HHS Public Access

Author manuscript

Annu Rev Neurosci. Author manuscript; available in PMC 2026 June 01.

Planning in the brain: It's not what you think it is

Marcelo G. Mattar[1], **Nathaniel D. Daw**[2]

[1]Department of Psychology and Center for Neural Science, New York University, New York, NY, USA

[2]Princeton Neuroscience Institute and Department of Psychology, Princeton University, Princeton, New Jersey, USA

Abstract

The neuroscience of planning has long been analogized to search algorithms in artificial intelligence (AI), which simulate future actions to guide immediate choices. We argue that advances in both neuroscience and AI suggest that planning is better understood to encompass a broader class of computations where mental simulation supports learning, often well before a decision is needed. We review three neurocomputational mechanisms that illustrate this shift. First, hippocampal replay resembles search but also often occurs prospectively or offline, likely training downstream circuits rather than directly guiding choice. Second, temporally abstract representations, such as grid cells, can enable planning without iterative search. Third, metalearning may shape how prefrontal dynamics implement task-specific planning strategies, echoing how AI systems learn to adapt across contexts. This view recasts the brain's planning machinery as a family of learning processes that leverage simulations to build representations and strategies, with forward search as one special case.

1. Introduction

In the brain sciences, the study of planning—indeed, even the idea that planning is a discrete function to be studied—has long been driven by analogy to a class of algorithms from artificial intelligence (AI) known as tree search. Most tree search algorithms in AI, from the earliest chess and checkers players (Shannon 1950, Samuel 1959) to modern refinements with superhuman performance (Campbell et al. 2002, Silver et al. 2016), center around forward search through the "tree" of future board positions. Psychologists have long envisioned that a similar simulation-based search mechanism might underlie biological decision-making in many behavioral domains beyond board games, endowing humans and even animals with the capability to flexibly discover novel courses of action appropriate to whatever circumstance they find themselves in (Tolman 1948, Dickinson 1985, Daw et al. 2005). These psychological theories rested on two key ideas drawn from a stylized version of their AI counterpart: that the brain learns a world model (or cognitive map) to mentally simulate actions, and that this simulation unfolds as a forward search from the current situation at the time of the decision (Figure 1a). While this framework aligned well with gross behavioral capabilities, the details of how such an algorithm would run in the brain remained unclear, largely because such hypothetical search occurs covertly and is difficult to infer from behavior alone. This limitation motivated a more recent turn to

neural measurements in an attempt to more directly observe such model-based simulation and search.

Here, we revisit this evidence, exploring the neuroscience of planning in light of recent progress in AI. In doing so, we argue that the classical view of planning as forward search to guide the next choice is, at best, too narrow: a rather unrepresentative special case of a more general family of computations. Instead, we argue that planning is better understood as any type of mental simulation that enables learning, often to improve behavior further in the future. In turn, the mechanisms of planning are themselves fundamentally sculpted by learning in ways that go well beyond the stereotyped search algorithms derived from a simple analogy with AI. This argument is grounded in recent developments in both AI and neuroscience.

The current explosion of AI dramatically demonstrates that the fundamental engine of intelligent behavior is learning from data (Sutton 2019, Kaplan et al. 2020, Hoffmann et al. 2022). In light of this, the usefulness of simulation models lies not only in helping to select the next action but also in their ability to generate limitless, synthetic experience from which one can learn to improve many future choices. Such learning need not, and typically does not, happen at choice times: On the contrary, having a model frees the learner from entrainment to real-world events and enables learning that is faster, safer, and more scalable than learning from real-world interaction. Indeed, in modern AI—most obviously in the case of large language models, but even in apparently purer examples of game tree search such as AlphaGo—the benefits of model-generated data disproportionately accrue through precomputation and learning during pretraining, rather than through situation-specific search when a decision is faced (Silver et al. 2016, Hamrick et al. 2020, Ruoss et al. 2024).

In psychology and neuroscience, a similar theme is emerging in planning research, emphasizing the critical role of precomputation, often via learning from model-generated data. In what follows, we review this theme in light of three (not mutually exclusive) neurocomputational mechanisms. First, we consider planning via iterative search, most often associated with neural replay. This is the case closest to the classical picture, but both behavioral and neural experiments indicate that the same mechanisms are often engaged prospectively to compute appropriate responses well before a choice is needed (Wimmer & Shohamy 2012, Gershman et al. 2014, Momennejad et al. 2018, Liu et al. 2021, Miller et al. 2022, Nicholas et al. 2025) (Figure 1b).

Second, we consider planning without search. Converging lines of work suggest that world models can also be built to support long-run predictions, enabling flexible planning without iterative simulation (Dayan 1993) (Figure 1c). Neurally, these theories intersect with ideas about the function of the grid cells of the entorhinal cortex, which represent physical space (and other metric feature spaces) with periodic functions that capture distant relationships (Hafting et al. 2005, Stachenfeld et al. 2017). Such planning-specialized representations must, of course, be built for each task, though the emerging picture of this process looks rather different than the classical picture of learning a one-step world model for search (Whittington et al. 2020, Piray & Daw 2025).

Finally, such learning is one example of the critical role of metalearning in planning. In machine learning, this idea formalizes how an agent's strategy for within-task learning can be adapted (metalearned) by outer-loop training across an ensemble of tasks. This process is again most dramatically exemplified by large language models, whose powerful learning in context is enabled by pretraining of the model weights (Brown et al. 2020). Neurally, this idea has been used to suggest how activity dynamics in the prefrontal cortex (PFC) can be shaped to enable bespoke computations for particular tasks, including planning (Wang et al. 2018) (Figure 1d). Coming full circle, this puts a modern spin on venerable psychological ideas of classic planning being supported by working memory in PFC, though again, the more computational view emphasizes that this type of planning is not generic but instead shaped by task-specific adaptations.

2. Planning as learning from simulated experience

2.1. Planning as forward search

The most intuitive form of planning, which dominated early research in both AI and cognitive science, is online forward search from the current situation. Partially inspired by recordings of human chess masters who verbally reported considering sequences of future moves (De Groot 1946), early AI researchers formalized planning as a search through the tree of possible future states, evaluating each to find an optimal path to the goal (Shannon 1950, Newell & Simon 1956). To manage the combinatorial explosion of possibilities—where each additional step multiplies the number of states to consider—these algorithms employed heuristic functions that efficiently estimate a state's value without exhaustive evaluation, guiding search toward promising trajectories while pruning less favorable branches (Russell & Norvig 2010). This paradigm achieved iconic success in systems such as IBM's Deep Blue, which defeated world champion Garry Kasparov in 1997 by searching millions of chess positions per second, guided by sophisticated evaluation functions (Campbell et al. 2002).

Behavioral evidence has long suggested that humans evaluate actions through a similar, albeit vastly more selective, forward search process. The "think-aloud" experiments mentioned above indicated that expert chess players explore only a small, carefully chosen subset of possible moves rather than exhaustively considering all options (De Groot 1946). Later work began to characterize more precisely the specific strategies humans use to manage limited cognitive resources: By analyzing choice patterns in small, carefully designed toy problems, researchers concluded that people adaptively limit their search depth to a fixed horizon (Keramati et al. 2016) and prune branches of the decision tree that are unlikely to yield rewards, sometimes to the point of overlooking genuinely beneficial options (Huys et al. 2012, 2015). Large-scale computational modeling of behavior in a much more complex two-player game, four-in-a-row, indicated that in this setting, human choices and reaction times are well-explained by a best-first search algorithm employing simple feature-based heuristics to focus computation on the most promising moves (van Opheusden et al. 2023).

Notably, this study further highlights a critical dependence on precomputed knowledge to guide the search and fill in values for parts of the tree not searched. For four-in-a-row,

the model of human play relied on a static evaluator—a heuristic function that can quickly estimate a board's value based on built-in knowledge about the value of different board features. Although less often emphasized in cognitive science analogies, similar static evaluators are a ubiquitous component of AI game players, from the earliest checkers players (Samuel 1959) to AlphaGo (Silver et al. 2016). In all these cases, the evaluators' parameters must be somehow learned or precomputed, which in the case of AlphaGo (whose deep network evaluator was pretrained over millions of games of selfplay) is arguably the secret sauce primarily responsible for its superhuman performance. Learning to evaluate a novel situation without further search, then, is one way in which the success of human planning emerges not from search alone but from the interplay between online simulation and offline learning—a theme that, as we see below, extends far beyond the traditional picture of forward search at decision time. This process is also similar to, but goes beyond, simply learning the value of specific, individual situations or actions, which in neuroscience is widely viewed as a model of automatization or habits (Daw et al. 2005).

Early attempts to localize the neural substrates for planning via search implicate a wide network of brain regions, perhaps because the underlying computations implicate many different cognitive functions (among them working memory, attention, reward, and cognitive control). Two areas—PFC and hippocampus—are particularly implicated. Patients with frontal lobe damage show severe impairments on planning tasks even when they can articulate the rules, suggesting that the PFC is necessary for such tasks (Shallice 1982). Functional neuroimaging confirms that activity in the dorsolateral PFC (among other areas) scales with planning difficulty (Duncan et al. 1996, Unterrainer & Owen 2006). How these areas contribute mechanistically has been less clear (and a point to which we return in Section 4), but key ideas include working memory (where the term working emphasizes the need to manipulate, not just maintain, the tree of future possibilities) and associated executive and cognitive control (Baddeley 1992).

The hippocampus appears to be involved in mental simulation in ways that (at least in rodents) appear to provide a more direct window into the content of the simulations themselves. During spatial navigation, hippocampal place cells exhibit theta sequences—rapid, ordered activations that represent upcoming locations, which may support online lookahead akin to the forward search in AI systems, albeit only at relatively short range (Johnson & Redish 2007, Kay et al. 2020, Comrie et al. 2024). During pauses or rest, hippocampal activity includes sporadic sharpwave ripples (SWRs), within which replay events often form longer, extended trajectories that start at the animal's current location and project toward known goals (Pfeiffer & Foster 2013). These replayed trajectories can even represent novel paths around known obstacles and have been shown to predict subsequent behavior in a content-specific manner (Gupta et al. 2010, Singer et al. 2013, Widloski & Foster 2022). Despite the challenges of relating rodent navigation to human multistep decision-making, these results point to the intriguing possibility that hippocampal sequences reflect forward planning, producing simulated trajectories that guide navigation.

2.2. The critical role of precomputation

Although studies of human behavior in complex games support the view of forward search as a core component of human deliberation, they also highlight its critical dependence on precomputed knowledge. As discussed above, this knowledge can be expressed as a heuristic function or static evaluator, which was used even in the earliest AI systems (Samuel 1959). Even dynamic programming, a classic algorithm developed by Bellman (1966) in the same era, is often used to precompute and cache, or store, solutions to subproblems rather than repeatedly solving them online. The modern deep learning era has only amplified this principle: Following AlphaGo, MuZero learns world models purely from self-play without knowing the game rules, while recent work shows that the knowledge gained from search can be distilled entirely into neural networks that play chess at high levels without any search at test time (Schrittwieser et al. 2019, Ruoss et al. 2024).

The critical role of precomputation to planning is a clear example of Sutton's (2019) "Bitter Lesson": In AI, general-purpose methods that scale with data and computation ultimately triumph over clever, handcrafted algorithms. From this viewpoint, the true power of world models lies not in enabling a few steps of lookahead at decision time but in their capacity to generate virtually unlimited synthetic training data. Unlike real-world experience, which is slow, risky, and limited in quantity, simulated experience can be produced safely, quickly, and in massive parallel batches, fundamentally changing the economics of learning. This perspective was most clearly formalized in Sutton's (1991) Dyna architecture, which elegantly unifies planning and learning into a single computational framework. In Dyna, an agent interleaves real, online experience with imagined trajectories generated from its world model, using both to update the same value functions and policies through identical learning rules. The simulated trajectories are not discarded after contributing to action selection but instead serve as synthetic training data that produces lasting changes in the agent's policy, blurring the traditional boundary between planning and learning.

One issue that is less obvious in constrained settings like games, but increasingly evident for more general-purpose agents, is that the apparent efficiency of precomputation and caching can become maladaptive when the environment changes, as cached information may no longer be suitable to the current task or environment. This tension between flexibility and efficiency—closely related to the model-based versus model-free distinction in reinforcement learning—suggests that successful agents need both rapid cached responses and the ability to recompute when those caches become stale, a balance we see reflected in the brain's planning mechanisms (Daw et al. 2005).

2.3. Planning as learning in biological organisms

The computational view of precomputation and caching relates to a longstanding challenge in studying biological planning: Behavior alone can demonstrate that mental simulation occurs but offers little evidence about when it happens. Specifically, behavioral psychologists have long used the ability of animals to flexibly adapt to changes in contingencies (nimble replanning without additional trial-and-error learning) to argue that they are capable of planning. For instance, in the classic demonstration of latent learning discussed by Tolman (1948), rats first explored a maze without reward, with food being

subsequently introduced at a particular location. The rats' ability to quickly find efficient routes suggested that they formed and planned with a cognitive map of the environment.

Although Tolman (and many subsequent authors) generally assumed that this planning occurred via forward simulation when they first confronted the probe choice point, this behavior actually leaves open crucial questions about timing. Did the rats mentally simulate routes during initial exploration, at the moment reward was introduced, during the delay before testing, or only when confronting the choice point? Modern behavioral studies have begun to address this ambiguity via more sophisticated designs and additional observables. When humans face analogous replanning problems, their eye movements (Konovalov & Krajbich 2016), response times (Momennejad et al. 2017), and patterns of dual task interference (Gershman et al. 2014) all suggest the crucial computations occurred before the crucial replanning decision was faced.

Human neuroimaging has provided more direct evidence that planning-related computations often occur outside the moment of decision, revealing neural signatures of offline trajectory evaluation. Using multivariate pattern analysis and functional MRI, studies have shown that hippocampal and cortical activity patterns associated with specific locations or items reactivate in ways that bridge the gap between separately learned associations, much like in planning. However, although this activity does sometimes extend forward at choice time (Doll et al. 2015, Nicholas et al. 2025), it also occurs at other times before choices are faced, such as when rewards are first introduced (Wimmer & Shohamy 2012, Nicholas et al. 2025) and in quiet rest periods (Momennejad et al. 2018). Critically, the content of spontaneous reactivations at all these times predicts subsequent flexible choice behavior: For instance, participants whose rest-period activity more strongly represents particular state transitions subsequently navigate more efficiently through those transitions, indicating that these offline computations have lasting effects on decision making. Similar but more detailed results have been found with magnetoencephalography (Kurth-Nelson et al. 2015, Liu et al. 2021). With its millisecond temporal resolution, this technique has captured the rapid sequential reactivation of states that form coherent trajectories through learned task spaces. Again, in these cases, subsequent choice behavior is predicted by reactivations when rewards are first learned, rather than later when the choice is faced. This evidence suggests that much of the computational work happens not during deliberation at choice points, but in the seemingly idle moments between active decisions.

The idea of precomputation also invites a reinterpretation of sequential activity in the rodent hippocampus and helps resolve longstanding challenges in that field. As with the human results, although forward replay suggests a role in planning in the classical, online sense, replay also occurs in many patterns and circumstances suggestive of a more general role in precomputation, of which immediate planning would merely be a special case. First, hippocampal SWRs (i.e., long-range replay) occur only when the animal is stationary and do not occur when the animal is actively moving, arguably when a choice is most needed (Wilson & McNaughton 1994). Interestingly, replay is prominent during quiet rest and sleep, when it can represent remote locations far from the animal's current position. Second, the predominant patterns of replay are quite different from those presumably used for forward simulation. For example, upon encountering a reward, replay is observed predominantly

in a backward direction, proceeding from the animal's current location toward recently visited locations (Foster & Wilson 2006, Diba & Buzsáki 2008). Such backward sequences are more easily interpreted as supporting local and nonlocal credit assignment—a form of precomputation—than as informing the animal's immediate choice. Third, even when replay does occur at decision points, the timing and content of such sequences are not consistently linked to subsequent choices, and can instead even lag behind behavioral learning (Carey et al. 2019, Gillespie et al. 2021). This disconnect between decision-time replay and immediate choice suggests that hippocampal sequences may serve broader functions than moment-by-moment planning, supporting learning processes that unfold over longer timescales.

In light of the evidence above, replay has been recently proposed to implement a biological version of the Dyna framework—generating synthetic experience that trains downstream circuits without additional real-world interaction (Mattar & Daw 2018). On this view, replayed trajectories at different times (at choice time versus rest) and in different patterns (e.g., forward versus backward) all serve a common function in helping the animal connect actions to their distal consequences. But since this activity often occurs in advance of a choice, its effects are primarily mediated via learning. This view aligns with the Complementary Learning Systems theory, where SWRs coordinate hippocampal-cortical dialog that transfers information to long-term storage, affecting behavior hours or days later (Wilson & McNaughton 1994, McClelland et al. 1995, Káli & Dayan 2004). Critically, offline replay goes far beyond repeating past experience: It constructs novel trajectories by combining segments from different experiences, explores never-visited locations, and stitches together separate memories to discover hidden relationships (Gupta et al. 2010, Liu et al. 2021). The set of compositional computations supported by replay, accordingly, appear to support diverse forms of inference, such as counterfactual reasoning about alternative choices, abstract schema formation, and associative chaining (Liu et al. 2019, Kurth-Nelson et al. 2023, Schwartenbeck et al. 2023). The fact that these varied computations, many of which do not seem to involve iterative search, all share a common replay mechanism supports the idea that forward search is a special case of a more general computation. In particular, replay is an active generative process, using the world model to derive new knowledge and prepare for novel situations through precomputation, of which online planning is only a special case.

If simulated experience is a valuable resource for learning, a critical question becomes what to simulate—a decision that appears to be strategically controlled rather than left to chance. Both computational theory and neural evidence converge on the principle that replay is prioritized based on expected learning benefit: Sequences are preferentially selected based on their potential to improve future performance (Moore & Atkeson 1993, Mattar & Daw 2018). In rodents, replay frequency increases for trajectories leading to large rewards, those involving surprising outcomes, and paths connecting the animal's current location to known goals (Singer & Frank 2009, Ólafsdóttir et al. 2018). The existence of such prioritization mechanisms suggests that the brain not only generates synthetic experience but actively curates it, implementing a form of metacontrol over its own learning process—a theme we explore further in Section 4's discussion of metalearning. This strategic selection of what to simulate represents another form of precomputation, where the brain invests its

computational resources during rest to prepare for future challenges, guided by principles that themselves may be shaped by experience.

3. Planning without search

3.1 Successor representations and temporal abstraction

Classic search iterates a one-step dynamics model (e.g., a single chess move or one step on a spatial grid) to simulate longer trajectories, evaluating the states encountered along the way. One way to shortcut this cost is to aggregate over these steps and batch them into longer steps or groups of states. Such a strategy is generally known as temporal abstraction. One version of this strategy, which has deep roots in AI and sustained interest in neuroscience, is called the successor representation (SR) (Dayan 1993). The SR stores, for each state, the aggregated number of encounters with each other state likely to follow it over some time-discount horizon. From this, in turn, it is straightforward to compute the start state's long-run value in a single computation via a weighted sum over successor encounters.

This approach provides much of the flexibility of search, such as the ability to revalue and replan when goals change (by reweighting the successor states), but all at once, without iterative search at choice time. In this way, the SR and its variants can capture many types of flexible replanning behaviors that animals are capable of, without the requirement for online search (Russek et al. 2017). This strategy is another example of precomputation to facilitate later planning: The SR can be viewed, roughly, as a set of value functions corresponding to situations in which different states become goals, and learning an SR is thus like planning ahead for all such eventualities (Sagiv et al. 2025). Of course, there is no free lunch; batching the transition states involves caching assumptions about the intervening events (e.g., intermediate state dynamics and choices), which limits the model's ability to replan if these change. For this reason, the SR can also be viewed as an intermediate strategy between model-based and model-free learning, caching less aggressively than the latter. Nevertheless, in practice, variants of the method can work quite well, and a number of newer variants have been suggested to mitigate many of these caching concerns (Barreto et al. 2016, Piray & Daw 2021).

There is considerable evidence that the brain adopts a similar strategy (Gershman 2018). This includes evidence from reaction times and neural signatures of long-run expectancy in sequence prediction tasks (Garvert et al. 2017, Lynn et al. 2020, Wittkuhn et al. 2024, Kahn et al. 2025), and slips of action in choice tasks thought to reflect caching long-run chunks (Momennejad et al. 2017, Russek et al. 2021).

3.2 Grid cells and temporal abstraction

The primary hypothesis for how such a mechanism is implemented neurally is via the grid cells of the entorhinal cortex, which famously tile space periodically at a range of frequencies (Hafting et al. 2005). Intuitively, the grid cells represent possible future locations —periodically, rather than timestep by timestep—with low-frequency grid cells capturing long-distance predictions about the future location. More formally, in the open field, the grid cell population corresponds to a set of basis functions (roughly, the eigenvectors of the

transition dynamics) that, among many other things, can be linearly combined to express the SR (Stachenfeld et al. 2017). [The combination itself might be implicit downstream or, in one prominent model, is expressed in the place cell system of the hippocampus (Stachenfeld et al. 2017).] In turn, this means that planning (i.e., computing expected future value from some situation, which is itself linear in the SR's columns) amounts to a weighted readout from the grid cells. Apart from this specific RL computation, the grid cell representation can support a range of other related types of vector mathematics over space, potentially subserving path integration, vector-based navigation, and intuitive planning (Bush et al. 2015, Banino et al. 2018, Baram et al. 2018, Yu et al. 2021), and also has implications (discussed below) for how the SR is learned.

3.3 Beyond planning

We previously stressed that replay is applicable to problems that go beyond planning in the sense of sequential forward prediction. The same is true of grid cell–like methods. For the same reasons that grid cells are useful for extrapolating trajectories over space, these representations are equally useful for extrapolation over nonspatial multidimensional feature spaces. For instance, a line of work suggests that grid cells (or grid-like neural signatures in neuroimaging) can represent 2D metric feature spaces other than physical location (Constantinescu et al. 2016, Aronov et al. 2017).

One appealing application of this capacity is to multiattribute choice. For instance, in a pair of studies, monkeys learned symbolic representations for different levels of reward probability and magnitude (Bongioanni et al. 2021, Veselic et al. 2025). These two features, together, define a two-dimensional space of options, the understanding of which could (and in practice does) enable monkeys to accurately evaluate even novel options, that is, combinations of probability and magnitude not previously trained. This is another example of a task that requires generalizing to novel situations without iterative future simulation. Neuroimaging and, later, neurophysiological recordings suggest that the space of options is represented by a grid code, though interestingly, in ventromedial PFC rather than hippocampus. Similar results have been observed in human neuroimaging (Nitsch et al. 2024). Vector math over the grid code, analogous to that for spatial reasoning, could generalize option value over the space.

3.4 Planning with less search

Although we have stressed that temporally abstract representations like the SR can obviate search, they can also work in conjunction with search to extend its scope. An SR can itself be viewed as a world model with a particular predictive timescale (formally, given by its discount factor), which can then be further extended simply by iterative search over its predictions. The ability to trade off the depth of the world model against the depth of search [known as gamma models in machine learning (Janner et al. 2020)] is one way to balance the computational benefits of temporal abstraction against the inflexibility of long-range caching. The idea that timescales can be freely traded off between a deeper world model and deeper search also helps to address the lack of a single characteristic timescale for real-world dynamics (Sutton 1995). Interestingly, precisely this type of hybrid SR-search mechanism independently arose in descriptive models of human episodic memory, known as temporal

context models (Howard & Kahana 2002, Gershman et al. 2012, Zhou et al. 2024). While the temporal patterns this gives rise to at recall time seem arbitrary in experiments on word list learning, one view is that these effects are another window into the brain's mechanisms for mental simulation via temporal abstraction and search.

Another way to view how temporal abstraction supports search is that given the right representation—specifically, a predictive one—planning becomes trivial. Learning the SR is, from one perspective, learning a particular type of world model, but it is also learning a state representation that is optimized for long-run prediction. From the latter perspective, learning supports and shapes planning more fundamentally than in the classic picture—a point developed even further in the next section.

The ways in which the SR, and related temporally abstract maps, are learned speak to the richness of these interactions. First, one of the ways in which replay goes beyond search is that it is likely used to build SRs (i.e., precompute policies for different possible future goals) as well as evaluate current courses of action (Russek et al. 2017, Wittkuhn et al. 2024, Sagiv et al. 2025). The view of the grid cells as a basis for temporally abstract dynamics prediction also gives rise to richer accounts of learning maps by recombining basis functions, including regularization and compositional assembly (Stachenfeld et al. 2017, Piray & Daw 2025). Other lines of work envision that the dynamics model is learned over latent states and can be reused flexibly across environments (Whittington et al. 2020, 2022; George et al. 2021).

4. Metalearning: learning how to plan

Much of what we have discussed so far concerns how planning is shaped by learning, for example, of task-appropriate static evaluators, long-run representations, or replay prioritizations. In short, planning is not a fixed algorithm that the brain executes in a uniform way across contexts. Instead, it is itself an adaptive process, shaped by experience to better meet the demands of new tasks.

In AI, these types of refinement can all be more generally formalized as instances of metalearning (Hospedales et al. 2022): adapting one's algorithms to a family of tasks. Metalearning interleaves two levels of learning: Task instances are repeated many times (an outer loop, like multiple chess games), and within each of these instances, behavior is governed by some inner-loop learning/planning algorithm. Learning in the outer loop refines and specializes learning in the inner loop. In artificial agents, metalearning dramatically improves planning efficiency by exploiting structural regularities across tasks (Ritter et al. 2018, Wang et al. 2018, Botvinick et al. 2019).

In the brain, metalearning (or more specifically, meta–reinforcement learning) has been proposed to be implemented in PFC, with sustained, recurrent activity dynamics implementing inner-loop learning, and their learning rules sculpted in task-specific ways by outer-loop, across-episode synaptic plasticity (Wang et al. 2018). In this model, recurrent neural networks (putatively in PFC) trained on related tasks learn shared statistical structure through slow synaptic changes, while fast within-task adaptation emerges from the

network's dynamic activity patterns. The network's hidden state serves as an active, adaptive working memory, maintaining task-specific information to enable rapid within-episode policy adjustment without additional synaptic modification (Sun et al. 2026).

When applied to problems requiring planning, such metalearning processes can learn task-specialized planning algorithms. For instance, decoding analyses suggest that a deep network static evaluator trained on chess implicitly learns some forward search through its activation dynamics (Jenner et al. 2024). Closer to the biological setting, when networks are trained on the two-step task—a standard assay for model-based planning in humans and animals—they spontaneously develop activity patterns that produce planning-equivalent behavior via their activation dynamics (Wang et al. 2018). This idea has been leveraged to explain why in the same task, in rats, trial-by-trial planning-equivalent choice updates are insensitive to manipulations of dopamine (a neurochemical thought to be involved in reinforcement learning at the level of synaptic plasticity, i.e., the outer loop here): perhaps because such inner-loop learning is instead accomplished in PFC activity (Akam et al. 2015, Blanco-Pozo et al. 2024).

PFC activity (Akam et al. 2015, Blanco-Pozo et al. 2024). These ideas suggest a more specific computational implementation for the classic but vague idea that working memory (thought to be implemented by sustained activation dynamics in PFC) plays a critical role in maintaining and manipulating information during planning tasks (Shallice 1982, Duncan et al. 1996). In this view, the particular recurrent dynamics needed to maintain and update the latent state information required to guide a planning policy are task specific, but the PFC can learn them through metalearning. These sculpted dynamics allow a single neural substrate to implement different planning algorithms depending on the context, explaining not only that the PFC is involved in planning (as lesion and imaging studies indicate) but also how it can adapt its planning policy while leveraging long-term knowledge.

A somewhat different aspect of metalearning is determining how a learned world model is used in planning: which computations to perform and when. Evidence suggests that humans adapt their planning depth, pruning heuristics, and subgoal selection to the demands of the task in resource-rational fashion, balancing decision quality against cognitive cost (Lieder & Griffiths 2017, Callaway et al. 2022). One key example is the metalearning of offline simulation policies, where the replay of some trajectories is prioritized over others to optimize future decisions. This view explains the diversity of replay sequences —forward, backward, remote—and the prevalence of each pattern in different behavioral scenarios as the result of a metacontroller that always replays the most useful experience at each moment (Mattar & Daw 2018). Recent modeling work extends this idea to explain how this metacontroller is obtained by treating planning itself as a metalearned policy of when to initiate mental simulations, or rollouts (Jensen et al. 2024). In these models, agents learn to balance the benefit of further thinking against the cost of delaying action, producing deliberation patterns that match behavioral data and bear qualitative similarities to hippocampal replay sequences.

The modeling work above suggests a link between meta-learned control over simulation and the hippocampal mechanisms for trajectory generation discussed earlier, highlighting

how metalearning shapes not only what we know about the world but also how we deploy cognitive resources to search within it. Ultimately, this metalearning perspective reveals that what appears as planning in the brain—whether forward search, backward replay, or SRs—may all be different manifestations of metalearned strategies for leveraging world models, with classical forward search being just one special case among many possible algorithms the brain has learned to implement.

5. Discussion

We have discussed a number of ways in which the classic scenario of planning—forward search at decision time—is better seen as a special case (and an idiosyncratic one) of more general computations for flexible choice. Instead, we argue that the key principle is precomputation: Learning from model-generated data (such as self-play or offline replay) sets the stage for later choices but also shapes the mechanisms for later in-the-moment planning. Metalearning offers a general computational formalism for these processes and a modern, mechanistically specific way of understanding how activity dynamics in PFC are crucial for implementing flexible choice.

Two important caveats represent key opportunities for future work. First, flexible, planning-like computations are likely not unitary. We have discussed several different candidate neural mechanisms, and the extent to which they are distinct versus (as we have at times suggested) integrated or interacting remains largely unknown. Another, likely somewhat distinct, aspect of planning in humans is more explicitly linguistically mediated reasoning. This has some resonance with PFC and working memory (e.g., articulatory loops) but is also clearly different from fast hippocampal replay dynamics. A related point is that we have somewhat breezily combined evidence from a variety of species, tasks, and measurement methodologies. How special is spatial navigation? How different are toy laboratory decision tasks from more elaborate, realistic ones? Does decoded reinstatement in human neuroimaging reflect anything related to hippocampal replay in animals? We know little about any of these questions.

Second, and even more importantly, we have mentioned at various points how precomputation is a double-edged sword. Changing tasks, contingencies, or goals may invalidate previously computed and stored information, which is indeed part of why in-the-moment planning is thought to be useful. Classic work has viewed this trade-off through a very simple dichotomy between unrealistically comprehensive model-based replanning and maximally narrow model-free automaticity (Dickinson 1985, Daw et al. 2005, Keramati et al. 2011). Even this has shed suggestive light on resource-rational automaticity and potential dysfunction, for example, in disorders of compulsion (Gillan et al. 2016). But how these trade-offs play out in richer and subtler circumstances where aspects of planning itself are shaped by metalearning are likely equally important but almost completely unexplored.

Acknowledgements

This work was supported by the National Institute of Mental Health of the National Institutes of Health under grant R01MH136875 (N.D.D.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Akam T, Costa R, Dayan P. 2015. Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. PLoS Comput. Biol 11(12):e1004648

Aronov D, Nevers R, Tank D. 2017. Mapping of a non-spatial dimension by the hippocampal/entorhinal circuit. Nature. 543:719–22 [PubMed: 28358077]

Baddeley A. 1992. Working memory. Science. 255(5044):556–59 [PubMed: 1736359]

Banino A, Barry C, Uria B, Blundell C, Lillicrap T, et al. 2018. Vector-based navigation using grid-like representations in artificial agents. Nature. 557:429–33 [PubMed: 29743670]

Baram AB, Muller TH, Whittington JCR, Behrens TEJ. 2018. Intuitive planning: global navigation through cognitive maps based on grid-like codes. bioRxiv

Barreto A, Dabney W, Munos R, Hunt JJ, Schaul T, et al. 2016. Successor features for transfer in reinforcement learning. Adv. Neural Inf. Process. Syst abs/1606.05312:

Bellman R. 1966. Dynamic programming. Science. 153(3731):34–37 [PubMed: 17730601]

Blanco-Pozo M, Akam T, Walton ME. 2024. Dopamine-independent effect of rewards on choices through hidden-state inference. Nat. Neurosci 27:286–97 [PubMed: 38216649]

Bongioanni A, Folloni D, Verhagen L, Sallet J, Klein-Flügge MC, Rushworth MFS. 2021. Activation and disruption of a neural mechanism for novel choice in monkeys. Nature. 591(7849):270–74 [PubMed: 33408410]

Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D. 2019. Reinforcement learning, fast and slow. Trends Cogn. Sci 23(5):408–22 [PubMed: 31003893]

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, et al. 2020. Language Models are Few-Shot Learners. Neural Inf Process Syst. abs/2005.14165:1877–1901

Bush D, Barry C, Manson D, Burgess N. 2015. Using grid cells for navigation. Neuron. 87(3):507–20 [PubMed: 26247860]

Callaway F, van Opheusden B, Gul S, Das P, Krueger PM, et al. 2022. Rational use of cognitive resources in human planning. Nat. Hum. Behav 6(8):1112–25 [PubMed: 35484209]

Campbell M, Hoane AJ Jr, Hsu F-H. 2002. Deep blue. Artif. Intell 134(1–2):57–83

Carey AA, Tanaka Y, van der Meer MAA. 2019. Reward revaluation biases hippocampal replay content away from the preferred outcome. Nat. Neurosci 22:1450–59 [PubMed: 31427771]

Constantinescu AO, O'Reilly JX, Behrens TEJ. 2016. Organizing conceptual knowledge in humans with a gridlike code. Science. 352(6292):1464–68 [PubMed: 27313047]

Daw N, Niv Y, Dayan P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci 8:1704–11 [PubMed: 16286932]

Dayan P. 1993. Improving generalization for temporal difference learning: The successor representation. Neural Comput. 5:613–24

De Groot AD. 1946. Het denken van den schaker [Thought and choice in chess]. Amsterdam: Noord Hollandsche

Diba K, Buzsáki G. 2008. Hippocampal network dynamics constrain the time lag between pyramidal cells across modified environments. J. Neurosci 28:13448–56 [PubMed: 19074018]

Dickinson A. 1985. Actions and habits: the development of behavioural autonomy. Philos. Trans. R. Soc. Lond 308(1135):67–78

Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND. 2015. Model-based choices involve prospective neural activity. Nat. Neurosci 18(5):767–72 [PubMed: 25799041]

Duncan J, Emslie H, Williams P, Johnson R, Freer C. 1996. Intelligence and the frontal lobe: the organization of goal-directed behavior. Cogn. Psychol 30(3):257–303 [PubMed: 8660786]

Foster DJ, Wilson MA. 2006. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature. 440(7084):680–83 [PubMed: 16474382]

Garvert MM, Dolan RJ, Behrens TE. 2017. A map of abstract relational knowledge in the human hippocampal-entorhinal cortex. Elife. 6:e17086

George D, Rikhye R, Gothoskar N, Guntupalli JS, Dedieu A, Lázaro-Gredilla M. 2021. Clone-structured graph representations enable flexible learning and vicarious evaluation of cognitive maps. Nat. Commun 12(1):2392 [PubMed: 33888694]

Gershman S. 2018. The successor representation: Its computational logic and neural substrates. J. Neurosci 38:7193–7200 [PubMed: 30006364]

Gershman SJ, Markman AB, Otto AR. 2014. Retrospective revaluation in sequential decision making: a tale of two systems. J. Exp. Psychol. Gen 143(1):182–94 [PubMed: 23230992]

Gershman SJ, Moore CD, Todd MT, Norman KA, Sederberg PB. 2012. The successor representation and temporal context. Neural Comput. 24(6):1553–68 [PubMed: 22364500]

Gillan CM, Kosinski M, Whelan R, Phelps EA, Daw ND. 2016. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. Elife. 5:

Gillespie AK, Astudillo Maya DA, Denovellis EL, Liu DF, Kastner DB, et al. 2021. Hippocampal replay reflects specific past experiences rather than a plan for subsequent choice. Neuron. 109(19):3149–63.e6 [PubMed: 34450026]

Gupta AS, van der Meer MAA, Touretzky DS, Redish AD. 2010. Hippocampal replay is not a simple function of experience. Neuron. 65(5):695–705 [PubMed: 20223204]

Hafting T, Fyhn M, Molden S, Moser M, Moser E. 2005. Microstructure of a spatial map in the entorhinal cortex. Nature. 436(7052):801–6 [PubMed: 15965463]

Hamrick JB, Friesen AL, Behbahani F, Guez A, Viola F, et al. 2020. On the role of planning in model-based deep reinforcement learning. arXiv [cs.AI]

Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, et al. 2022. Training Compute-Optimal Large Language Models. arXiv [cs.CL]

Hospedales T, Antoniou A, Micaelli P, Storkey A. 2022. Meta-learning in neural networks: A survey. IEEE Trans. Pattern Anal. Mach. Intell 44(9):5149–69 [PubMed: 33974543]

Howard MW, Kahana MJ. 2002. A distributed representation of temporal context. J. Math. Psychol 46(3):269–99

Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP. 2012. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput. Biol 8(3):e1002410

Huys QJM, Lally N, Faulkner P, Eshel N, Seifritz E, et al. 2015. Interplay of approximate planning strategies. Proc. Natl. Acad. Sci. U. S. A 112(10):3098–3103 [PubMed: 25675480]

Janner M, Mordatch I, Levine S. 2020. Generative temporal difference learning for infinite-horizon prediction. arXiv [cs.LG]

Jenner E, Kapur S, Georgiev V, Allen C, Emmons S, Russell S. 2024. Evidence of learned look-ahead in a chess-playing neural network. Work. Pap.

Jensen KT, Hennequin G, Mattar M. 2024. A recurrent network model of planning explains hippocampal replay and human behavior. Nature Neuroscience. 27:1340–48 [PubMed: 38849521]

Johnson A, Redish A. 2007. Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci 27:12176–89 [PubMed: 17989284]

Kahn AE, Bassett DS, Daw ND. 2025. Trial-by-trial learning of successor representations in human behavior. bioRxiv. 2024.11. 07.622528

Káli S, Dayan P. 2004. Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. Nat. Neurosci 7(3):286–94 [PubMed: 14983183]

Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, et al. 2020. Scaling laws for neural language models. ArXiv. abs/2001.08361:

Kay K, Chung JE, Sosa M, Schor JS, Karlsson MP, et al. 2020. Constant sub-second cycling between representations of possible futures in the hippocampus. Cell. 180(3):552–67.e25 [PubMed: 32004462]

Keramati M, Dezfouli A, Piray P. 2011. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput. Biol 7(5):e1002055

Keramati M, Smittenaar P, Dolan RJ, Dayan P. 2016. Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. Proc. Natl. Acad. Sci. U. S. A 113(45):12868–73 [PubMed: 27791110]

Konovalov A, Krajbich I. 2016. Gaze data reveal distinct choice processes underlying model-based and model-free reinforcement learning. Nat. Commun 7(1):12438 [PubMed: 27511383]

Kurth-Nelson Z, Barnes G, Sejdinovic D, Dolan R, Dayan P. 2015. Temporal structure in associative retrieval. Elife. 4:

Kurth-Nelson Z, Behrens T, Wayne G, Miller K, Luettgau L, et al. 2023. Replay and compositional computation. Neuron. 111(4):454–69 [PubMed: 36640765]

Lieder F, Griffiths T. 2017. Strategy Selection as Rational Metareasoning. Psychol. Rev 124:762–94 [PubMed: 29106268]

Liu Y, Dolan RJ, Kurth-Nelson Z, Behrens TEJ. 2019. Human replay spontaneously reorganizes experience. Cell. 178(3):640–52.e14 [PubMed: 31280961]

Liu Y, Mattar MG, Behrens TEJ, Daw ND, Dolan RJ. 2021. Experience replay is associated with efficient nonlocal learning. Science. 372(6544):eabf1357

Lynn CW, Kahn AE, Nyema N, Bassett D. 2020. Abstract representations of events arise from mental errors in learning and memory. Nat. Commun 11(1):2313 [PubMed: 32385232]

Mattar MG, Daw ND. 2018. Prioritized memory access explains planning and hippocampal replay. Nat. Neurosci 21(11):1609–17 [PubMed: 30349103]

McClelland JL, McNaughton BL, O'Reilly RC. 1995. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychol. Rev 102(3):419–57 [PubMed: 7624455]

Miller KJ, Botvinick MM, Brody CD. 2022. Value representations in the rodent orbitofrontal cortex drive learning, not choice. Elife. 11:e64575

Momennejad I, Otto AR, Daw ND, Norman KA. 2018. Offline replay supports planning in human reinforcement learning. Elife. 7:

Momennejad I, Russek EM, Cheong JH, Botvinick MM, Daw ND, Gershman SJ. 2017. The successor representation in human reinforcement learning. Nat. Hum. Behav 1(9):680–92 [PubMed: 31024137]

Moore AW, Atkeson CG. 1993. Prioritized sweeping: Reinforcement learning with less data and less time. Mach. Learn 13(1):103–30

Newell A, Simon H. 1956. The logic theory machine-A complex information processing system. IRE Trans. Inf. Theory 2:61–79

Nicholas J, Daw ND, Shohamy D. 2025. Proactive and reactive construction of memory-based preferences. Nat. Commun 16:

Nitsch A, Garvert MM, Bellmund JLS, Schuck NW, Doeller CF. 2024. Grid-like entorhinal representation of an abstract value space during prospective decision making. Nat. Commun 15(1):1198 [PubMed: 38336756]

Ólafsdóttir HF, Bush D, Barry C. 2018. The role of hippocampal replay in memory and planning. Curr. Biol 28(1):R37–50 [PubMed: 29316421]

Pfeiffer BE, Foster DJ. 2013. Hippocampal place-cell sequences depict future paths to remembered goals. Nature. 497(7447):74–79 [PubMed: 23594744]

Piray P, Daw ND. 2021. Linear reinforcement learning in planning, grid fields, and cognitive control. Nat. Commun 12(1):4942 [PubMed: 34400622]

Piray P, Daw ND. 2025. Reconciling flexibility and efficiency: medial entorhinal cortex represents a compositional cognitive map. Nat. Commun 16(1):7444 [PubMed: 40796544]

Ritter S, Wang J, Kurth-Nelson Z, Jayakumar S, Blundell C, et al. 2018. Been there, done that: Meta-learning with episodic recall. . 4354–63

Ruoss A, Del'etang G'egoire, Medapati S, Grau-Moya J, Li WK, et al. 2024. Amortized planning with large-scale transformers: A case study on chess. Neural Inf Process Syst. 37:65765–90

Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. 2017. Predictive representations can link model-based reinforcement learning to model-free mechanisms. PLoS Comput. Biol 13(9):e1005768

Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. 2021. Neural evidence for the successor representation in choice evaluation. bioRxiv

Russell SJ, Norvig P. 2010. Artificial intelligence: a modern approach

Sagiv Y, Akam T, Witten IB, Daw ND. 2025. Between planning and map-building: Prioritizing replay when future goals are uncertain. Neuron

Samuel AL. 1959. Some studies in machine learning using the game of checkers. IBM J. Res. Dev

Schrittwieser J, Antonoglou I, Hubert T, Simonyan K, Sifre L, et al. 2019. Mastering Atari, Go, chess and shogi by planning with a learned model. Nature. 588:604–9

Schwartenbeck P, Baram A, Liu Y, Mark S, Muller T, et al. 2023. Generative replay underlies compositional inference in the hippocampal-prefrontal circuit. Cell. 186(22):4885–97.e14 [PubMed: 37804832]

Shallice T. 1982. Specific impairments of planning. Philos. Trans. R. Soc. Lond. B Biol. Sci 298(1089):199–209 [PubMed: 6125971]

Shannon CE. 1950. Programming a computer for playing chess. Lond. Edinb. Dublin Philos. Mag. J. Sci 41(314):256–75

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, et al. 2016. Mastering the game of Go with deep neural networks and tree search. Nature. 529(7587):484–89 [PubMed: 26819042]

Singer AC, Carr MF, Karlsson MP, Frank LM. 2013. Hippocampal SWR activity predicts correct decisions during the initial learning of an alternation task. Neuron. 77(6):1163–73 [PubMed: 23522050]

Singer AC, Frank LM. 2009. Rewarded outcomes enhance reactivation of experience in the hippocampus. Neuron. 64(6):910–21 [PubMed: 20064396]

Stachenfeld KL, Botvinick M, Gershman S. 2017. The hippocampus as a predictive map. Nat. Neurosci 20(11):1643–53 [PubMed: 28967910]

Sutton R. 2019. The Bitter Lesson. https://heartyhaven.github.io/files/bitter_lesson.pdf

Sutton RS. 1991. Dyna, an integrated architecture for learning, planning, and reacting. SIGART Newsl. 2(4):160–63

Sutton RS. 1995. TD models: Modeling the world at a mixture of time scales. In Machine Learning Proceedings 1995, pp. 531–39. Elsevier

Tolman E. 1948. Cognitive maps in rats and men. Psychol. Rev 55(4):189–208 [PubMed: 18870876]

Unterrainer JM, Owen AM. 2006. Planning and problem solving: from neuropsychology to functional neuroimaging. J. Physiol. Paris 99(4–6):308–17 [PubMed: 16750617]

van Opheusden B, Kuperwajs I, Galbiati G, Bnaya Z, Li Y, Ma W. 2023. Expertise increases planning depth in human gameplay. Nature. 618(7967):1000–1005 [PubMed: 37258667]

Veselic S, Muller TH, Gutierrez E, Behrens TEJ, Hunt LT, et al. 2025. A cognitive map for value-guided choice in the ventromedial prefrontal cortex. Cell. 188(12):3259–73.e22 [PubMed: 40262608]

Wang JX, Kurth-Nelson Z, Kumaran D, Tirumala D, Soyer H, et al. 2018. Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci 21(6):860–68 [PubMed: 29760527]

Whittington JCR, McCaffary D, Bakermans JJW, Behrens TEJ. 2022. How to build a cognitive map. Nat. Neurosci 25(10):1257–72 [PubMed: 36163284]

Whittington JCR, Muller TH, Mark S, Chen G, Barry C, et al. 2020. The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. Cell. 183(5):1249–63.e23 [PubMed: 33181068]

Widloski J, Foster DJ. 2022. Flexible rerouting of hippocampal replay sequences around changing barriers in the absence of global place field remapping. Neuron. 110(9):1547–58.e8 [PubMed: 35180390]

Wilson MA, McNaughton BL. 1994. Reactivation of hippocampal ensemble memories during sleep. Science. 265(5172):676–79 [PubMed: 8036517]

Wilson RC, Takahashi YK, Schoenbaum G, Niv Y. 2014. Orbitofrontal cortex as a cognitive map of task space. Neuron. 81(2):267–79 [PubMed: 24462094]

Wimmer GE, Shohamy D. 2012. Preference by association: how memory mechanisms in the hippocampus bias decisions. Science. 338(6104):270–73 [PubMed: 23066083]

Wittkuhn L, Krippner LM, Koch C, Schuck NW. 2022. Replay in human visual cortex is linked to the formation of successor representations and independent of consciousness. bioRxiv

Yu C, Behrens TEJ, Burgess N. 2020. Prediction and generalisation over directed actions by grid cells. arXiv [q-bio.NC]

Zhou C, Talmi D, Daw N, Mattar M. 2024. Episodic retrieval for model-based evaluation in sequential decision tasks. Psychol. Rev 132(1):18–49 [PubMed: 39869686]

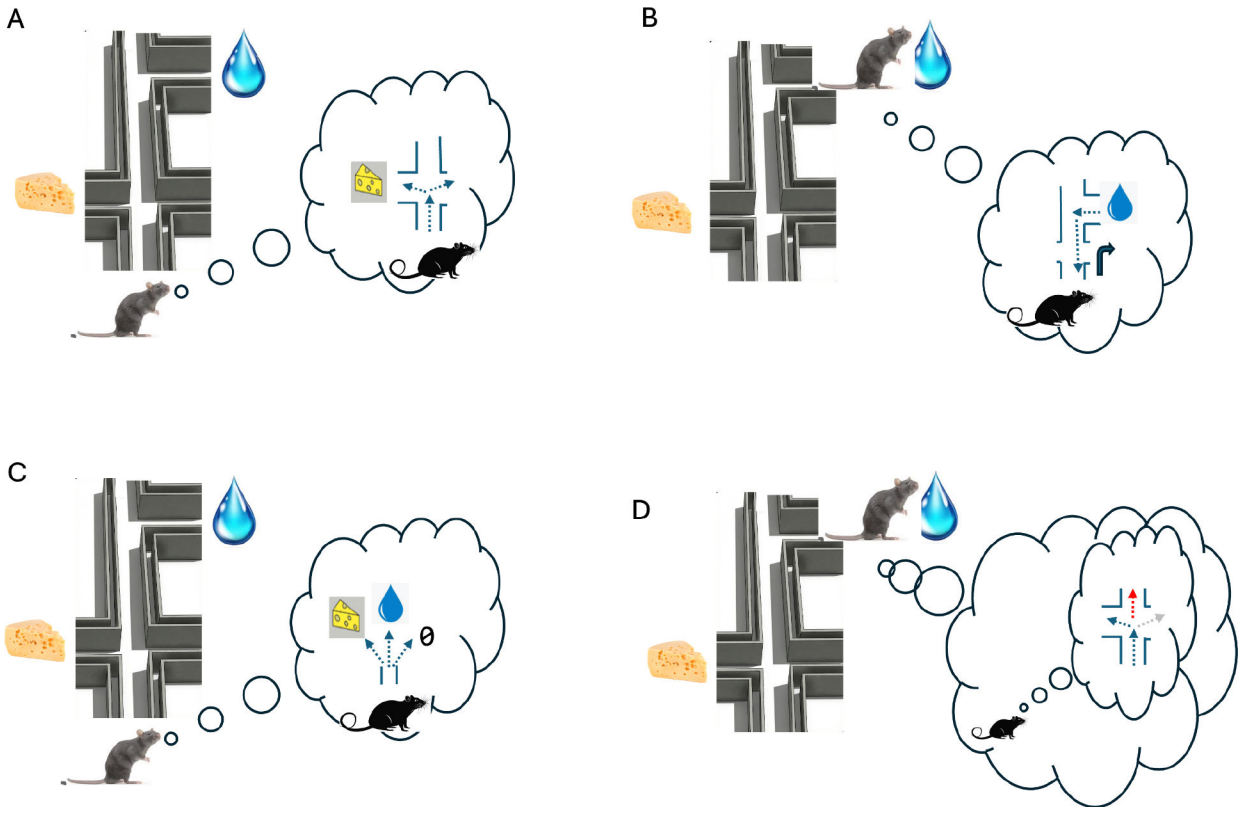

Figure 1:

Four views of planning in the brain. Each panel illustrates a distinct mechanism by which the brain can support flexible, goal-directed behavior, using the common scenario of a rat navigating a maze to find food or shelter.

(A) Online forward search. At a decision point, the agent simulates future trajectories by iterating a one-step world model, evaluating branches to select the best action.

(B) Pre-planning via offline simulation. Before a decision is needed, the agent generates simulated trajectories (forward, backward, or remote) that train downstream value representations, with effects mediated through learning rather than direct action selection.

(C) Planning without iterative search. Temporally abstract representations such as the successor representation aggregate long-run predictions, enabling rapid option evaluation in a single computation rather than through step-by-step simulation.

(D) Metalearning: Across task episodes, learning shapes recurrent dynamics and control policies to implement task-specialized planning strategies, determining how and when each type of simulation is deployed.

Solid paths indicate executed behavior; dashed paths indicate simulated/replayed transitions; goal icons denote reward locations.