:10.1038/74880 (2000).

51 Stoll, F. M., Fontanier, V. & Procyk, E. Specific frontal neural dynamics contribute to decisions to check. *Nat Commun* **7**, 11990, doi:10.1038/ncomms11990 (2016).

52 Singer, A. C. & Frank, L. M. Rewarded outcomes enhance reactivation of experience in the hippocampus. *Neuron* **64**, 910-921, doi:10.1016/j.neuron.2009.11.016 (2009).

53 Papale, A. E., Zielinski, M. C., Frank, L. M., Jadhav, S. P. & Redish, A. D. Interplay between Hippocampal Sharp-Wave-Ripple Events and Vicarious Trial and Error Behaviors in Decision Making. *Neuron* **92**, 975-982, doi:10.1016/j.neuron.2016.10.028 (2016).

54 Gupta, A. S., van der Meer, M. A., Touretzky, D. S. & Redish, A. D. Hippocampal replay is not a simple function of experience. *Neuron* **65**, 695-705, doi:10.1016/j.neuron.2010.01.034 (2010).

55 Singer, A. C., Carr, M. F., Karlsson, M. P. & Frank, L. M. Hippocampal SWR activity predicts correct decisions during the initial learning of an alternation task. *Neuron* **77**, 1163-1173, doi:10.1016/j.neuron.2013.01.027 (2013).

56 Jadhav, S. P., Kemere, C., German, P. W. & Frank, L. M. Awake hippocampal sharp-wave ripples support spatial memory. *Science* **336**, 1454-1458, doi:10.1126/science.1217230 (2012).

57 Foster, D. J. & Wilson, M. A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680-683, doi:10.1038/nature04587 (2006).

58 Ambrose, R. E., Pfeiffer, B. E. & Foster, D. J. Reverse Replay of Hippocampal Place Cells Is Uniquely Modulated by Changing Reward. *Neuron* **91**, 1124-1136, doi:10.1016/j.neuron.2016.07.047 (2016).

59 Davidson, T. J., Kloosterman, F. & Wilson, M. A. Hippocampal replay of extended experience. *Neuron* **63**, 497-507, doi:10.1016/j.neuron.2009.07.027 (2009).

60 Olafsdottir, H. F., Barry, C., Saleem, A. B., Hassabis, D. & Spiers, H. J. Hippocampal place cells construct reward related sequences through unexplored space. *Elife* **4**, e06063, doi:10.7554/eLife.06063 (2015).

61 Miller, K. J. & Venditto, S. J. Multi-Step Planning in the Brain. *PsyArXiv preprints*, doi:10.31234/osf.io/kv86m (2020).

62 Kurth-Nelson, Z., Economides, M., Dolan, R. J. & Dayan, P. Fast Sequences of Non-spatial State Representations in Humans. *Neuron* **91**, 194-204, doi:10.1016/j.neuron.2016.05.028 (2016).

63 Momennejad, I., Otto, A. R., Daw, N. D. & Norman, K. A. Offline replay supports planning in human reinforcement learning. *Elife* **7**, doi:10.7554/eLife.32548 (2018).

64 Schuck, N. W. & Niv, Y. Sequential replay of nonspatial task states in the human hippocampus. *Science* **364**, doi:10.1126/science.aaw5181 (2019).

65 Liu, Y., Dolan, R. J., Kurth-Nelson, Z. & Behrens, T. E. J. Human Replay Spontaneously Reorganizes Experience. *Cell* **178**, 640-652 e614, doi:10.1016/j.cell.2019.06.012 (2019).

66 Liu, Y., Mattar, M. G., Behrens, T. E. J., Daw, N. D. & Dolan, R. J. Experience replay supports non-local learning. *bioRxiv*, 2020.2010.2020.343061, doi:10.1101/2020.10.20.343061 (2020).

67 van Opheusden, B. & Ma, W. J. Tasks for aligning human and machine planning. *Current Opinion in Behavioral Sciences* **29**, 127-133, doi:10.1016/j.cobeha.2019.07.002 (2019).

68 Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proc Natl Acad Sci U S A* **105**, 10687-10692, doi:10.1073/pnas.0802631105 (2008).

69 Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35**, 1798-1828, doi:10.1109/tpami.2013.50 (2013).

70 Radulescu, A., Niv, Y. & Ballard, I. Holistic Reinforcement Learning: The Role of Structure and Attention. *Trends in Cognitive Sciences* **23**, 278-292, doi:10.1016/j.tics.2019.01.010 (2019).

71 Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M. & Behrens, T. E. J. Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*, doi:10.1016/j.neuron.2020.11.024 (2020).

72 Schulz, E., Franklin, N. T. & Gershman, S. J. Finding structure in multi-armed bandits. *Cogn Psychol* **119**, 101261, doi:10.1016/j.cogpsych.2019.101261 (2020).

73 Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D. & Meder, B. Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour* **2**, 915-924, doi:10.1038/s41562-018-0467-4 (2018).

74 Harlow, H. F. The formation of learning sets. *Psychol Rev* **56**, 51-65, doi:10.1037/h0062474 (1949).

75 Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat Neurosci* **21**, 860-868, doi:10.1038/s41593-018-0147-8 (2018).

76 Browning, P. G., Easton, A. & Gaffan, D. Frontal-temporal disconnection abolishes object discrimination learning set in macaque monkeys. *Cereb Cortex* **17**, 859-864, doi:10.1093/cercor/bhk039 (2007).

77 M'Harzi, M. *et al.* Effects of selective lesions of fimbria-fornix on learning set in the rat. *Physiol Behav* **40**, 181-188, doi:10.1016/0031-9384(87)90205-8 (1987).

78 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, doi:10.1016/j.neuron.2016.08.019 (2016).

79 Dayan, P. Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation* **5**, 613-624, doi:10.1162/neco.1993.5.4.613 (1993).

80 Singh, S., James, M. R. & Rudary, M. R. in *Proceedings of the 20th conference on Uncertainty in artificial intelligence* 512–519 (AUAI Press, Banff, Canada, 2004).

81 Gershman, S. J. The Successor Representation: Its Computational Logic and Neural Substrates. *J Neurosci* **38**, 7193-7200, doi:10.1523/JNEUROSCI.0151-18.2018 (2018).

82 Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat Neurosci* **20**, 1643-1653, doi:10.1038/nn.4650 (2017).

83 Mehta, M. R., Quirk, M. C. & Wilson, M. A. Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron* **25**, 707-715, doi:10.1016/s0896-6273(00)81072-7 (2000).

84 Momennejad, I. *et al.* The successor representation in human reinforcement learning. *Nat Hum Behav* **1**, 680-692, doi:10.1038/s41562-017-0180-8 (2017).

85 Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput Biol* **13**, e1005768, doi:10.1371/journal.pcbi.1005768 (2017).

86 Gardner, M. P. H., Schoenbaum, G. & Gershman, S. J. Rethinking dopamine as generalized prediction error. *Proc Biol Sci* **285**, doi:10.1098/rspb.2018.1645 (2018).

87 Barnes, T. D., Kubota, Y., Hu, D., Jin, D. Z. & Graybiel, A. M. Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature* **437**, 1158-1161, doi:10.1038/nature04053 (2005).

88 van der Meer, M. A., Johnson, A., Schmitzer-Torbert, N. C. & Redish, A. D. Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron* **67**, 25-32, doi:10.1016/j.neuron.2010.06.023 (2010).

89 Koechlin, E. & Summerfield, C. An information theoretical approach to prefrontal executive function. *Trends Cogn Sci* **11**, 229-235, doi:10.1016/j.tics.2007.04.005 (2007).

90 Brunec, I. K. & Momennejad, I. Predictive Representations in Hippocampal and Prefrontal Hierarchies. *BioRxiv*, 786434 (2019).

91 Mahadevan, S. & Maggioni, M. Proto-value Functions: A Laplacian Framework for Learning Representation and Control in Markov Decision Processes. *Journal of Machine Learning Research* **8**, 2169-2231 (2007).

92 Machado, M. C., Bellemare, M. G. & Bowling, M. Count-based exploration with the successor representation. *arXiv preprint arXiv:1807.11622* (2018).

93 Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M. & Norman, K. A. 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**, doi:10.1098/rstb.2016.0049 (2017).

94 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.e1223, doi:10.1016/j.cell.2020.10.024 (2020).

95 Piray, P. & Daw, N. D. A common model explaining flexible decision making, grid fields and cognitive control. *bioRxiv*, 856849 (2019).

96 Russo, E. & Durstewitz, D. Cell assemblies at multiple time scales with arbitrary lag constellations. *Elife* **6**, doi:10.7554/eLife.19428 (2017).

97 Durstewitz, D. A state space approach for piecewise-linear recurrent neural networks for identifying computational dynamics from neural measurements. *PLoS Comput Biol* **13**, e1005542, doi:10.1371/journal.pcbi.1005542 (2017).

98 Pipa, G., Wheeler, D. W., Singer, W. & Nikolic, D. NeuroXidence: reliable and efficient analysis of an excess or deficiency of joint-spike events. *J Comput Neurosci* **25**, 64-88, doi:10.1007/s10827-007-0065-3 (2008).

99 Benchenane, K. *et al.* Coherent theta oscillations and reorganization of spike timing in the hippocampal- prefrontal network upon learning. *Neuron* **66**, 921-936, doi:10.1016/j.neuron.2010.05.013 (2010).

100 Quaglio, P., Yegenoglu, A., Torre, E., Endres, D. M. & Grun, S. Detection and Evaluation of Spatio-Temporal Spike Patterns in Massively Parallel Spike Train Data with SPADE. *Front Comput Neurosci* **11**, 41, doi:10.3389/fncom.2017.00041 (2017).

101 Grossberger, L., Battaglia, F. P. & Vinck, M. Unsupervised clustering of temporal patterns in high-dimensional neuronal ensembles using a novel dissimilarity measure. *PLoS Comput Biol* **14**, e1006283, doi:10.1371/journal.pcbi.1006283 (2018).

102 van de Ven, G. M., Trouche, S., McNamara, C. G., Allen, K. & Dupret, D. Hippocampal Offline Reactivation Consolidates Recently Formed Cell Assembly Patterns during Sharp Wave-Ripples. *Neuron* **92**, 968-974, doi:10.1016/j.neuron.2016.10.020 (2016).

103 Gridchyn, I., Schoenenberger, P., O'Neill, J. & Csicsvari, J. Assembly-Specific Disruption of Hippocampal Replay Leads to Selective Memory Deficit. *Neuron* **106**, 291-300 e296, doi:10.1016/j.neuron.2020.01.021 (2020).

104 Oettl, L. L. *et al.* Phasic dopamine reinforces distinct striatal stimulus encoding in the olfactory tubercle driving dopaminergic reward prediction. *Nat Commun* **11**, 3460, doi:10.1038/s41467-020-17257-7 (2020).

105 Gomperts, S. N., Kloosterman, F. & Wilson, M. A. VTA neurons coordinate with the hippocampal reactivation of spatial experience. *Elife* **4**, doi:10.7554/eLife.05360 (2015).

106 Kurth-Nelson, Z. & Redish, A. D. Don't let me do that! - models of precommitment. *Front Neurosci* **6**, 138, doi:10.3389/fnins.2012.00138 (2012).

107 Kolling, N., Scholl, J., Chekroud, A., Trier, H. A. & Rushworth, M. F. S. Prospection, Perseverance, and Insight in Sequential Behavior. *Neuron* **99**, 1069-1082 e1067, doi:10.1016/j.neuron.2018.08.018 (2018).

108 Goel, V. & Grafman, J. Are the frontal lobes implicated in "planning" functions? Interpreting data from the Tower of Hanoi. *Neuropsychologia* **33**, 623-642, doi:10.1016/0028-3932(95)90866-p (1995).

109 Burgess, P. W. Strategy application disorder: the role of the frontal lobes in human multitasking. *Psychol Res* **63**, 279-288, doi:10.1007/s004269900006 (2000).

110 Holroyd, C. B., Ribas-Fernandes, J. J. F., Shahnazian, D., Silvetti, M. & Verguts, T. Human midcingulate cortex encodes distributed representations of task progress. *Proc Natl Acad Sci U S A* **115**, 6398-6403, doi:10.1073/pnas.1803650115 (2018).

111 Averbeck, B. B., Sohn, J. W. & Lee, D. Activity in prefrontal cortex during dynamic selection of action sequences. *Nat Neurosci* **9**, 276-282, doi:10.1038/nn1634 (2006).

112 Collins, A. & Koechlin, E. Reasoning, learning, and creativity: frontal lobe function and human decision-making. *PLoS Biol* **10**, e1001293, doi:10.1371/journal.pbio.1001293 (2012).

113 Donoso, M., Collins, A. G. & Koechlin, E. Human cognition. Foundations of human reasoning in the prefrontal cortex. *Science* **344**, 1481-1486, doi:10.1126/science.1252254 (2014).

114 Juechems, K. *et al.* A Network for Computing Value Equilibrium in the Human Medial Prefrontal Cortex. *Neuron* **101**, 977-987 e973, doi:10.1016/j.neuron.2018.12.029 (2019).

115 Huys, Q. J. *et al.* Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput Biol* **8**, e1002410, doi:10.1371/journal.pcbi.1002410 (2012).

116 Lally, N. *et al.* The Neural Basis of Aversive Pavlovian Guidance during Planning. *J Neurosci* **37**, 10215-10229, doi:10.1523/JNEUROSCI.0085-17.2017 (2017).

117 Huys, Q. J. *et al.* Interplay of approximate planning strategies. *Proc Natl Acad Sci U S A* **112**, 3098-3103, doi:10.1073/pnas.1414219112 (2015).

118 Keramati, M., Smittenaar, P., Dolan, R. J. & Dayan, P. Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proc Natl Acad Sci U S A* **113**, 12868-12873, doi:10.1073/pnas.1609094113 (2016).

119 Gittins, J. C. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)* **41**, 148-164 (1979).

120 Russo, D., Van Roy, B., Kazerouni, A., Osband, I. & Wen, Z. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038* (2017).

121 MacIver, M. A., Patankar, N. A. & Shirgaonkar, A. A. Energy-Information Trade-Offs between Movement and Sensing. *PLoS Computational Biology* **6**, doi:10.1371/journal.pcbi.1000769 (2010).

122 Chen, C., Murphey, T. D. & MacIver, M. A. Tuning movement for sensing in an uncertain world. *Elife* **9**, e52371, doi:10.7554/eLife.52371 (2020).

123 Khamassi, M., Quilodran, R., Enel, P., Dominey, P. F. & Procyk, E. Behavioral Regulation and the Modulation of Information Coding in the Lateral Prefrontal and Cingulate Cortex. *Cereb Cortex* **25**, 3197-3218, doi:10.1093/cercor/bhu114 (2015).

124 Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore-exploit dilemma. *J Exp Psychol Gen* **143**, 2074-2081, doi:10.1037/a0038199 (2014).

125 Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876-879, doi:10.1038/nature04766 (2006).

126 Boorman, E. D., Behrens, T. E., Woolrich, M. W. & Rushworth, M. F. How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* **62**, 733-743, doi:10.1016/j.neuron.2009.05.014 (2009).

127 Blanchard, T. C. & Gershman, S. J. Pure correlates of exploration and exploitation in the human brain. *Cogn Affect Behav Neurosci* **18**, 117-126, doi:10.3758/s13415-017-0556-2 (2018).

128 Badre, D., Doll, B. B., Long, N. M. & Frank, M. J. Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron* **73**, 595-607, doi:10.1016/j.neuron.2011.12.025 (2012).

129 Zajkowski, W. K., Kossut, M. & Wilson, R. C. A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife* **6**, doi:10.7554/eLife.27430 (2017).

130 Warren, C. M. *et al.* The effect of atomoxetine on random and directed exploration in humans. *PLoS One* **12**, e0176034, doi:10.1371/journal.pone.0176034 (2017).

131 Tervo, D. G. R. *et al.* Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell* **159**, 21-32, doi:10.1016/j.cell.2014.08.037 (2014).

132 Blanchard, T. C., Hayden, B. Y. & Bromberg-Martin, E. S. Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron* **85**, 602-614, doi:10.1016/j.neuron.2014.12.050 (2015).

133 White, J. K. *et al.* A neural network for information seeking. *Nat Commun* **10**, 5168, doi:10.1038/s41467-019-13135-z (2019).

134 Hunt, L. T. *et al.* Triple dissociation of attention and decision computations across prefrontal cortex. *Nat Neurosci* **21**, 1471-1481, doi:10.1038/s41593-018-0239-5 (2018).

135 Ayala-Orozco, B. *et al.* Lévy walk patterns in the foraging movements of spider monkeys (Ateles geoffroyi). *Behavioral Ecology and Sociobiology* **55**, 223-230, doi:10.1007/s00265-003-0700-6 (2004).

136 Sims, D. W. *et al.* Scaling laws of marine predator search behaviour. *Nature* **451**, 1098-1102, doi:10.1038/nature06518 (2008).

137 Viswanathan, G. M., Raposo, E. P. & da Luz, M. G. E. Lévy flights and superdiffusion in the context of biological encounters and random searches. *Physics of Life Reviews* **5**, 133-150, doi:10.1016/j.plrev.2008.03.002 (2008).

138 McNamee, D., Stachenfeld, K. L., Botvinick, M. & Gershman, S. J. in *Society for Neuroscience* (San Diego, CA, 2018).

139 Stella, F., Baracskay, P., O'Neill, J. & Csicsvari, J. Hippocampal Reactivation of Random Trajectories Resembling Brownian Diffusion. *Neuron* **102**, 450-461 e457, doi:10.1016/j.neuron.2019.01.052 (2019).

140 Wosniack, M. E., Santos, M. C., Raposo, E. P., Viswanathan, G. M. & da Luz, M. G. E. The evolutionary origins of Levy walk foraging. *PLoS Comput Biol* **13**, e1005774, doi:10.1371/journal.pcbi.1005774 (2017).

141 Skaggs, W. E., McNaughton, B. L., Wilson, M. A. & Barnes, C. A. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6**, 149-172, doi:10.1002/(SICI)1098-1063(1996)6:2<149::AID-HIPO6>3.0.CO;2-K (1996).

142 Fiebelkorn, I. C., Pinsk, M. A. & Kastner, S. A Dynamic Interplay within the Frontoparietal Network Underlies Rhythmic Spatial Attention. *Neuron* **99**, 842-853 e848, doi:10.1016/j.neuron.2018.07.038 (2018).

143 Kleinfeld, D., Deschenes, M. & Ulanovsky, N. Whisking, Sniffing, and the Hippocampal theta-Rhythm: A Tale of Two Oscillators. *PLoS Biol* **14**, e1002385, doi:10.1371/journal.pbio.1002385 (2016).

144 Stewart, N., Hermens, F. & Matthews, W. J. Eye Movements in Risky Choice. *Journal of Behavioral Decision Making* **29**, 116-136, doi:10.1002/bdm.1854 (2016).

145 Hunt, L. T., Rutledge, R. B., Malalasekera, W. M., Kennerley, S. W. & Dolan, R. J. Approach-Induced Biases in Human Information Sampling. *PLoS Biol* **14**, e2000638, doi:10.1371/journal.pbio.2000638 (2016).

146 Kobayashi, K., Ravaioli, S., Baranes, A., Woodford, M. & Gottlieb, J. Diverse motives for human curiosity. *Nat Hum Behav* **3**, 587-595, doi:10.1038/s41562-019-0589-3 (2019).

147 Hayden, B. Y. Economic choice: the foraging perspective. *Current Opinion in Behavioral Sciences* **24**, 1-6, doi:10.1016/j.cobeha.2017.12.002 (2018).

148 Sweis, B. M. *et al.* Sensitivity to "sunk costs" in mice, rats, and humans. *Science* **361**, 178-181, doi:10.1126/science.aar8644 (2018).

149 Juechems, K. & Summerfield, C. Where Does Value Come From? *Trends Cogn Sci* **23**, 836-850, doi:10.1016/j.tics.2019.07.012 (2019).

150 Nilsson, D. E. Evolution: An Irresistibly Clear View of Land. *Curr Biol* **27**, R715-R717, doi:10.1016/j.cub.2017.05.082 (2017).

Funding acknowledgements

L.T.H. was supported by a Henry Dale Fellowship from the Royal Society and Wellcome (208789/Z/17/Z). N.D. was supported by NIDA R01DA038891, and NSF IIS-1822571, both part of the CRCNS program. M. A. M. was funded by NSF Brain Initiative ECCS-1835389. E. R. was supported by a Ch. and H. Schaller Foundation and the Boehringer Ingelheim Foundation grant 'Complex Systems'. E.P. and C.R.E.W. are supported by the French National Research Agency within the framework of the labex CORTEX ANR-11-LABX-0042 of Université de Lyon, and grant ANR-19-CE37-0008 NORAD and ANR-18-CE37-0016-01 PREDYCT. E.P. is employed by the *Centre National de la Recherche Scientifique*. J.S. is funded by a MRC Skills Development Fellowship (MR/N014448/1). N.K. is funded by a fellowship from the BBSRC (BB/R010803/1).