Neuron

CellPress
OPEN ACCESS

Perspective

Planning in the brain

Marcelo G. Mattar[1,2,*] and Máté Lengyel[2,3]
[1]Department of Cognitive Science, University of California, San Diego, San Diego, CA, USA
[2]Computational and Biological Learning Lab, Department of Engineering, University of Cambridge, Cambridge, UK
[3]Center for Cognitive Computation, Department of Cognitive Science, Central European University, Budapest, Hungary
*Correspondence: mmattar@ucsd.edu
https://doi.org/10.1016/j.neuron.2021.12.018

SUMMARY

Recent breakthroughs in artificial intelligence (AI) have enabled machines to plan in tasks previously thought to be uniquely human. Meanwhile, the planning algorithms implemented by the brain itself remain largely unknown. Here, we review neural and behavioral data in sequential decision-making tasks that elucidate the ways in which the brain does—and does not—plan. To systematically review available biological data, we create a taxonomy of planning algorithms by summarizing the relevant design choices for such algorithms in AI. Across species, recording techniques, and task paradigms, we find converging evidence that the brain represents future states consistent with a class of planning algorithms within our taxonomy—focused, depth-limited, and serial. However, we argue that current data are insufficient for addressing more detailed algorithmic questions. We propose a new approach leveraging AI advances to drive experiments that can adjudicate between competing candidate algorithms.

INTRODUCTION

Life is full of choices: from the barely noticeable, such as selecting the right muscle contractions to reach for a glass, through the mundane, such as picking the best wine for an occasion, to the momentous, such as choosing a partner for life. Making all those choices is difficult because their consequences tend to only pan out over time, often well after we have made the choice. Thus, the result of any one choice will typically depend on a whole sequence of choices we have made on the way: a sequence of motor commands for grasping the glass; a sequence of turns for navigating to the wine shop; and a sequence of decisions between going on a date and proposing to get married. Despite the challenges of sequential decision making, humans and other animals are often able to perform impressively well in such tasks. How we achieve this performance, however, is poorly understood.

A common approach for making sequential decisions is by "planning," a process that considers actions and their sequential interdependence in terms of the desirability of their outcomes. By forecasting the long-term consequences of candidate actions, planning allows agents to flexibly adapt their behavior in response to changes both in the environment and in their goals. For this reason, planning is a fundamental component of intelligent behavior in both biological and artificial agents. However, in practice, there is a price to pay: planning exactly and exhaustively for all eventualities may require vast resources (time, memory, computational power; see curse of dimensionality; Bellman, 1957a). Thus, the key question is: how to perform planning *efficiently* under realistic resource limitations without making undue sacrifices on performance and flexibility—for the brain must be using such efficient planning strategies.

Although planning has been a classical subject of investigation in psychology, it remains one of the most elusive cognitive processes at the neural level. A main factor limiting progress has been the relatively impoverished nature of tasks typically used to study the neuroscience of decision making. Those tasks often focus on only a single step of decision, unlike the tasks we typically encounter in our lives, which are, instead, sequential in nature. Of course, the simplicity of single-step tasks has the advantage of greater tractability. Indeed, in single-step tasks, planning amounts to the selection of a single action in terms of the desirability of its immediate outcome. Leveraging this simplicity, neuroscientists have been able to dissociate the neural bases of planned (goal-directed) versus reflexive (habitual) behavior (Balleine and Dickinson, 1998). Single-step tasks have also allowed much progress in characterizing the computations and neural signals associated with related processes, such as perceptual decision making (Gold and Shadlen, 2007) and economic choice (Padoa-Schioppa and Assad, 2006). However, single-step tasks have limited ecological validity, resting on the assumption that agents and their actions have no effect in the world beyond the immediate reward they obtain. As a result, these tasks do not tap into some of the key challenges of real-life planning, which often unfolds over longer time horizons. Unfortunately, deploying more complex and cognitively demanding paradigms risks losing the interpretability of behavioral and neural measurements (Rust and Movshon, 2005). We suggest that the best way to solve this impasse, as is often the case, is to have formally defined computational hypotheses. In the case of planning, these hypotheses take the form of candidate-planning algorithms that perform competently on tasks of realistic complexity and offer specific, experimentally testable predictions (LaValle, 2006).

Please cite this article in press as: Mattar and Lengyel, Planning in the brain, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.12.018

Fortunately, there now exist many examples of such algorithms. Recent advances in artificial intelligence (AI) have given us important clues as to the key ingredients of planning algorithms that work well in practice (Silver et al., 2017a, 2018; Hafner et al., 2019a; Schrittwieser et al., 2020). These algorithms now achieve near- or even super-human performance on a number of difficult, large-scale sequential decision-making tasks, rendering them relevant for understanding how the brain may implement planning on tasks of similar complexity. While these algorithms remain unable to perform well in the largely unconstrained settings in which humans operate, we argue that they might give hints as to what is really important for superior performance. Our reasoning is that each planning algorithm developed in AI can be thought of as a recipe for converting goals and knowledge of the world into actions. As such, these algorithms share a number of commonalities in the way they produce an output, but they also differ in a number of important ways related to the richness of the prior knowledge they employ, how they prioritize different steps of computation, and how they incorporate the results of these computations into future decisions. Collectively, the various algorithms developed in AI span a sizable space of the different ways in which planning can be accomplished in principle.

Critically, the usefulness of AI algorithms for understanding biological planning does not rest on the assumption that the exact same algorithms are used by artificial and biological organisms. Indeed, there are known fundamental differences between the two. For instance, while the chess-playing system Deep Blue was able to defeat Garry Kasparov 3.5–2.5, it achieved this performance by evaluating 200,000,000 positions per second (Campbell et al., 2002), whereas Kasparov achieved a similar performance, presumably evaluating only a handful of positions per second (de Groot, 1978). Nonetheless, we argue that it is the fundamental design decisions and consequent algorithmic motifs that are critical for developing efficient planning algorithms in AI that should also apply to the algorithms that the brain might implement. In other words, by specifying a number of ways in which planning can be accomplished *in principle*, we suggest that AI algorithms can also help neuroscientists formalize how planning can be achieved *in practice* by the brain, which in turn facilitates the generation of clear hypotheses that can then be tested experimentally. A similar symbiosis between AI and cognitive neuroscience has, in fact, already been highly fruitful in the discovery of the core algorithms underlying reward-based learning in the brain (Schultz et al., 1997), leading to insights into the neural basis of habitual behavior and its relationship to goal-directed control (Daw et al., 2005). What we propose, then, is to leverage this approach to planning algorithms more broadly (Daw and Dayan, 2014).

Therefore, in this perspective, we advocate for a closer symbiosis between the fields of neuroscience and AI to shed light on how the brain performs planning. We start by reviewing the essential components of planning and laying out a map for the most relevant dimensions of the space of planning algorithms through the lessons learned from AI. Equipped with this map, we present, for each dimension, the most relevant algorithms in AI while discussing how existing behavioral and neuroscientific data from humans and animals can be used to rule in and out candidate algorithms. We find that while this research has identified a number of relevant brain regions and some general algorithmic motifs, the existing data remain largely unable to discriminate various algorithmic details. Finally, we use this systematic approach to suggest important areas for further research and propose an approach that could narrow the space of potential planning algorithms that the brain may use.

REFLEXIVE VERSUS PLANNING AGENTS

In this article, we define planning as the process of selecting an action or sequence of actions in terms of the desirability of their outcomes. Note that this definition includes, as a special case, the selection of actions in terms of their immediate, one-step outcomes (in which case the label "goal-directed behavior" is commonly used; Balleine and Dickinson, 1998). At the other extreme, when actions are chosen based on the desirability of distant outcomes in the future, our definition also includes the special case when even the motivational state determining the desirability of those outcomes is predicted to be different from the agent's current motivational state. This latter case has proved particularly useful for establishing planning behavior in studies of prospective simulations in non-human animals (Suddendorf and Corballis, 1997; Raby et al., 2007).

According to our definition, planning involves two fundamental computations: (1) estimating the outcomes of one's actions and (2) assessing the utility of such outcomes. These computations are some of the most fundamental ingredients of intelligence, and as such, planning has always been a central topic in AI. Indeed, some of the first AI programs ever written—the Logic Theorist and the General Problem Solver developed by Newell and Simon in the 1950s—were, in essence, planning systems (Newell and Simon, 1956; Newell et al., 1959). Since then, planning has emerged as one of the major AI methods for action selection in autonomous agents. Autonomous agents, biological or artificial, continuously perform the perception-action loop (Russell and Norvig, 2010) (Figures 1A and 1B), perceiving and assessing their current situation in the environment (the "state") and then selecting and executing an appropriate action, thereby modifying their future situation in the environment (a "state transition"). The goal of the agent when selecting actions is defined either in terms of a specific state that needs to be achieved (e.g., having our wine glass full) or, more generally, as the maximization of a graded quantity ("reward") to be accumulated over time (e.g., the hedonistic value of consecutive sips from our wine) (Sutton and Barto, 2018).

Planning-based methods for autonomous behavior contrast with, and sometimes complement, "reflex"-based methods (Figure 1A). Reflex-based methods are those in which the agent directly maps states to actions (Russell and Norvig, 2010)—an approach typically associated with classical (model-free) forms of reinforcement learning (Sutton and Barto, 2018). The mapping from states to actions, called "policy," can be represented directly or indirectly. When it is represented indirectly, the agent instead represents the long-term desirability ("value") of states, such that the policy amounts to choosing the action in each state that leads the agent to a higher value state. Policies or values can be either hardwired (as in most Pavlovian behaviors; but see

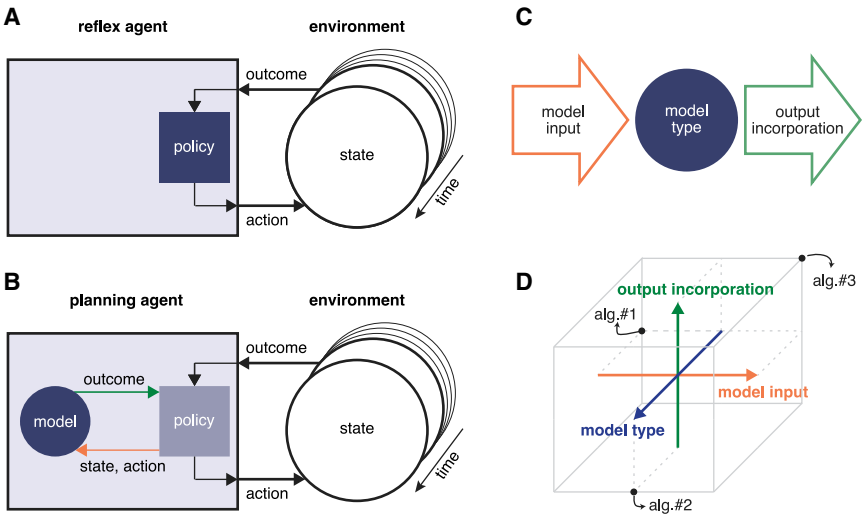

Figure 1. Components of a planning algorithm

(A and B) Action-perception loop: the agent (left) perceives the outcome (sensory observations and rewards, including the attainment of goals; center top) generated by the current state of the environment (right) and executes an action (center bottom), resulting in the environment transitioning to a new state (note arrow of time in bottom right). (A) A reflexive agent selects actions based on a stored policy, which maps the current state (inferred from observations of the environment) to an output action. This policy is gradually updated based on interactions with the environment.

(B) A planning agent selects actions based on the desirability of their expected consequences. The agent stores an internal model, which predicts the outcomes (rewards and state transitions) resulting from any given state and action. The internal model is learned based on interactions with the environment. The policy, then, can be updated (temporarily or permanently) based on interactions with the internal model via the process of "planning." Thus, planning amounts to querying the model to evaluate candidate actions to inform current or future decisions. The orange and green arrows, respectively, indicate inputs and outputs to the internal model.

(C) A closer look at the three main ingredients of planning algorithms (based on B): the model (center, purple) that determines the type of output returned to a query; model input (left, orange) representing the query itself; and output incorporation (right, green) representing how the model outputs are used to inform future behavior.

(D) The space of planning algorithms spanning the three dimensions corresponding to the main ingredients listed in (C). Different planning algorithms correspond to different combinations of design choices for these ingredients, and as such, to different "points" in this space (black dots show hypothetical examples).

Derman et al., 2018) or learned incrementally via interaction with the environment (e.g., when developing habitual behavior). However, reflex agents are often inflexible: in dynamic environments where action outcomes or goals change over time, the same policy that was relevant and thus stored at some point may not apply anymore. Planning agents can succeed in such dynamic situations by directly considering the consequences of each action, as long as this knowledge is current (Figure 1B).

The distinction between reflex and planning agents has many parallels in behavioral and neural sciences, though under different labels, such as automatic versus deliberate control, habitual versus goal-directed behavior, model-free versus model-based decision making, and type I versus type II reasoning (Anderson, 1982; Adams and Dickinson, 1981; Dickinson and Balleine, 2002; Kahneman, 2011; Daw et al., 2005; Dayan, 2009). A common observation across all of these fields is that humans and other animals exhibit signatures of *both* reflexive and planning-based computations in their behavior. Thus, a useful (if not essential) first step in studying planning algorithms in the brain is to make sure that subjects' behavior in a task reflects the latter rather than the former. A straightforward way of achieving this is to use tasks unlikely to be solved by reflexive behavior, such as tasks that are new to subjects and are explained only by verbal descriptions or a map so that subjects need to solve them without having experienced them previously. An example of such tasks is the Tower of Hanoi, a mathematical puzzle in which the objective is to move a stack of disks from a starting position to a goal position while respecting a set of constraints on such moves. In this task, humans are often able to select promising actions even the very first time a state is encountered (before a policy could be learned from experience), suggesting that the behavior is unlikely to be reflexive (Kotovsky et al., 1985).

In non-human animals, the study of planning is complicated by the fact that such animals are usually unable to learn how to solve stereotypical planning tasks without extensive training, making it difficult to rule out an automatized, reflexive strategy. Thus, a more nuanced approach is necessary to dissect reflexive and planning-based computations. The predominant paradigm for assessing whether behavior is reflexive (habitual) or planned (goal-directed) involves manipulating the two fundamental computations underlying planning, namely, degradation of action-outcome contingencies (affecting the estimation of action outcomes) and devaluation of rewards (affecting the utility of such outcomes). If a given behavior is sensitive to such manipulations, it can be inferred to be goal directed, akin to that expressed by a planning agent. If subjects are not sensitive to such manipulations, the behavior is considered to be habitual, as would be expected from a reflexive agent (Balleine and Dickinson, 1998).

BRAIN AREAS INVOLVED IN PLANNING

To study how planning is implemented by the brain, it is useful to know which brain regions are most relevant to the computations underlying planning. However, different planning tasks may present the brain with fundamentally different computational challenges (e.g., motor control versus navigation, dealing with continuous versus discrete state spaces while planning on short versus long-time horizons, respectively) and may consequently engage different sensory, cognitive, and motor systems. Accordingly, different planning tasks will often engage different brain regions. Note that the involvement of a brain region in the execution of a planning task does not automatically imply that this region mediates planning; conversely, even brain areas that seem to have a critical involvement in planning may also contribute to processes other than planning. Nonetheless,

some regions are consistently recruited in the vast majority of planning tasks (Dolan and Dayan, 2013), and have been explicitly distinguished from those critical for reflexive behavior. These areas are our primary focus here.

First, the prefrontal cortex (PFC) has been identified as the region most directly associated with planning (Goel and Grafman, 1995; Unterrainer and Owen, 2006). Without a functioning prefrontal cortex, humans appear to be largely "stimulus bound," executing behaviors in a reflexive manner and neglecting any potential goal (Duncan et al., 1996). When performing sequential decision tasks, patients with frontal lesions often fail to achieve task goals (Penfield and Evans, 1935) and express difficulties in planning and rule-guided behavior (Shallice, 1982; Owen, 1997; Hoshi et al., 2000) despite understanding the requirements of the task. In converging neuroimaging results, tasks that require planning, such as the Tower of Hanoi or its variants, seem to activate a number of frontal regions in humans (Goel and Grafman, 1995; Fincham et al., 2002), and PFC activation increases with planning difficulty (Anderson et al., 2005; Owen, 1997). Note that the involvement of PFC in planning and decision making does not contradict another large body of work implicating PFC in the active maintenance of information for subsequent use (Curtis and D'Esposito, 2003), as a crucial component of planning is the active maintenance of current goals in working memory (Miller and Cohen, 2001).

In addition to PFC, another brain structure deeply involved in planning is the striatum, a subcortical structure in the basal ganglia that receives input from various areas of the brain, including the frontal cortex. While the circuitry involving the striatum is most commonly associated with the less-flexible, habitual system, this region is also involved in flexible, planning-based strategies. Indeed, both lesion and recording studies in rodents suggest that the dorsomedial portion of the striatum is essential for goal-directed behavior (Albin et al., 1989; Obeso et al., 2009; Ragozzino et al., 2002; Yin et al., 2005a; Balleine et al., 2007). For instance, lesions to dorsomedial striatum seem to abolish the sensitivity of behavior to outcome devaluation while leaving reflexive behavior largely intact (Yin et al., 2005b). In primates, a similar subdivision between the caudate nucleus and the putamen maps onto the rodent dorsomedial and dorsolateral striatum, with the caudate being the most closely associated with goal-directed behavior and the putamen with motor preparation, learning, and execution (Simon and Daw, 2011; Wunderlich et al., 2012).

While the approaches and results described above have been instrumental in delineating the conditions under which a behavior is due to reflexive or planning-based computations, and the importance of the PFC and dorsomedial striatum in mediating the latter, they are not adequate for identifying the specific algorithm(s) the brain uses for planning. To address this more granular level of description, two more ingredients are needed. First, sufficiently rich behavioral paradigms need to be employed. In particular, tasks with a single step of decision (often used in animal studies) are usually not ideal, as all planning algorithms predict the same response, i.e., the selection of the action that leads to the largest reward. Instead, tasks should ideally involve multiple steps of decisions for which different planning algorithms predict different behavioral and neural signatures. Therefore, in the following, we will focus mainly on empirical data obtained using such sequential decision tasks. The second necessary ingredient is the systematic understanding of how different planning algorithms map onto such behavioral and neural signatures. For this, the classical approach would be to focus on a select subset of specific algorithms and test them via model fits (e.g., van Opheusden et al., 2017). Here, we take a more systematic approach and identify the fundamental design decisions and consequent algorithmic motifs underlying essentially all planning algorithms. This allows us to build a comprehensive "map" of planning algorithms and look for empirical evidence for the main dimensions of algorithmic variability (see also Daw 2012). The rest of this paper develops such a map and critically evaluates what we do and do not know about where the planning algorithms of the brain are located in this map.

INTERNAL MODELS IN PLANNING

To develop an algorithmic map of planning, we first need to understand its basic ingredients. The most critical ingredient, used in all planning algorithms, is an internal model of the environment: the agent's representation of how the environment responds to the agent's actions. When the agent interrogates the model with a particular action in a particular state, the model produces a prediction of the outcome—the resulting state and reward, or whether a goal is achieved (Figure 1B). Using the internal model, planning agents can learn from counterfactuals ("if I performed action X, outcome Y might happen"), inferring what action sequences are most appropriate without having to try each one out in the environment (Sutton and Barto, 2018). This is particularly useful when the actions being considered are costly or may lead to catastrophic consequences.

The performance of a planning algorithm can be only as good as the internal model it uses. An incorrect internal model can give rise to biased beliefs about action outcomes, which, in turn, may lead to suboptimal or even pathological behavior (Talvitie, 2017). Thus, interacting with the world is essential for learning and improving the internal model itself. When the model is learned by interacting with the world, each experience can be used to improve the model, which then, indirectly, can inform multiple future decisions. In this way, planning allows the agent to make fuller use of a given amount of experience than other, e.g., reflexive, strategies (Sutton and Barto, 2018). The learning of an internal model is an interesting and challenging AI problem on its own right (Ha and Schmidhuber, 2018; Schrittwieser et al., 2020), and a major theme in the field of model-based reinforcement learning (Sutton and Barto, 2018). While model learning has far-reaching implications for cognitive neuroscience (Gläscher et al., 2010; Behrens et al., 2018; Fiser et al., 2010), it is beyond the scope of this perspective.

In biological organisms, the internal model has been an object of study for decades (Craik, 1943). In general, it may be used to infer and predict environmental states at many different levels of granularity, from low-level sensory attributes to high-level objects and scene-descriptors (Koblinger et al., 2021). Accordingly, several brain areas may be involved in implementing the brain's internal model of the environment, including even primary sensory areas (Berkes et al., 2011). Nevertheless, planning will

usually depend primarily on the higher-level variables inferred by the internal model (we rarely need to make decisions depending on the activation of specific pixels in our retina). In line with this, brain areas specifically associated with representing these higher-level variables largely overlap with those required for planning and take up a sizable fraction of the mammalian brain.

The internal model is sometimes equated to the concept of a "cognitive map," a representation of one's spatial environment akin to the type of information obtainable from a map (Tolman, 1948). The encoding, as well as learning, of such a cognitive map is most commonly attributed to the hippocampus and the surrounding medial temporal lobe. Indeed, in rodents, the hippocampus is causally involved in certain types of planning behavior (Miller et al., 2017), and units in the ventral hippocampus, a region which is strongly connected to those supporting reward processing, mediate a form of goal-oriented search in mice (Ruediger et al., 2012). The hippocampus's role in configural learning (O'Reilly and Rudy, 2001) and spatial learning (Doeller and Burgess, 2008) also points to its involvement in representing internal models. Most prominently, the hippocampus has a critical role in representing one's location in the environment, both spatial (O'Keefe and Nadel, 1978) and otherwise (Aronov et al., 2017; Behrens et al., 2018). In turn, this representation supports memory and guides future action (Shohamy and Daw, 2015).

The hippocampus is also commonly associated with episodic memory (Scoville and Milner, 1957). Theoretical work suggests that the episodic memory system can contribute to planning when one needs to retrieve a contextually appropriate response (or a previously rewarded action, or sequence of actions) encoded only once or a few times, or whenever one needs to combine multiple experiences obtained at different times to build a hypothetical scenario (Lengyel and Dayan, 2008; Shohamy and Daw, 2015; Gershman and Daw, 2017; Mattar and Daw, 2018). According to these proposals, the hippocampus encodes an episodic type of internal model where, for each state and action, one or more memories of the resulting outcomes are retrieved. Behavioral and neuroimaging results support the involvement of the hippocampus in event-based decisions—human choices are influenced by individual experiences made in similar contexts (Bornstein et al., 2017), an effect accompanied by hippocampal activation and reduced in patients with hippocampal lesions (Bornstein and Norman, 2017; Vikbladh et al., 2019).

Besides the hippocampus, prefrontal regions are also thought to encode (parts of) the internal model. For instance, signals related to expected reward have been identified in the primate orbitofrontal cortex (OFC) (Padoa-Schioppa and Assad, 2006; Valentin et al., 2007), a subregion of the PFC. Activity in OFC neurons is largely determined by an expected outcome (Schultz et al., 2000; Wallis and Miller, 2003; Roesch and Olson, 2004), particularly when the outcome is determined by the animal's own choice (Padoa-Schioppa and Assad, 2006; Rudebeck and Murray, 2014; Ballesta et al., 2020). Accordingly, lesioning or inactivating OFC impairs behavior relying on the inference of outcomes in rodents (Gallagher et al., 1999; Jones et al., 2012; Gremel and Costa, 2013) and non-human primates (Rudebeck et al., 2013). Moreover, the OFC is involved in representing not only rewarding outcomes but also expected future stimuli (Sadacca et al., 2018; Pauli et al., 2019) and the task space more generally (Wilson et al., 2014). Interestingly, anticipatory representations in OFC disappear after hippocampal lesions (Ramus et al., 2007), suggesting that the OFC receives input from the hippocampus related to internal models. The neighboring medial PFC also seems to code for the expected reward associated with chosen actions (Luk and Wallis, 2013). Together, the OFC and medial PFC precisely represent the types of information required for planning: the state and reward expected to result from a certain action. Note, however, that this does not mean that these are the only regions involved in planning, as the internal model, while critical, is but a single component of the planning process. A brain area not implementing an internal model may still contribute to planning through other processes, such as providing inputs to, and evaluating the outputs of the internal model.

In summary, an internal model of the environment is the central component of any planning algorithm. However, while all planning algorithms make use of an internal model, each individual algorithm may do so in a different way. Therefore, we propose to use the internal model as an anchor with respect to which we define the dimensions of the space of planning algorithms (Figure 1C). In particular, we propose that planning algorithms vary primarily along three dimensions (Figure 1D): (1) the type of internal model used; (2) the input with which the internal model is interrogated, i.e., which actions or action sequences are evaluated and in which order; and (3) how the output of the internal model is used to inform future actions. In the following, we expand on these three dimensions, reviewing along the way several of the best-known planning algorithms in AI (Figure 2), including classical as well as some more recently developed state-of-the-art algorithms, and how they fit in these dimensions, as well as how they relate to the relevant neural and behavioral data.

DESIGN CHOICE 1: INTERNAL MODEL TYPE

The first major dimension along which planning algorithms differ is the type of internal model they use. The simplest types of internal models are those suitable for deterministic environments, allowing the agent to predict with certainty the state of its environment after any sequence of actions (Figure 3A). These models are most commonly used in the branch of AI known as "classical planning," which describes planning problems involving deterministic action outcomes and a unique, known initial state (Figure 2). Classical planning algorithms were common in the early days of AI and are well suited to solving tasks such as route planning in mazes, logical board games (e.g., chess and Go), or puzzles (e.g., Rubik's cube and the Tower of Hanoi). Formally, given an input state and an action, a deterministic model returns the resulting state and reward. If the goal of the agent is to reach a specific state (e.g., a single color on each face of Rubik's cube), the model also informs the agent whether that goal is achieved. In such cases, a plan (sometimes also called "solution") is a sequence of actions that takes the agent from its current situation to a goal state, and the process of finding solutions is called "search" (Russell and Norvig, 2010; Korf, 1987). Because deterministic models specify the outcomes of each action with certainty, the planned sequence of

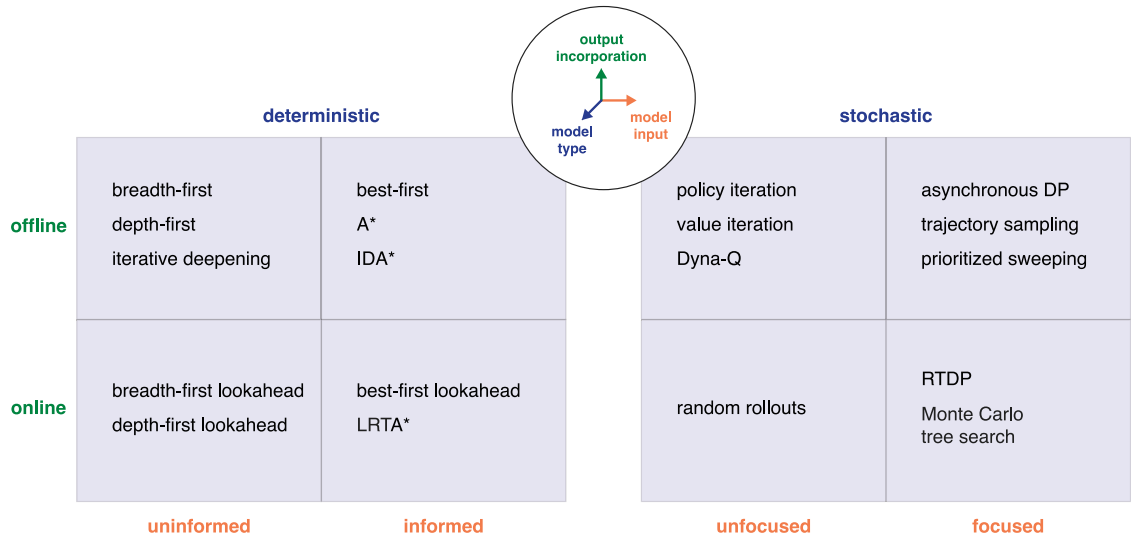

Figure 2. Mapping the space of planning algorithms in AI

Each example algorithm is mapped onto the space of algorithms spanned by three dimensions (Figure 1D, see also inset at top center), determining their main design choices: (1) the type of internal model used (deterministic versus stochastic; purple); (2) the input with which the internal model is interrogated (e.g., uninformed versus informed; orange); and (3) the way the output of the internal model is used to inform future actions (online versus offline; green). References: breadth-first search (Moore, 1959); depth-first search (Nilsson, 2014); iterative deepening (Korf, 1985); best-first search (Dechter and Pearl, 1985); A* (Hart et al., 1968); iterative deepening A* (IDA*; Korf, 1985); lookahead algorithms (Geffner, 2013); learning real-time A* (LRTA*; Korf, 1990); policy iteration (Howard, 1960); value iteration (Bellman, 1957b); Dyna-Q (Sutton, 1991); prioritized sweeping (Moore and Atkeson, 1993); trajectory sampling (Barto et al., 1995); asynchronous dynamic programing (asynchronous DP; Bertsekas, 1982); random rollouts (Tesauro and Galperin, 1996); real-time dynamic programing (RTDP; Barto et al., 1995); Monte Carlo tree search (Coulom, 2007).

actions can be executed "with eyes closed," without having to consider inputs on the way (a process also known as "open-loop control" in optimal control).

Stochastic internal models (often formalized in the framework of Markov decision processes, MDPs) are more expressive than deterministic models, capturing the agent's uncertainty about how the environment will respond to its actions (Bellman, 1957b; Sutton and Barto, 2018; Figure 3B). This uncertainty can arise either due to fundamentally stochastic processes in the environment (aleatory uncertainty, e.g., a coin flip) or due to partial knowledge of the environment (epistemic uncertainty, e.g., the unknown outcome of a lever press when tried for the first time), referred to as "risk" and "uncertainty," respectively, in economic decision making. Thus, stochastic models allow agents to plan in much more general and realistic settings, such as those involving noisy, random, or unknown physical environments. Formally, given an input state and action, stochastic models output either a single sample of the resulting outcome ("sample models") or the probabilities with which a set of outcomes might occur ("distribution models"). In either case, because it is impossible for the agent to know in advance the exact effects of an action, it is no longer possible to rely on open-loop control as in classical planning. Therefore, in this setting, it is often not helpful to pre-compute a sequence of actions. Instead, what needs to be computed is the best action at any particular state, a policy, which thus gives the agent more flexibility to choose each future action depending on future, yet unknown, outcomes ("closed-loop" or "optimal feedback control"). Examples of tasks where planning requires a stochastic model and closed-loop control are board games involving a chance component, such as dice rolls, route planning with unpredictable traffic, and most other realistic tasks.

Naturally, humans and other animals are able to consider stochastic outcomes when planning (Daw et al., 2011; Keramati et al., 2016; Miller et al., 2017; Akam et al., 2021; Miranda et al., 2020). Studies on the control of eye and arm movements have provided particularly striking demonstrations of stochastic internal (so-called forward) models underlying motor planning (Harris and Wolpert, 1998; Todorov and Jordan, 2002). Thus, planning in the brain must be closed loop, at least when necessary (an assumption typically incorporated in computational models of planning). Nevertheless, in some tasks with deterministic outcomes, neural data suggest that entire sequences of actions may be planned in advance. For instance, when monkeys are asked to plan a sequence of actions to achieve a goal, recordings from the lateral PFC during the preparatory phase of a sequential decision task contain information about the entire planned sequence of actions (Averbeck et al., 2002; Mushiake et al., 2006), even when the planned sequence involves mistakes (Averbeck et al., 2002). This suggests that monkeys deploy open-loop control at least in some tasks, which could be plausibly implemented by a deterministic model.

Given the important consequences of the type of internal model for narrowing the space of possible planning algorithms (Figure 1C), it is surprising how little direct evidence there is in favor of either deterministic or stochastic models. Part of this difficulty arises from the fact that the alignment between open- versus closed-loop and deterministic versus stochastic models is not perfect. For instance, a deterministic model may be used to replan in every time step. Similarly, even if suboptimal, a

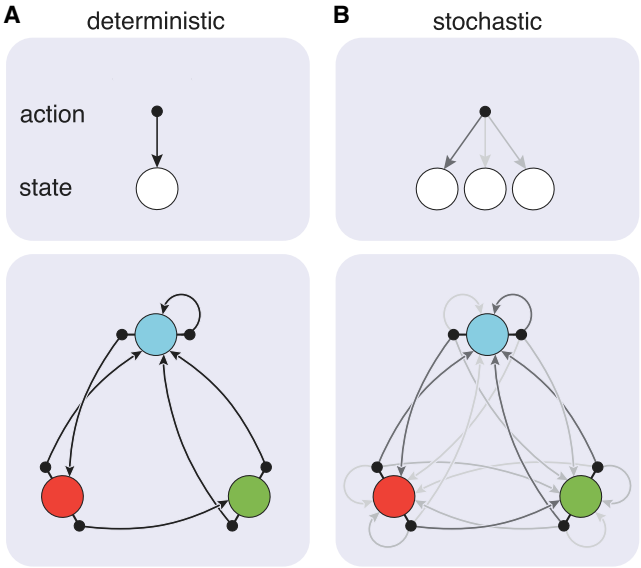

Figure 3. Types of internal models

Large white or colored circles denote states, and small black circles denote actions. Top parts of both panels show an example isolated action and the resulting states; bottom panels show a simple environment with three states and the actions available in each, causing transitions between them.

(A) In a deterministic model, outcomes can be predicted with certainty—in particular, one action always leads to the same state. By interrogating such a model with a particular action in a particular state, the agent can predict with certainty the resulting state and reward, or whether a goal is achieved.

(B) A stochastic model captures the agent's uncertainty about how the environment responds to its actions (in particular, one action may lead to many states with different probabilities, represented as different shades of gray here). By interrogating such a model with a particular action in a particular state, the agent can obtain either a single sample of the resulting outcome or the probabilities with which a set of outcomes might occur.

stochastic model may be used to commit to a sequence of actions and produce open-loop behavior (either because a sample model is used or because an action sequence is evaluated at once as a chunk, as we discuss below). Thus, examining only if behavior is open loop or closed loop is not sufficient for inferring the type of internal model used. To obtain the required evidence, one should not only use tasks with stochastic transitions, which can only be faithfully captured by a stochastic internal model but also particularly tasks in which open- and closed-loop planning differ in the values they compute and thus the actions they favor (for an example, see Friedrich and Lengyel, 2016). This would then allow distinguishing between these planning strategies (and conversely, between types of internal models) by measuring the extent to which the actions or values of each is reflected in overt behavior or correlated with neural activity, respectively.

In addition to the deterministic versus stochastic distinction, internal models can also vary in terms of the timescale on which their output is expressed. For example, the action of "donating to charity" can have an immediate (negative) outcome for one's bank account and a delayed (positive) outcome for one's level of happiness. However, choosing the "right" timescale for predictions is non-trivial. The successor representation (SR) offers a principled solution to this problem by actually side stepping the need to commit to one particular timescale and instead providing blended predictions including all timescales (with larger weights given to imminent versus distant outcomes). Such a representation allows multi-step planning tasks to be solved using a single-step algorithm, greatly reducing the computational complexity of planning (Dayan, 1993). The SR has found some support in behavioral studies with humans (Momennejad et al., 2017; Russek et al., 2017) and has also been argued to explain distortions in the response of hippocampal place cells that arise after extensive experience (Stachenfeld et al., 2017). Whereas planning with the SR is theoretically simple, re-planning can be sometimes challenging because of the fact that predictions of longer-horizon outcomes implicitly depend on the agent's previous policy. This inconvenience can be corrected by allowing the previous policy to exert only a weak bias on the inferred action values (Piray and Daw, 2021).

Finally, internal models can also vary in terms of how much states are abstracted away from raw observations. This is particularly relevant in problems with partial observability, where states and observations might not coincide. For instance, a planning agent in a video-game task can use states expressed directly in terms of a vector of pixel values or expressed in the more abstract terms of the physical parameters of objects (Wahlström et al., 2015; Hafner et al., 2019b). Recent work has proposed that internal models in human planning might use an abstraction that merges "microscopic" states (i.e., states closer to raw observations) when they have similar future consequences (Lehnert et al., 2020).

DESIGN CHOICE 2: ORDER OF COMPUTATION (MODEL INPUT)

The second major way in which planning algorithms differ is in terms of which actions are evaluated and in what order. That is, once the form of the internal model is determined, one must specify the inputs with which the model is interrogated (Figure 1C). The iterative selection of actions to evaluate amounts to computing a decision tree from the internal model. A decision tree has an initial state as the root node, actions represented as branches, and the states reachable from the initial state, according to the internal model, represented as leaf nodes (Figure 4A). Each leaf node can then be recursively expanded according to a strategy specific to the algorithm.

Exhaustive planning—evaluating all branches (action sequences) of the decision tree—is impractical in all but the simplest settings, as it typically requires vast amounts of computational resources (imagine evaluating all possible outcomes in chess, i.e., all board configurations). And even in settings where that might be possible, it is often important that the inputs to the model—the actions or action sequences to be evaluated—are ordered appropriately, so that the best outcome can be achieved with minimal amounts of time and computation. In the decision tree framework, this prioritization is expressed in terms of how the algorithm chooses the root node first and the leaf nodes to be expanded subsequently (Figure 4A). A fundamental distinction along this dimension is that between "uninformed" and "informed" algorithms, according to whether or not they make these choices

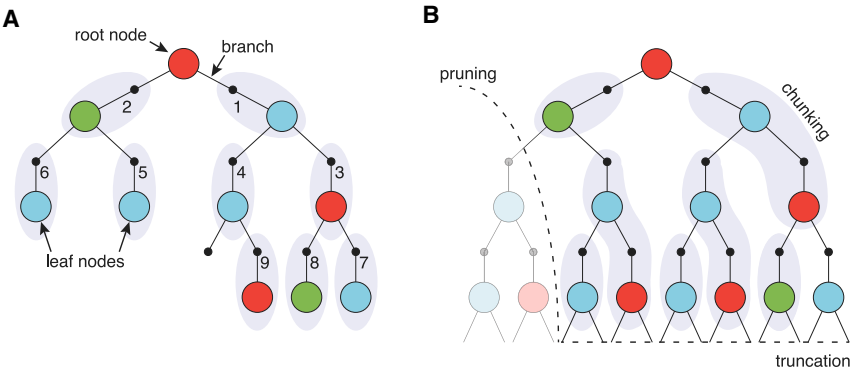

Figure 4. Expansion of a decision tree

(A) Starting from a root node, a decision tree is expanded by iteratively interrogating the internal model to infer the outcomes of any action sequence. For illustration, the expansion of a deterministic internal model (Figure 3A) is shown. Shaded ovals indicate the action being expanded, and numbers indicate the order of expansion (here, representing a breadth-first algorithm). Actions (small black circles) available in each state (large colored circles) form branches (sub-trees). The deepest unexpanded nodes are called leaf nodes.

(B) Different strategies can be used to prioritize the computations involved in expanding the tree. Three example strategies are shown (see also main text): pruning, whereby unpromising branches of the tree are ignored; truncation, whereby the tree is only expanded up to a maximum depth (a maximum number of actions in a sequence); and chunking, whereby multiple actions are clustered into an "option" (action-state sequences joined by shaded background) that can be evaluated as if it were a single action.

based on estimating how favorable a particular branch of the decision tree is (in terms of cumulative rewards or costs).

If the expansion of nodes of the decision tree is not guided by estimates of favorability, planning is uninformed (or unfocused) and amounts to systematically expanding nodes of the decision tree according to some predetermined schedule or strategy. For example, uninformed algorithms may start from the current state and iteratively expand the shallowest unexpanded node (a strategy known as breadth-first search) or alternatively expand the deepest unexpanded node (a strategy known as depth-first search). These algorithms are particularly relevant when the internal model is deterministic, and the task is defined in terms of a discrete goal state (e.g., winning the game, or reaching a goal location)—the setting of classical planning (Figure 2). In such cases, these algorithms are sometimes called "blind search" because of the way they behave—the expansion of the tree happens without any knowledge of the goal location, and once the goal state is reached, a solution is identified, and the search process may be halted. In addition to expanding forward from the current state, search algorithms may also expand backward from the goal state or expand simultaneously in both directions.

One of the most effective approaches in AI is to use the agent's current state as the root node (because of its imminent relevance) and expand from there only up to a limited depth, a strategy known as "lookahead" (Figure 4B; Korf, 1990). When expansion is "truncated" in this way, action values are computed by summing the reward (or cost) accrued up to the truncated state with the heuristic value of the truncated state, which serves as a stand-in for the reward (or cost) to be accrued thereafter (see below). In board games, such as chess, rather than evaluating entire sequences of actions until the end of the game, an algorithm might evaluate a given action by assessing the predicted state of the game after a few moves, using a static evaluator that counts the different number of pieces on the board as a heuristic—a stand-in for the likelihood of victory (Campbell et al., 2002). In general, the deeper the lookahead is, the better the corresponding approximations are, but also the slower the planning process becomes (Geffner, 2013). An extreme version of lookahead is to expand a single-step decision "tree" (i.e., evaluating an action in terms of its immediate outcomes). This strategy is particularly fruitful in stochastic environments, where algorithms can improve the agent's policy by expanding a large number of such single-step trees, starting from all possible states (as in dynamic programming algorithms, such as value iteration or policy iteration) or from a random sample of states (as in the Dyna-Q algorithm).

If exhaustive planning is rarely possible in artificial agents, it ought to be even less likely in biological agents, where the rate of evaluation is thought to be much lower (de Groot, 1978). Indeed, a common observation in behavioral studies of planning is that humans often use lookahead to prioritize computation when making a decision (Keramati et al., 2016). The typical depth of expansion in humans, found across multiple behavioral studies, is approximately 3–6 steps (Snider et al., 2015; Arad and Rubinstein, 2012; Krusche et al., 2018; Huys et al., 2015; van Opheusden et al., 2017). Yet, even within a single task and subject, expansion depth is not a completely fixed algorithmic property, but rather variable according to immediate needs. Subjects trade off expansion depth with the frequency of recomputation, arbitrating between a shallow lookahead after each action or a deeper lookahead (pre-committing to a short action sequence) every few actions (Snider et al., 2015), pointing to the existence of some type of computational budget that can be re-allocated based on task demands. In line with this hypothesis, increased time pressure tends to result in a shallower lookahead (Keramati et al., 2016; van Opheusden et al., 2017), pointing to a speed-accuracy trade-off where a deeper expansion leads to more accurate evaluations.

To better prioritize a limited computational budget, informed (or focused) planning algorithms leverage estimates of the favorability of branches for choosing where to expand the decision tree. These estimates are typically obtained by using a "heuristic," a quick-and-dirty estimate of the distance or cost separating a state from a goal state or of the cumulative reward expected to follow from a state (and, in fact, the same value used as a stand-in when truncating the decision tree in lookahead algorithms; Pearl, 1984; Geffner, 2013). By using a heuristic, informed planning algorithms are able to focus computation on promising states and actions (or conversely, prune unpromising parts of the decision tree, Figure 4B). This prioritization saves vast amounts of computation, for example, when planning a route between New

York and San Diego, heuristics can prevent the evaluation of routes through Montreal that might have been considered by uninformed methods. Indeed, most uninformed algorithms have an informed counterpart which can achieve more efficient results by using a heuristic to focus computation (Figure 2). For instance, the most straightforward heuristic-based algorithm for deterministic environments, best-first search, improves upon breadth-first search by selecting nodes to expand greedily based on the heuristic value (as opposed to based solely on the node depth as in breadth-first search; Dechter and Pearl, 1985).

An even more effective approach, formalized in the highly successful A* (''A-star'') search algorithm (Hart et al., 1968), is to select nodes based not only on the heuristic value of the candidate node but also on the cost separating the root and the candidate node. This algorithm, in fact, forms the basis of many contemporary route planning services, such as that used in Google Maps. Heuristics are also incredibly useful for focusing computation in stochastic environments. Classic examples of informed algorithms that can be deployed with either deterministic or stochastic models are asynchronous dynamic programming, a focused version of conventional dynamic programming (Korf, 1990; Barto et al., 1995), prioritized sweeping, which prioritizes the selection of Dyna updates (Moore and Atkeson, 1993), and Monte Carlo tree search (MCTS; Coulom, 2007). MCTS is an algorithm commonly used in board games that evaluates actions iteratively, by performing multiple simulations of how the future might unfold (called rollouts) and using the results of early rollouts to focus later ones. Indeed, many of the recent AI breakthroughs in board games can be traced back to a clever selection of heuristics to guide the forward simulations in MCTS (Silver et al., 2016, 2017b, 2018).

Given the high performance and accuracy of informed algorithms in AI, one might expect that biological organisms might use similar strategies when planning. A good example of this is the widespread use of a Euclidean distance-based heuristic in animal (and human) navigation (Gallistel, 1990), putatively underpinned by the active maintenance and updating of Euclidean distance to the goal by the hippocampal formation (Howard et al., 2014; Epstein et al., 2017). More generally, behavioral data suggest that humans will use a range of focused strategies whenever possible. For example, humans typically avoid selecting a sequence of actions potentially leading to a transient large loss, even when considering the full set of future outcomes that could lead to a larger gain (Huys et al., 2012, 2015). This suggests that subjects ''prune'' entire branches of the tree during planning. Interestingly, when the full tree is presented visually along with the associated rewards, subjects do not seem to prune but take into account the entire set of prospective possibilities (Snider et al., 2015), suggesting that subjects only prune when internal evaluation is needed—another evidence for the flexible adjustment of algorithmic features. Notice that pruning and truncation are not mutually exclusive strategies, with the former reducing the width and the latter reducing the depth of the decision tree that needs to be considered (Figure 4B).

Another way to alleviate the computational burden of planning by well-chosen inputs is to combine a sequence of actions into units (sometimes called ''options'') that can then be evaluated as an individual model input (Figure 4B). In AI, options have played a key role in speeding up reinforcement learning and planning (Sutton et al., 1999b), becoming one of the key research areas for developing autonomous agents (Barto and Mahadevan, 2003). The challenge, of course, is to define such options in a way that proves useful. In humans, options also appear to play a key role in planning (Solway et al., 2014). In particular, actions that have consistently followed from one another are ''chunked'' as a unit that can be later evaluated independently, a mechanism that has been hypothesized to give rise to habits (Dezfouli and Balleine, 2013). For example, a detailed analysis of reaction time data using a computational model of evidence integration revealed that, indeed, sequences of two actions, rather than individual actions, were treated as independent units to be evaluated (Solway and Botvinick, 2015). Once a series of actions have been combined into a chunk, their computed values can be stored and later reused, a strategy known as ''memoization'' (Huys et al., 2015).

In humans, these various algorithmic motifs (e.g., lookahead, pruning, chunking) all seem to play a role in reducing the computational costs of planning by prioritizing the inputs to the model in humans (Huys et al., 2015). However, the results above also highlight an intrinsic difficulty in discovering the specific planning algorithm(s) used by biological agents: sometimes multiple algorithms with very different inner workings predict similar, if not identical, behavioral responses. Additionally, sometimes one strategy may masquerade as another (e.g., best-first search may indirectly implement pruning, since expanding greedily generally amounts to not expanding paths beyond a large loss). To overcome this limitation, rather than focusing only on the output behavior (the end result of the algorithm), one can attempt to observe the algorithm *at work* by directly recording neural activity associated with its latent variables (i.e., the expanded branches of the decision tree; Box 1).

As an example, direct neural recordings have allowed us to resolve an important distinction in this dimension: whether different branches of the decision tree are evaluated serially or in parallel (Cisek, 2012; Hunt and Hayden, 2017). This distinction is less relevant for AI because, in general, the specification of most AI algorithms requires serializing the evaluations as sequences of steps in computer code (though parallelization might be possible with multithreading). In biological organisms, however, the distinction is important to consider as it constrains the types of circuits capable of implementing planning. Interestingly, many computational models typically assume that multiple options can be evaluated in parallel (Wong and Wang, 2006; Balleine et al., 2007; Collins and Frank, 2014), but neural data might suggest otherwise. In a task involving multi-attribute choice, recordings from OFC neurons in non-human primates correlate primarily with the value of the currently fixated option rather than all options, representing different values as the focus of attention is shifted (Hunt et al., 2018). This pattern of results also mimics the finding that OFC primarily represents the value of the attended option at any given time (Rich and Wallis, 2016). While these tasks do not require expanding a decision tree and, as such, cannot fully rule out the parallel evaluation of different branches, they nonetheless suggest that, at least in OFC, action evaluation might proceed in a serial manner.

Please cite this article in press as: Mattar and Lengyel, Planning in the brain, Neuron (2021), https://doi.org/10.1016/j.neuron.2021.12.018

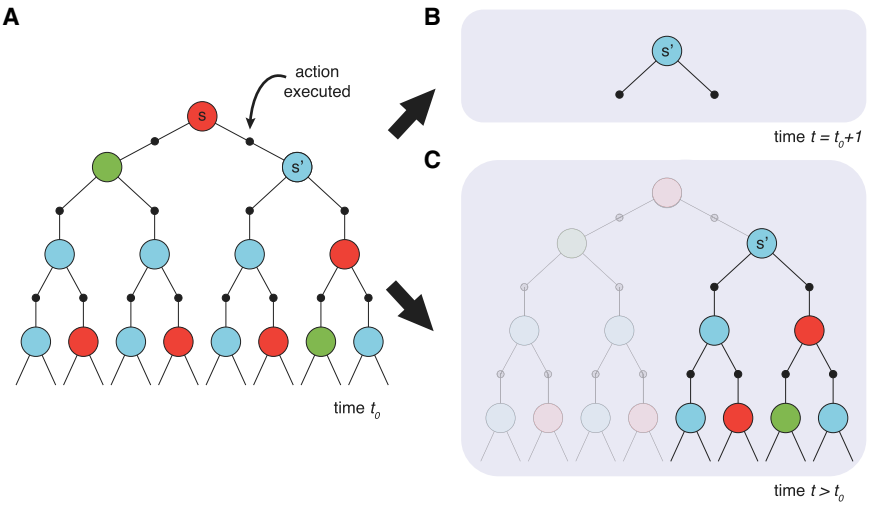

Figure 5. Use and re-use of the results of planning

(A) A decision tree expanded at time t_0 using any of the methods described in the main text. We consider here a setting where the agent expands from state s, which has s' as one of its successor state.

(B) In online algorithms (i.e., decision-time planning), planning and acting are interleaved. The results of a tree expansion are used only for choosing the immediate next action and are discarded once the consequent state transition has happened. For this reason, online algorithms typically focus computation on the current state (e.g., lookahead). In the example shown, state s was expanded at t_0 because it was the current state at the time, and by executing the appropriate action, a transition to state s' happened. Upon transitioning to a new state s', at $t = t_0 + 1$, planning needs to start afresh.

(C) In offline algorithms (i.e., planning in the background), planning and acting are performed in separate phases. The results of a tree expansion are, by necessity, stored for later use. Thus, upon transitioning to a new state s', agents can re-use previous computations with minimal cost when selecting actions. Note that, while we represent here the re-use of a plan after a single transition, planning computations can also be stored and reused at a much later time.

DESIGN CHOICE 3: INCORPORATION OF OUTPUTS

The third major way in which planning algorithms differ is in how planning ultimately informs imminent or future behavior. In other words, upon estimating action outcomes and their utility, how are the results translated into improved behavior? The most important distinction here is between online and offline algorithms, depending on whether the results of planning are used immediately or saved for later (Figure 5). In "online" (also known as "decision time") algorithms, planning and acting are interleaved, such that the results of planning (Figure 5A) are used only for choosing the immediate next action and are discarded once the consequent state transition has occurred (Figure 5B). However, time is always at a premium when planning online: we rarely have unlimited time to choose our actions (for example, in real-time game play). Therefore, online planning almost always needs to be approximate. The single most common approximation used in online planning algorithms is to focus computation on the agent's current state and on the evaluation of immediately available actions for imminent behavior (i.e., to use lookahead, see design choice 2). This strategy is, in fact, used to construct online counterparts to classic offline algorithms, such as the learning real-time A* algorithm based on A* (Figure 2).

Most studies of planning in biological organisms tend to focus on online planning because of its experimental tractability: it is much more straightforward to link a choice to cognitive and neural processes that immediately preceded it than to processes that occurred at any earlier point in time. Indeed, all of the empirical results presented so far in this paper concern online planning. Historically, perhaps the first report of a behavioral signature associated with online planning was Tolman's description of vicarious trial and error (VTE), almost a century ago (Tolman, 1938). When rodents reach a decision point in a maze (e.g., a bifurcation), they often pause and orient back and forth between the alternative paths, as if they were deliberating between the choices (Redish, 2016). Interestingly, VTE tends to disappear after extensive training on a task (Redish, 2016), consistent with a reduced need for planning in extensively trained tasks (Daw et al., 2005).

More recently, VTE was found to happen concurrently with theta sequences in the hippocampus (Foster and Wilson, 2007; Johnson and Redish, 2007). The hippocampal formation can potentially offer a direct window into planning algorithms, given the relatively straightforward mapping between states and locations in navigation (O'Keefe and Nadel, 1978) and the well-known encoding of location in the activity of hippocampal place and entorhinal grid cells (Moser et al., 2017). While place cells typically respond when the animal is in a specific location in the environment (the place field of the cell), during VTE (but also during freezing in fear and active movement), these cells fire in sequences encoding a short trajectory connecting the animal's previous location, its current location, and a few future locations (Foster and Wilson, 2007; Johnson and Redish, 2007). Notably, at decision points with two alternative paths, theta sequences encode the alternative paths in constant alternation at approximately 8 Hz (i.e., one path per 125 ms, akin to a short rollout) (Kay et al., 2020). This raises the intriguing possibility that this cycling might underlie the evaluation of different candidate paths, as in a lookahead-like online planning algorithm (Papale et al., 2016; Pezzulo et al., 2019).

Another way to manage the problem of time pressure that haunts planning is to abandon online algorithms altogether and instead, use "offline" algorithms in which planning and acting are performed in separate phases. In the planning phase, a plan (an action sequence in the deterministic case or a policy in the stochastic case) is pre-computed and then stored. This stored plan can then be consulted with minimal computational (and, thus, time) cost during the action phase (Figure 5C). Classical planning methods (using deterministic internal models), such as breadth-first or depth-first search, are typically used in the offline setting where a whole sequence of actions (a solution) must be found before the agent starts acting. In stochastic

Neuron
Perspective

CellPress
OPEN ACCESS

Box 1. Using behavioral and neural data to adjudicate between planning algorithms

To discover how the brain implements planning, we argue that neuroscientists should specify their computational hypotheses in the form of planning algorithms. As we argued previously, each planning algorithm can be thought of as a recipe for converting goals and knowledge of the world into actions (Figure 6A). Thus, to adjudicate between alternative algorithms, one would ideally examine both the resulting actions predicted by each algorithm and the actual recipes that gave rise to these actions and compare those with behavioral and neural data recorded from animals, respectively. Below, we describe how this approach would work in practice through a simple illustrative example.

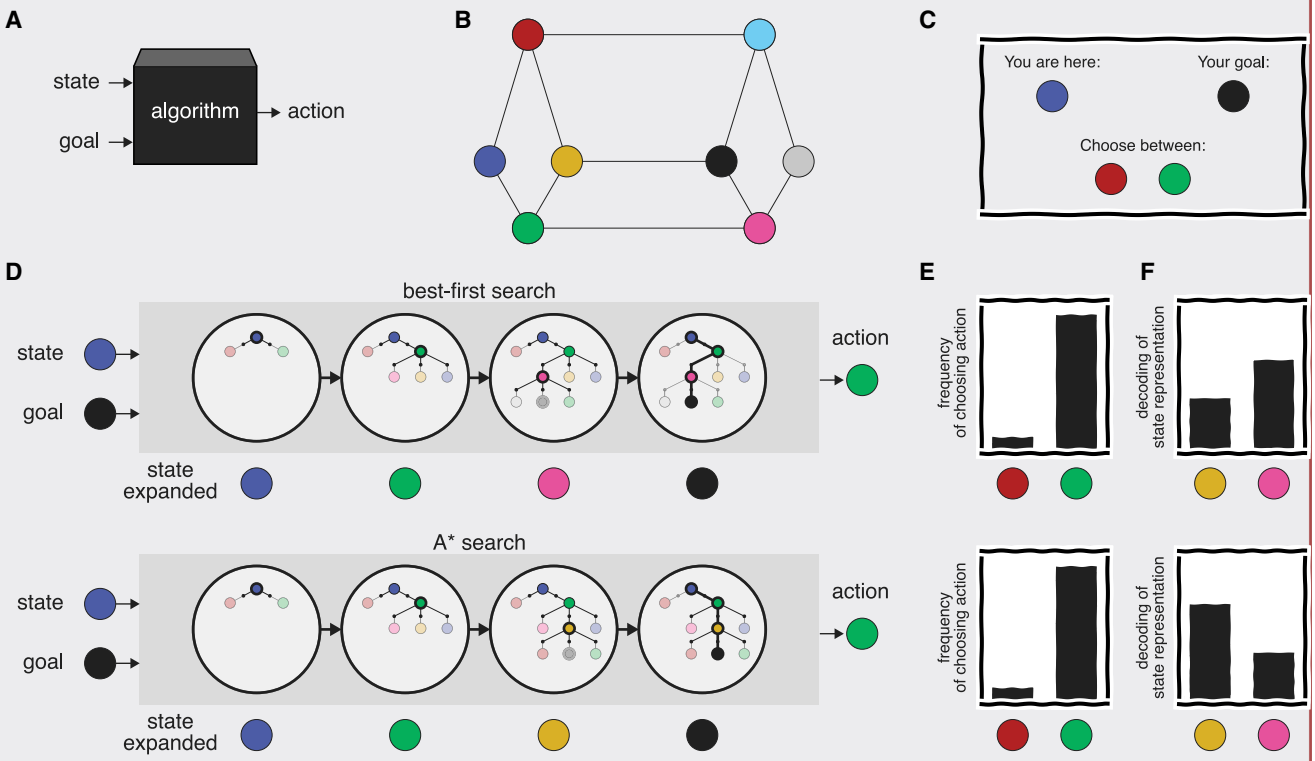

Figure 6. Example approach for using behavioral and neural data to adjudicate between planning algorithms

(A) A computational hypotheses about planning can be formalized as an algorithm (represented as a black box): a procedure that, for a given input state and goal, outputs an action or action sequence.

(B) A state-space representation of a simplified navigation task. Each colored node represents a different state, and each edge represents a possible transition between states (via an action). The length of an edge indicates the distance between the nodes it connects, which we formalize as a cost associated with that transition. Since this representation is assumed to be known by the agent, it is also the agent's internal model of the task.

(C) Example trial of the task. The agent is presented with a start and a goal state. The agent needs to identify the action sequence requiring the minimum total traveled distance from the start to the goal, and to select the first action of this sequence.

(D) Expansion of a decision tree according to two alternative algorithms: best-first search (top) and A* search (bottom). Each algorithm iteratively expands a decision tree starting from the start state until a path to the goal state is found. Each white circle represents one iteration of expansion, with the state expanded indicated beneath it. Note that the third state expanded by best-first search (pink, top) is different from the third state expanded by A* (yellow, bottom).

(E) Example behavioral data predicted by each algorithm across multiple trials equivalent to (C). Each bar indicates the relative frequency of selecting each possible action. Aside from a small amount of randomness, both algorithms predict the selection of the action leading to the green state.

(F) Example neural analysis. Neural data collected during presentations of distinctive sensory cues associated with each state can be used to train a state classifier (e.g., using multi-voxel pattern analysis in functional imaging [Norman et al., 2006] or linear decoding of population recordings in electrophysiological experiments [Dayan and Abbott, 2001]). This classifier can then be used to reveal which states are likely to have been expanded during planning. Each bar indicates the output of a binary classifier applied to distinguish between the yellow and pink states. For a best-first search agent, the classifier is expected to provide greater evidence for the pink state (expanded) versus yellow (not expanded) during planning. Conversely, for an A* search agent, the classifier is expected to provide greater evidence for the yellow (expanded) versus pink (not expanded) during planning.

We consider here the adjudication between two simple search (classical planning) algorithms: best-first search and A* search. Both these algorithms are informed, offline, and use deterministic models, so their distinction is rather subtle (Figure 2). Indeed, these algorithms only differ with respect to the order of computation (design choice 2). Here, we examine these algorithms in a simple multi-step task with eight states and deterministic transitions, analogous to a spatial navigation task (Figure 6B). In this example task, when presented with a start and a goal state, the subject must identify the action sequence that requires the minimum traveled distance from start to goal (Figure 6C). Note that, while these algorithms compute entire action sequences every time they are executed, the observable behavior of the subject only depends on the first action of the sequence computed at a given decision

point (as a new sequence may be planned as soon as the next decision point is reached). For simplicity, we assume that the subjects have complete knowledge of the states and transitions defining the task (i.e., the internal model).

Let us consider an example trial in this task. Given a state and a goal, each algorithm follows a different recipe for expanding a decision tree until an action is selected (Figure 6D). Best-first search always expands the node with highest heuristic value, which, in this case, amounts to the node with the lowest estimated (e.g., Euclidean) distance to the goal. A*, instead, expands the node with the lowest estimated distance from the start to the goal, i.e., it considers both the distance from the start and the estimated distances to the goal.

To distinguish which of the two algorithms provides a better account of subjects' behavior, the simplest approach is to compare the actions selected by each. If the task is sufficiently rich, the actions selected by each algorithm might be different. In such cases, subjects' behavior can be compared with the actions selected by each algorithm. By repeating this procedure across multiple trials, one algorithm may be found to provide a better fit to subject behavior than the other.

However, if the task is not sufficiently rich or the difference between the two algorithms is too subtle, the actions selected by each could be identical, as in our example trial (Figure 6E). Reaction times may also offer useful clues, as even if two algorithms arrive at the same decision, they may take a different amount of computations, e.g., different number of iterations, to do so. For serial algorithms, this will be reflected in different reaction times. However, sometimes even this cannot distinguish between two algorithms (Figure 6E, both algorithms take 4 iterations to reach the goal). In such cases, neural data can be used to adjudicate between the candidate algorithms. For example, if the competing algorithms expand different nodes, neural data can be used to classify/decode these states during planning, providing greater evidence for one algorithm over the other (Figure 6F).

settings, where optimization is performed over policies rather than action sequences, offline planning can be used to improve the agent's policy, which can then be used for selecting multiple actions in the future. This process, sometimes known as "background" planning, is typified by algorithms, such as dynamic programming and Dyna, which compute policies in remote states and store them for later use (Figure 2). However, the time savings at action time of such offline algorithms come with the disadvantage that the previously computed plans may become less useful over time, (e.g., when they involve states that are currently irrelevant or unreachable). Worse even, precomputed plans may also be invalid (e.g., if the environment, or our knowledge of it, has changed since the pre-computation was performed), thus losing some of the behavioral flexibility that is the main appeal of planning in the first place.

Behavioral evidence for offline planning in biological organisms is often difficult to obtain due to the lack of direct correspondence between the planning process and the improved behavior that results from it. However, a carefully designed behavioral experiment with humans suggests that processes occurring during a rest phase can impact a subject's choices in a later test phase (Gershman et al., 2014). Using a 2-step task, subjects learned different parts of an internal model (first-stage outcomes and second-stage rewards) in different phases of the experiment. Choice behavior was consistent with a planning agent that combined the different pieces of the internal model in their decision, indicating a planning computation. Interestingly, increasing the subject's cognitive load during the learning of the second-stage rewards but long before the decision, reduced the degree to which this combination occurred. In contrast, increasing cognitive load immediately prior to the final choice did not affect performance. This suggests that the cognitive load manipulation interfered with the planning processes even before subjects were asked to make a decision, suggesting the existence of an offline planning mechanism.

Associating neural data with offline planning is also challenging because of the inherent difficulty of relating neural activity to a change in behavior that is only measured at a later point in time. One way of establishing this relationship is to identify neural activity related to a state other than the agent's current state (a nonlocal state) and which is unlikely to be relevant for decisions the animal is currently taking (if any). Here, again, hippocampal place cells may shed light onto this process. In "hippocampal replay," similar to the theta sequences described above, place cell activity represents coherent sequences of spatial locations akin to those experienced by the animal during navigation. For example, hippocampal replay often forms trajectories that start at the animal's location and can extend forward toward a goal location, with the represented trajectory being predictive of upcoming behavior (Pfeiffer and Foster, 2013; Singer et al., 2013). Thus, like theta sequences, hippocampal replay may be involved in online planning. However, hippocampal replay differs from theta sequences in several important ways that make it also suitable for representing offline planning. For instance, unlike theta sequences, which happen during active locomotion when online planning would be required (Foster and Wilson, 2007; Pezzulo et al., 2019), hippocampal replay often happens at times when there are no apparent decisions that need to be planned, such as during moments of rest (Diba and Buzsáki, 2007) and sleep (Wilson and McNaughton, 1994; Lee and Wilson, 2002). Also, unlike theta sequences, hippocampal replay can extend not only forward but also backward in space (Foster and Wilson, 2006), recapitulating the animal's previous steps. This may be a signature of backward planning from the current location, which can be useful for future decisions, i.e., offline planning, but hardly for current decisions, i.e., for online planning. Hippocampal replay can also span much longer distances (Davidson et al., 2009) than those spanned by theta sequences and may represent locations in different environments altogether (Karlsson and Frank, 2009). Therefore, while theta sequences appear to be involved strictly in online planning, the representational content of hippocampal replay is highly suggestive of a potential role in not only online but possibly also offline planning (Mattar and Daw, 2018; Pezzulo et al., 2019).

To go beyond a suggestive role, however, one must establish a direct link between nonlocal activity and actual changes in the corresponding nonlocal policy, as reflected in behavior. Some evidence points in that direction. For example, when replay is interrupted electrophysiologically, animals exhibit dysfunctional value learning (Girardeau et al., 2009) and deficits in behaviors that cannot be executed with a reflexive strategy (Jadhav et al., 2012). By optogenetically disrupting the replay of specific memory traces, animals exhibit deficits in learning the association of rewards to those locations (Gridchyn et al., 2020). Moreover, awake replay activity predicts correct choices on a trial-by-trial basis (Singer et al., 2013), and sleep replay of an unexplored arm predicts later preference for that arm (Ólafsdóttir et al., 2015). More recently, hippocampal replay was found to mediate nonlocal credit assignment in rodents (Barron et al., 2020), and backward replay (measured with magnetoencephalography) was similarly found to mediate nonlocal credit assignment in humans (Liu et al., 2021). The role of replay in offline planning suggested by these various studies can be captured by a normative computational model associating place cell activity with nonlocal action evaluation (Mattar and Daw, 2018).

However, other empirical evidence paints a more controversial role of replay on planning. For example, replay can sometimes over-represent trajectories unrelated to the upcoming behavior, more so than currently relevant locations (Gupta et al., 2010). Moreover, these trajectories more often include previously than currently rewarding states (Carey et al., 2019; Gillespie et al., 2021). Thus, to establish a conclusive relationship between replay and planning, existing computational models will need to be revised to explain these apparently contradictory results. More generally, offline planning is but one possible computational manifestation of memory consolidation, inasmuch as it involves the retrieval of specific memories and their integration into a longer-term, more semantic representation (here, the value function or policy). Indeed, hippocampal replay has been traditionally interpreted as a signature of memory consolidation or maintenance (McClelland et al. 1995; Káli and Dayan 2004). It will be interesting to develop and test specific predictions that distinguish between all these different and perhaps complementary computational accounts of replay.

To be able to flexibly trade off time against accuracy, many of the algorithms we considered so far can produce useful (albeit not necessarily optimal) behavior even when their execution is interrupted prematurely (i.e., they belong to the class of *anytime* algorithms; Horvitz, 1987; Russell and Norvig, 2010). In other words, in these algorithms, planning is not an all-or-none affair, where you have either a complete plan or nothing. Instead, they incrementally improve their plan (and thus their performance) with more processing time. There are fundamentally two distinct mechanisms for achieving this. First, there exist "deterministic" algorithms that iteratively refine their estimates of action values or optimal actions (policy), e.g., by increasing the depth of lookahead or by implementing some form of gradient-based optimization. These include policy gradients (Sutton et al., 1999a), several incarnations of generalized policy iteration (Sutton and Barto, 2018), and some neural network algorithms (Friedrich and Lengyel, 2016). Second, there exist "stochastic" algorithms that, on each run, only return a single noisy estimate of values or actions. This may happen due to using a sampling-based stochastic internal model (as in design choice 1) or stochastically choosing actions for evaluation (as in the rollout-based algorithms described within design choice 2). In either case, the outputs of several runs therefore need to be averaged to obtain a more reliable estimate (an idea that underlies the now ubiquitous planning algorithm MCTS and its many variants; Coulom, 2007; Silver et al., 2016). It is this averaging that can then be performed iteratively and that essentially corresponds to a form of *internal* evidence accumulation (where the evidence to be accumulated is the sequence of noisy estimates provided by subsequent runs of the stochastic planning algorithm). For example, such an evidence accumulation process, performed in parallel for different paths of a decision tree, has been proposed to account for the dynamics of deliberation in online planning (Solway and Botvinick, 2015). More generally, it remains an interesting open question as to which class of iterative anytime algorithms is responsible for the ubiquitously observed speed-accuracy trade-offs in biological planning.

PLANNING OUR WAY FORWARD

We have described a new taxonomy for the three core components of a planning algorithm and, oriented by this taxonomy, reviewed what is currently known and unknown about how planning is implemented in the brain. In dimension one, we found that the brain is capable of using stochastic internal models but will resort to deterministic ones if the task permits. In dimension two, we found that the brain uses heuristics to focus computation on the most relevant, depth-limited action sequences and likely evaluates candidate actions serially and not simultaneously. Finally, in dimension three, we reviewed evidence for both online and offline planning in different settings and for an iterative (rather than all-or-none) process where performance incrementally improves as more time is allowed for planning.

In light of recent breakthroughs in AI research, particularly in the field of planning (Silver et al., 2017a, 2018; Hafner et al., 2019a; Schrittwieser et al., 2020), one might be compelled to think that such breakthroughs were made possible by entirely distinct approaches to planning that would not fit into this taxonomy. However, such breakthroughs are better seen as concrete examples of algorithms that work well in practice and less as a reinvention of how planning is possible. In fact, state-of-the-art AI algorithms, like any of the classical algorithms, can still be described using the same three fundamental dimensions. What is new is that these algorithms exploit neural networks and advances in deep learning to better select which nodes to expand (dimension 2) (Silver et al., 2016, 2017b), make efficient use of cached computation (dimension 3) (Hamrick et al., 2019), and most recently, to learn an efficient internal model from experience (dimension 1) (Schrittwieser et al., 2020). It is yet unclear whether the specific neural architectures used in these algorithms have direct relevance for neuroscience (though some examples point in that direction, e.g., Banino et al. 2018; Wang et al. 2018). Nevertheless, research on the neurobiology of planning could certainly take inspiration at a more abstract, algorithmic level from these advances. Recent work on experience replay has started doing just that (Cazé et al., 2018; Mattar and

Daw, 2018; Igata et al., 2021), but a more systematic approach is needed to take full advantage of new developments in AI planning.

Ultimately, the usefulness of our taxonomy is measured by the extent to which it clarifies the various ways in which planning can be accomplished by any agent, be it artificial or biological. Therefore, our hope is that this taxonomy will, in the future, catalyze the generation of clear and testable experimental hypotheses, e.g., by providing a systematic framework for adopting advances in AI to computational models of planning in the brain. A more immediate benefit, however, is to reveal the main limitations in our current understanding of planning in the brain. These limitations stem from three main sources: (1) behavioral tasks and data have, for the most part, been too simple to differentiate algorithmic details; (2) the algorithmic motifs inferred so far have been too specific to the tasks used; (3) neural data have so far contributed surprisingly few constraints as to how these algorithms might be actually implemented in the brain. Below, we make concrete suggestions on how to overcome each of these three limitations, highlighting some recent work pointing in these directions.

Embracing task complexity

First, we suggest that richer behavioral paradigms are needed to further narrow the space of candidate planning algorithms. While existing experimental paradigms were successful in identifying a set of motifs (truncation, pruning, etc.), those motifs can still arise as components of a large number of distinct algorithms. To achieve a higher level of precision, experimenters should carefully design tasks to be maximally sensitive to specific algorithmic differences. As discussed previously, these tasks should, at a minimum, require multiple steps of decisions. Specifically, tasks should have enough complexity (e.g., stochastic and dynamic environments or sufficiently large decision trees) to differentiate between different algorithms. A recent study represents a step in this direction, studying human behavior in a two-player deterministic game where players compete to create four-in-a-row on a 4-by-9 board (van Opheusden et al., 2017). This game, which resembles a complex version of tic-tac-toe, contains upwards of 10^{17} possible states, making it substantially more complex than most tasks previously used to study biological planning. By comparing choices made by humans against those made by many planning algorithms, it was possible to infer that motifs, such as truncation and pruning are important to account for human behavior. However, the complexity of the game further allowed the authors to test the ability of multiple specific algorithms to explain human behavior. The authors proposed that a specific candidate—"lookahead best-first search" based on a feature-based heuristic—matched human behavior better than any other algorithm considered. Using sufficiently complex behavioral paradigms, it might also be possible to contrast human behavior against state-of-the-art algorithms in AI, such as those involving variants of MCTS (Silver et al., 2016, 2017b, 2018), and to discover the ways in which human behavior surpasses or falls short of AI performance.

While complex game-based paradigms have the potential to distinguish between specific algorithms, they are often limited to be used with humans and thus mostly non-invasive recordings. In addition, they leave open the question of how much the results obtained with them (and in particular, the specific planning algorithms they identify) generalize to other, more ecologically relevant tasks. Thus, future investigations should attempt to combine task complexity with ecological relevance in both humans and other animals (Krakauer et al., 2017; Hunt et al., 2021). In this regard, navigation could be an ideal testbed in which rich and exquisitely well-controlled behavioral paradigms exist (Tolman, 1948; O'Keefe and Nadel, 1978) in combination with established neural correlates (Moser et al., 2017; Bellmund et al., 2018), and modern high-throughput recording methods (Jun et al., 2017; Hong and Lieber, 2019). Although deterministic internal models are sufficient for standard navigation tasks (with the walls of a maze being fixed), more recent paradigms allow the introduction of stochasticity and thus pave the way to identifying a more general class of algorithms (Wood et al., 2018; Duvelle et al., 2021). The link between navigational signals in neural responses and planning has only recently been studied systematically and thus offers new insights not only into planning but also into the navigational systems of the brain (Pfeiffer and Foster, 2013).

The advantage of using navigation-based paradigms for investigating planning need not be limited to studying navigation in physical spaces (Behrens et al., 2018). A number of studies have now used behavioral paradigms directly modeled after (simplified) navigation tasks but in which subjects navigated more abstract cognitive spaces (Eichenbaum and Cohen, 2014; Constantinescu et al., 2016). These studies also revealed that neural responses in the hippocampus and entorhinal cortex, commonly studied in spatial navigation tasks, may also encode "locations" in these more abstract, cognitive spaces (Constantinescu et al., 2016; Aronov et al., 2017; Whittington et al., 2020), thereby implementing a *bona fide* state representation in the formal sense of reinforcement learning (Figures 1A and 1B). Thus, these results illustrate how the neural bases of more general forms of planning can be studied. Indeed, recent results have indicated offline planning in abstract state spaces similar to that seen in physical spaces (Liu et al., 2021). However, such abstract navigation paradigms have so far had relatively minimal planning requirements, often using passive navigation (in which subjects are exposed to a trajectory in the abstract space rather than allowed to choose one) or simple state transition structures. Therefore, in line with our suggestion above, to be able to distinguish between different planning algorithms, future studies will need to challenge subjects with more complex planning tasks in these abstract state spaces.

Motor control presents another highly ecological domain that has long been studied and shown to engage sophisticated stochastic internal models (McNamee and Wolpert, 2019). Recent work has also started revealing neural correlates of motor planning (Churchland et al., 2006; Averbeck et al., 2002; Mushiake et al., 2006). However, so far little attention has been given to identifying specific algorithms. Planning in motor control poses particularly strong constraints on the space of feasible algorithms due to its inherently large state and action spaces (continuous limb positions, joint angles, and muscle activations), multiple sources of noises and delays, and strong time pressures—making it a direction particularly ripe for investigating

the neural bases of planning in a domain that remains challenging for even today's AI algorithms (Arulkumaran et al., 2017). Also important is the specification of how planning in the space of motor commands relates to planning in the space of abstract states, particularly since the two types of plans appear to be represented in distinct brain regions (Mushiake et al., 2006).

One particular challenge is to design tasks in which subjects cannot rely on reflexive strategies. This creates a tension when working with animals: sufficiently high performance typically requires extensive training, which in turn can easily lead to habitual behavior based on reflexive strategies (Akam et al., 2015). Trial-by-trial changes in outcome-reward associations (e.g., by devaluation; Balleine and Dickinson, 1998) or goals (e.g., in a navigation-like setting; Mushiake et al., 2006) can promote planning-based strategies but still cannot exclude more sophisticated reflexive strategies that use temporal abstraction (such as the SR; Dayan, 1993; Russek et al., 2017). Ultimately, tasks in which transition probabilities between states can change (such as a short-cut opening; Wood et al., 2018; Duvelle et al., 2021) will be necessary. Moreover, such changes will need to occur repeatedly to allow the collection of sufficient amounts of behavioral data from each subject. In turn, for the performance to remain high in the face of a continually changing transition structure, subjects need to be (pre-)trained, such that they acquire more abstract knowledge about transitions that can be generalized to novel states. Paradigms using navigational schemata may be a good starting point for this (Tse et al., 2007).

Ultimately, with the advent of more sophisticated behavioral paradigms, it also becomes increasingly imperative to maximize the richness of the recorded behavioral data, beyond just registering final choices and reaction times, as in classical studies. Modern high-throughput multimodal recordings and analysis methods (Mathis et al., 2018) could involve the simultaneous registering of VTE-like behavior, eye movements, pupil dilation, stiffness, and more—all suggested to be linked to planning and its component processes (Redish, 2016; Callaway et al., 2021; Zénon, 2019; Franklin and Wolpert, 2011) but rarely studied in a unified analysis framework.

Going beyond the taxonomy

Second, as we increase the richness of behavioral paradigms, we may be faced with additional limitations: what if the algorithms used by the brain are adapted to the specific task at hand, thus varying from task to task? Can we ever hope to consider all viable candidate algorithms? And what if the brain uses fundamentally different algorithms than those used in AI? These questions lead to another important direction for further progress, which is to consider more expressive algorithms.

An approach that has been particularly fruitful is to define a meta-algorithm that receives a task as an input and produces as the output a planning algorithm that is appropriate for that task—under some quality metric and resource constraints. This approach can lead to different algorithms to different tasks while maintaining the elegance and compactness of a single, generative normative approach. The resulting algorithm can then be interpreted in light of the taxonomy and, indeed, compared across tasks. For instance, starting from the assumption that people make rational use of their limited cognitive resources, one can derive rational models and algorithms, an approach known as resource-rational (Lieder and Griffiths, 2019). Applying this framework to planning, recent studies have found that human choices are better described by a resource-rational planning strategy than by more conventional algorithms (such as those from Figure 2), and moreover, the details of the planning strategies used are adapted to the structure of the task at hand (Callaway et al., 2018; Correa et al., 2020). A similar normative approach has also been helpful in explaining neural data associated with planning, where the content and directionality of hippocampal replay were shown to arise naturally from optimal ordering of Dyna updates. Such optimal ordering of planning operations reproduces various physiological observations, such as the existence of both forward and reverse replay sequences, the greater representation of rewards and the agent's location, and different effects of experience (Mattar and Daw, 2018).

Another systematic approach for generating possible planning algorithms involves re-framing planning as a problem of probabilistic inference—computing a posterior distribution over actions, conditioned on reaching the goal state or maximizing reward (Attias, 2003; Toussaint and Storkey, 2006). This framework, called planning-as-inference, leads to a parsimonious explanation for numerous behavioral and neural findings related to goal-directed behavior and connects planning to other cognitive systems in sensory and motor domains, already understood within a probabilistic framework (Botvinick and Toussaint, 2012; Solway and Botvinick, 2012). Interestingly, classic planning algorithms have been "rediscovered" as the solution to an inference problem (e.g., policy iteration can be understood as computing optimal policies via expectation maximization; Toussaint and Storkey, 2006). This raises the intriguing possibility that novel planning algorithms might also be discovered using the planning-as-inference framework, particularly for problems where exhaustive planning (corresponding to exact inference) is too costly. For example, methods for approximate inference based on sampling might give rise to rational approximations to the exact planning solution, with the additional advantage of having potentially plausible neural instantiations (Fiser et al., 2010; Echeveste et al., 2020). Therefore, we suggest that in the future, either of these approaches (resource-rational planning or planning-as-inference) could be used to populate the space of planning algorithms in a systematic way, going beyond those that happen to be included in today's AI textbooks (Figure 2).

In addition to principled approaches for generating planning algorithms, the very definitions of actions and states can also be reconsidered. We have already discussed the learning of action chunks (the core idea underlying hierarchical reinforcement learning; Barto and Mahadevan, 2003), but more efficient representations of the state space can also be learned (Akam et al., 2015; McNamee et al., 2016). While learned latent states are usually specific to a certain context (and thus, by definition, less generalizable), they can also be optimized for a whole set of interrelated tasks (Wang et al., 2018). Good abstractions of the state space can also be identified by combining states that are equivalent in terms of both the transition and reward functions, a process that leverages the SR for state-space compression

(Lehnert et al., 2020). Therefore, reconsidering the building blocks of a planning agent (Figure 1B) is a fundamental yet often overlooked step, as the advantages and computations of each algorithm—and in turn their neural instantiations—depend directly on how their elements are defined.

In this paper, we have intentionally omitted a discussion of how the internal model might be learned. This is because once the internal model used by the agent is known, the process that gave rise to that model (i.e., learning) is no longer relevant to the planning algorithms used. This rationale, however, obscures the inherent difficulty in inferring the agent's internal model. In many experimental approaches, the agent will be required to learn the model (or parts thereof) during the experiment itself. This results in uncertainty, both from the agent's point of view (i.e., epistemic uncertainty about action outcomes) and from the scientist's point of view (i.e., epistemic uncertainty about the agent's learned internal model). Incorrect assumptions about the agent's internal model can, in turn, masquerade as algorithmic processes, as different algorithms using different internal models may produce identical behaviors (see our discussion on inferring the internal model type). One way to circumvent this issue is to probe subjects for the internal model used with an orthogonal task, e.g., an outcome prediction task. Alternatively, one can jointly infer the internal model and the parameters of the behavioral model (Houlsby et al., 2013; Wu et al., 2020). More broadly, considering the process by which the model is learned can have important consequences for the identification of the planning algorithms (Gläscher et al., 2010; Behrens et al., 2018).

More insights with neural data

Third, as the very definition of a planning algorithm is revised, so are the ways of interpreting the relevant data. Here, again, we propose that behavioral data alone will not suffice, since it reveals little more than the final result of the planning process. Thus, our final proposal for further progress is that the richer behavioral paradigms and analysis methods proposed above should be accompanied by concurrent neural recordings in brain areas of relevance. Given the relatively straightforward mapping between states and place cells (Eichenbaum, 2017), the hippocampus can be used as a window into planning, ruling algorithms in and out based on how well they predict the activation of remote states. However, high-quality neural data from other relevant areas such as the prefrontal cortex and basal ganglia should also be used. Indeed, despite the uncontested causal involvement of these regions in planning behavior resulting from lesion work (Duncan et al., 1996; Unterrainer and Owen, 2006), little is known about how planning algorithms are implemented in their dynamics and, particularly, how they interact with one another and with the hippocampus. It is, of course, essential to acknowledge that different systems in the brain may be optimized for different computations (e.g., a common view is that the hippocampus might be specialized for spatial planning). Nonetheless, a fuller picture of how planning algorithms are implemented by the brain will only be achieved by considering simultaneously the involvement of the complete planning network (Wunderlich et al., 2012; Dolan and Dayan, 2013; Patai and Spiers, 2021), and how the dynamics of their activity ultimately implement the computations underlying planning.

ACKNOWLEDGMENTS

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 726090 to M.L.), the Wellcome Trust (Investigator Award 212262/Z/18/Z to M.L.), and a Newton International Fellowship of the Royal Society (grant 181426 to M.G.M.).

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Adams, C., and Dickinson, A. (1981). Actions and habits: variations in associative representations during instrumental learning. In Information Processing in Animals: Memory Mechanisms (Lawrence Erlbaum Associates), pp. 143–166.

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

Akam, T., Rodrigues-Vaz, I., Marcelo, I., Zhang, X., Pereira, M., Oliveira, R.F., Dayan, P., and Costa, R.M. (2021). The anterior cingulate cortex predicts future States to mediate model-based action selection. Neuron *109*, 149–163.

Albin, R.L., Young, A.B., and Penney, J.B. (1989). The functional anatomy of basal ganglia disorders. Trends Neurosci. *12*, 366–375.

Anderson, J.R. (1982). Acquisition of cognitive skill. Psychol. Rev. *89*, 369–406.

Anderson, J.R., Albert, M.V., and Fincham, J.M. (2005). Tracing problem solving in real time: fMRI analysis of the subject-paced Tower of Hanoi. J. Cogn. Neurosci. *17*, 1261–1274.

Arad, A., and Rubinstein, A. (2012). The 11–20 money request game: A level-k reasoning study. Am. Econ. Rev. *102*, 3561–3573.

Aronov, D., Nevers, R., and Tank, D.W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. Nature *543*, 719–722.

Arulkumaran, K., Deisenroth, M.P., Brundage, M., and Bharath, A.A. (2017). Deep reinforcement learning: A brief survey. IEEE Signal Process. Mag. *34*, 26–38.

Attias, H. (2003). Planning by probabilistic inference. In Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics R4 (PMLR), pp. 9–16.

Averbeck, B.B., Chafee, M.V., Crowe, D.A., and Georgopoulos, A.P. (2002). Parallel processing of serial movements in prefrontal cortex. Proc. Natl. Acad. Sci. USA *99*, 13172–13177.

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

Balleine, B.W., Delgado, M.R., and Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. J. Neurosci. *27*, 8161–8165.

Ballesta, S., Shi, W., Conen, K.E., and Padoa-Schioppa, C. (2020). Values encoded in orbitofrontal cortex are causally related to economic choices. Nature *588*, 450–453.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M.J., Degris, T., Modayil, J., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. Nature *557*, 429–433.

Barron, H.C., Reeve, H.M., Koolschijn, R.S., Perestenko, P.V., Shpektor, A., Nili, H., Rothaermel, R., Campo-Urriza, N., O'Reilly, J.X., Bannerman, D.M., et al. (2020). Neuronal computation underlying inferential reasoning in humans and mice. Cell *183*, 228–243.

Barto, A.G., and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. Discrete Event Dyn. Syst. *13*, 41–77.

Barto, A.G., Bradtke, S.J., and Singh, S.P. (1995). Learning to act using real-time dynamic programming. Artif. Intell. *72*, 81–138.

Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A.B., Stachenfeld, K.L., and Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. Neuron *100*, 490–509.

Bellman, R. (1957a). Dynamic Programming (Princeton University Press).

Bellman, R. (1957b). A Markovian decision process. J. Math. Mech. *6*, 679–684.

Bellmund, J.L.S., Gärdenfors, P., Moser, E.I., and Doeller, C.F. (2018). Navigating cognition: spatial codes for human thinking. Science *362*, eaat6766.

Berkes, P., Orbán, G., Lengyel, M., and Fiser, J. (2011). Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science *331*, 83–87.

Bertsekas, D. (1982). Distributed dynamic programming. IEEE Trans. Automat. Contr. *27*, 610–616.

Bornstein, A.M., and Norman, K.A. (2017). Reinstated episodic context guides sampling-based decisions for reward. Nat. Neurosci. *20*, 997–1003.

Bornstein, A.M., Khaw, M.W., Shohamy, D., and Daw, N.D. (2017). Reminders of past choices bias decisions for reward in humans. Nat. Commun. *8*, 15958.

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

Callaway, F., Lieder, F., Das, P., Gul, S., Krueger, P.M., and Griffiths, T. (2018). A resource-rational analysis of human planning. In Proceedings of the 40th Annual Meeting of the Cognitive Science Society (Cognitive Science Society), pp. 178–183.

Callaway, F., Rangel, A., and Griffiths, T.L. (2021). Fixation patterns in simple choice reflect optimal information sampling. PLoS Comput. Biol. *17*, e1008863.

Campbell, M., Hoane, A.J., Jr., and Hsu, F.-h. (2002). Deep blue. Artif. Intell. *134*, 57–83.

Carey, A.A., Tanaka, Y., and van der Meer, M.A.A. (2019). Reward revaluation biases hippocampal replay content away from the preferred outcome. Nat. Neurosci. *22*, 1450–1459.

Cazé, R., Khamassi, M., Aubin, L., and Girard, B. (2018). Hippocampal replays under the scrutiny of reinforcement learning models. J. Neurophysiol. *120*, 2877–2896.

Churchland, M.M., Santhanam, G., and Shenoy, K.V. (2006). Preparatory activity in premotor and motor cortex reflects the speed of the upcoming reach. J. Neurophysiol. *96*, 3130–3146.

Cisek, P. (2012). Making decisions through a distributed consensus. Curr. Opin. Neurobiol. *22*, 927–936.

Collins, A.G., and Frank, M.J. (2014). Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. Psychol. Rev. *121*, 337–366.

Constantinescu, A.O., O'Reilly, J.X., and Behrens, T.E.J. (2016). Organizing conceptual knowledge in humans with a gridlike code. Science *352*, 1464–1468.

Correa, C.G., Ho, M.K., Callaway, F., and Griffiths, T.L. (2020). Resource-rational task decomposition to minimize planning costs. arXiv preprint arXiv:2007.13862.

Coulom, R. (2007). Efficient selectivity and backup operators in Monte-Carlo tree search. In Computers and Games (Springer), pp. 72–83.

Craik, K. (1943). The Nature of Explanation (Cambridge University Press).

Curtis, C.E., and D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. Trends Cogn. Sci. *7*, 415–423.

Davidson, T.J., Kloosterman, F., and Wilson, M.A. (2009). Hippocampal replay of extended experience. Neuron *63*, 497–507.

Daw, N.D. (2012). Model-based reinforcement learning as cognitive search: neurocomputational theories. In Cognitive Search: Evolution, Algorithms and the Brain (MIT Press), pp. 195–208.

Daw, N.D., and Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. Philos. Trans. R. Soc. Lond. B Biol. Sci. *369*, 20130478.

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

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron *69*, 1204–1215.

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

Dayan, P. (2009). Goal-directed control and its antipodes. Neural Netw. *22*, 213–219.

Dayan, P., and Abbott, L.F. (2001). Theoretical Neuroscience (MIT Press).

de Groot, A.D. (1978). Thought and Choice in Chess (de Gruyter Mouton).

Dechter, R., and Pearl, J. (1985). Generalized best-first search strategies and the optimality of A*. J. Assoc. Comput. Mach. *32*, 505–536.

Derman, R.C., Schneider, K., Juarez, S., and Delamater, A.R. (2018). Sign-tracking is an expectancy-mediated behavior that relies on prediction error mechanisms. Learn. Mem. *25*, 550–563.

Dezfouli, A., and Balleine, B.W. (2013). Actions, action sequences and habits: evidence that goal-directed and habitual action control are hierarchically organized. PLoS Comput. Biol. *9*, e1003364.

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

Dickinson, A., and Balleine, B. (2002). The role of learning in the operation of motivational systems. In Stevens' Handbook of Experimental Psychology: Learning, Motivation, and Emotion (John Wiley & Sons), pp. 497–533.

Doeller, C.F., and Burgess, N. (2008). Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. Proc. Natl. Acad. Sci. USA *105*, 5909–5914.

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

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

Duvelle, É., Grieves, R.M., Liu, A., Jedidi-Ayoub, S., Holeniewska, J., Harris, A., Nyberg, N., Donnarumma, F., Lefort, J.M., Jeffery, K.J., et al. (2021). Hippocampal place cells encode global location but not connectivity in a complex space. Curr. Biol. *31*, 1221–1233.

Echeveste, R., Aitchison, L., Hennequin, G., and Lengyel, M. (2020). Cortical-like dynamics in recurrent circuits optimized for sampling-based probabilistic inference. Nat. Neurosci. *23*, 1138–1149.

Eichenbaum, H. (2017). On the integration of space, time, and memory. Neuron *95*, 1007–1018.

Eichenbaum, H., and Cohen, N.J. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function? Neuron *83*, 764–770.

Epstein, R.A., Patai, E.Z., Julian, J.B., and Spiers, H.J. (2017). The cognitive map in humans: spatial navigation and beyond. Nat. Neurosci. *20*, 1504–1513.

Fincham, J.M., Carter, C.S., Van Veen, V., Stenger, V.A., and Anderson, J.R. (2002). Neural mechanisms of planning: a computational analysis using event-related fMRI. Proc. Natl. Acad. Sci. USA *99*, 3346–3351.

Fiser, J., Berkes, P., Orbán, G., and Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn. Sci. *14*, 119–130.

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

Foster, D.J., and Wilson, M.A. (2007). Hippocampal theta sequences. Hippocampus *17*, 1093–1099.

Franklin, D.W., and Wolpert, D.M. (2011). Computational mechanisms of sensorimotor control. Neuron *72*, 425–442.

Friedrich, J., and Lengyel, M. (2016). Goal-directed decision making with spiking neurons. J. Neurosci. *36*, 1529–1546.

Gallagher, M., McMahan, R.W., and Schoenbaum, G. (1999). Orbitofrontal cortex and representation of incentive value in associative learning. J. Neurosci. *19*, 6610–6614.

Gallistel, C.R. (1990). The Organization of Learning (MIT Press).

Geffner, H. (2013). Computational models of planning. Wiley Interdiscip. Rev. Cogn. Sci. *4*, 341–356.

Gershman, S.J., and Daw, N.D. (2017). Reinforcement learning and episodic memory in humans and animals: an integrative framework. Annu. Rev. Psychol. *68*, 101–128.

Gershman, S.J., Markman, A.B., and Otto, A.R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. J. Exp. Psychol. Gen. *143*, 182–194.

Gillespie, A.K., Maya, D.A.A., Denovellis, E.L., Liu, D.F., Kastner, D.B., Coulter, M.E., Roumis, D.K., Eden, U.T., and Frank, L.M. (2021). Hippocampal replay reflects specific past experiences rather than a plan for subsequent choice. Neuron *109*, 3149–3163.

Girardeau, G., Benchenane, K., Wiener, S.I., Buzsáki, G., and Zugaro, M.B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. Nat. Neurosci. *12*, 1222–1223.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron *66*, 585–595.

Goel, V., and Grafman, J. (1995). Are the frontal lobes implicated in "planning" functions? Interpreting data from the Tower of Hanoi. Neuropsychologia *33*, 623–642.

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

Gremel, C.M., and Costa, R.M. (2013). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. Nat. Commun. *4*, 2264.

Gridchyn, I., Schoenenberger, P., O'Neill, J., and Csicsvari, J. (2020). Assembly-specific disruption of hippocampal replay leads to selective memory deficit. Neuron *106*, 291–300.

Gupta, A.S., van der Meer, M.A., Touretzky, D.S., and Redish, A.D. (2010). Hippocampal replay is not a simple function of experience. Neuron *65*, 695–705.

Ha, D., and Schmidhuber, J. (2018). World models. arXiv preprint arXiv:1803.10122.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. (2019a). Dream to control: learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603.

Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., and Davidson, J. (2019b). Learning latent dynamics for planning from pixels. In Proceedings of the 36th International Conference on Machine Learning (PMLR), pp. 2555–2565.

Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Pfaff, T., Weber, T., Buesing, L., and Battaglia, P.W. (2019). Combining Q-learning and search with amortized value estimates. arXiv preprint arXiv:1912.02807.

Harris, C.M., and Wolpert, D.M. (1998). Signal-dependent noise determines motor planning. Nature *394*, 780–784.

Hart, P.E., Nilsson, N.J., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. IEEE Trans. Syst. Sci. Cybern. *4*, 100–107.

Hong, G., and Lieber, C.M. (2019). Novel electrode technologies for neural recordings. Nat. Rev. Neurosci. *20*, 330–345.

Horvitz, E. (1987). Reasoning about beliefs and actions under computational resource constraints. In Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence (AAAI), pp. 429–444.

Hoshi, E., Shima, K., and Tanji, J. (2000). Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. J. Neurophysiol. *83*, 2355–2373.

Houlsby, N.M., Huszár, F., Ghassemi, M.M., Orbán, G., Wolpert, D.M., and Lengyel, M. (2013). Cognitive tomography reveals complex, task-independent mental representations. Curr. Biol. *23*, 2169–2175.

Howard, L.R., Javadi, A.H., Yu, Y., Mill, R.D., Morrison, L.C., Knight, R., Loftus, M.M., Staskute, L., and Spiers, H.J. (2014). The hippocampus and entorhinal cortex encode the path and Euclidean distances to goals during navigation. Curr. Biol. *24*, 1331–1340.

Howard, R.A. (1960). Dynamic Programming and Markov Processes (MIT Press).

Hunt, L.T., and Hayden, B.Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. Nat. Rev. Neurosci. *18*, 172–182.

Hunt, L.T., Malalasekera, W.M.N., de Berker, A.O., Miranda, B., Farmer, S.F., Behrens, T.E.J., and Kennerley, S.W. (2018). Triple dissociation of attention and decision computations across prefrontal cortex. Nat. Neurosci. *21*, 1471–1481.

Hunt, L.T., Daw, N.D., Kaanders, P., MacIver, M.A., Mugan, U., Procyk, E., Redish, A.D., Russo, E., Scholl, J., Stachenfeld, K., et al. (2021). Formalizing planning and information search in naturalistic decision-making. Nat. Neurosci. *24*, 1051–1064.

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

Huys, Q.J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S.J., Dayan, P., and Roiser, J.P. (2015). Interplay of approximate planning strategies. Proc. Natl. Acad. Sci. USA *112*, 3098–3103.

Igata, H., Ikegaya, Y., and Sasaki, T. (2021). Prioritized experience replays on a hippocampal predictive map for learning. Proc. Natl. Acad. Sci. USA *118*, e2011266118.

Jadhav, S.P., Kemere, C., German, P.W., and Frank, L.M. (2012). Awake hippocampal sharp-wave ripples support spatial memory. Science *336*, 1454–1458.

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

Jones, J.L., Esber, G.R., McDannald, M.A., Gruber, A.J., Hernandez, A., Mirenzi, A., and Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. Science *338*, 953–956.

Jun, J.J., Steinmetz, N.A., Siegle, J.H., Denman, D.J., Bauza, M., Barbarits, B., Lee, A.K., Anastassiou, C.A., Andrei, A., Aydın, Ç., et al. (2017). Fully integrated silicon probes for high-density recording of neural activity. Nature *551*, 232–236.

Kahneman, D. (2011). Thinking, Fast and Slow (MacMillan).

Káli, S., and Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. Nat. Neurosci. *7*, 286–294.

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

Kay, K., Chung, J.E., Sosa, M., Schor, J.S., Karlsson, M.P., Larkin, M.C., Liu, D.F., and Frank, L.M. (2020). Constant sub-second cycling between representations of possible futures in the hippocampus. Cell *180*, 552–567.

Keramati, M., Smittenaar, P., Dolan, R.J., and Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. Proc. Natl. Acad. Sci. USA *113*, 12868–12873.

Koblinger, Á., Fiser, J., and Lengyel, M. (2021). Representations of uncertainty: where art thou? Curr. Opin. Behav. Sci. *38*, 150–162.

Korf, R.E. (1985). Depth-first iterative-deepening. Artif. Intell. *27*, 97–109.

Korf, R.E. (1987). Planning as search: a quantitative approach. Artif. Intell. *33*, 65–88.

Korf, R.E. (1990). Real-time heuristic search. Artif. Intell. *42*, 189–211.

Kotovsky, K., Hayes, J.R., and Simon, H.A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. Cogn. Psychol. *17*, 248–294.

Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. Neuron *93*, 480–490.

Krusche, M.J., Schulz, E., Guez, A., and Speekenbrink, M. (2018). Adaptive planning in human search. bioRxiv, 268938.

LaValle, S.M. (2006). Planning Algorithms (Cambridge University Press).

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

Lehnert, L., Littman, M.L., and Frank, M.J. (2020). Reward-predictive representations generalize across tasks in reinforcement learning. PLoS Comput. Biol. *16*, e1008317.

Lengyel, M., and Dayan, P. (2008). Hippocampal contributions to control: the third way. In Advances in Neural Information Processing Systems 20 (MIT Press), pp. 889–896.

Lieder, F., and Griffiths, T.L. (2019). Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. Behav. Brain Sci. *43*, 1–60.

Liu, Y., Mattar, M.G., Behrens, T.E.J., Daw, N.D., and Dolan, R.J. (2021). Experience replay is associated with efficient nonlocal learning. Science *372*, eabf1357.

Luk, C.-H., and Wallis, J.D. (2013). Choice coding in frontal cortex during stimulus-guided or action-guided decision-making. J. Neurosci. *33*, 1864–1871.

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., and Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci. *21*, 1281–1289.

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

McClelland, J.L., McNaughton, B.L., and O'Reilly, R.C. (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*, 419–457.

McNamee, D., and Wolpert, D.M. (2019). Internal models in biological control. Annu. Rev. Control Robot. Auton. Syst. *2*, 339–364.

McNamee, D., Wolpert, D.M., and Lengyel, M. (2016). Efficient state-space modularization for planning: theory, behavioral and neural signatures. In Advances in Neural Information Processing Systems 29 (MIT Press), pp. 4511–4519.

Miller, E.K., and Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. Annu. Rev. Neurosci. *24*, 167–202.

Miller, K.J., Botvinick, M.M., and Brody, C.D. (2017). Dorsal hippocampus contributes to model-based planning. Nat. Neurosci. *20*, 1269–1276.

Miranda, B., Malalasekera, W.M.N., Behrens, T.E., Dayan, P., and Kennerley, S.W. (2020). Combined model-free and model-sensitive reinforcement learning in non-human primates. PLoS Comput. Biol. *16*, e1007944.

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

Moore, A.W., and Atkeson, C.G. (1993). Prioritized sweeping: reinforcement learning with less data and less time. Mach. Learn. *13*, 103–130.

Moore, E.F. (1959). The shortest path through a maze. In Proceedings of an International Symposium on the Theory of Switching (Harvard University Press), pp. 285–292.

Moser, E.I., Moser, M.-B., and McNaughton, B.L. (2017). Spatial representation in the hippocampal formation: a history. Nat. Neurosci. *20*, 1448–1464.

Mushiake, H., Saito, N., Sakamoto, K., Itoyama, Y., and Tanji, J. (2006). Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. Neuron *50*, 631–641.

Newell, A., and Simon, H. (1956). The logic theory machine – a complex information processing system. IEEE Trans. Inform. Theory *2*, 61–79.

Newell, A., Shaw, J.C., and Simon, H.A. (1959). Report on a general problem solving program. In Proceedings of the International Conference on Information Processing (UNESCO), pp. 256–264.

Nilsson, N.J. (2014). Principles of Artificial Intelligence (Morgan Kaufmann).

Norman, K.A., Polyn, S.M., Detre, G.J., and Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. *10*, 424–430.

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

O'Reilly, R.C., and Rudy, J.W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. Psychol. Rev. *108*, 311–345.

Obeso, J.A., Jahanshahi, M., Alvarez, L., Macias, R., Pedroso, I., Wilkinson, L., Pavon, N., Day, B., Pinto, S., Rodríguez-Oroz, M.C., et al. (2009). What can man do without basal ganglia motor output? The effect of combined unilateral subthalamotomy and pallidotomy in a patient with Parkinson's disease. Exp. Neurol. *220*, 283–292.

Ólafsdóttir, H.F., Barry, C., Saleem, A.B., Hassabis, D., and Spiers, H.J. (2015). Hippocampal place cells construct reward related sequences through unexplored space. eLife *4*, e06063.

Owen, A.M. (1997). Cognitive planning in humans: neuropsychological, neuroanatomical and neuropharmacological perspectives. Prog. Neurobiol. *53*, 431–450.

Padoa-Schioppa, C., and Assad, J.A. (2006). Neurons in the orbitofrontal cortex encode economic value. Nature *441*, 223–226.

Papale, A.E., Zielinski, M.C., Frank, L.M., Jadhav, S.P., and Redish, A.D. (2016). Interplay between hippocampal sharp-wave-ripple events and vicarious trial and error behaviors in decision making. Neuron *92*, 975–982.

Patai, E.Z., and Spiers, H.J. (2021). The versatile wayfinder: prefrontal contributions to spatial navigation. Trends Cogn. Sci. *25*, 520–533.

Pauli, W.M., Gentile, G., Collette, S., Tyszka, J.M., and O'Doherty, J.P. (2019). Evidence for model-based encoding of Pavlovian contingencies in the human brain. Nat. Commun. *10*, 1099.

Pearl, J. (1984). Heuristics: Intelligent Search Strategies for Computer Problem Solving (Addison-Wesley Publishing Company).

Penfield, W., and Evans, J. (1935). The frontal lobe in man: a clinical study of maximum removals. Brain *58*, 115–133.

Pezzulo, G., Donnarumma, F., Maisto, D., and Stoianov, I. (2019). Planning at decision time and in the background during spatial navigation. Curr. Opin. Behav. Sci. *29*, 69–76.

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

Piray, P., and Daw, N.D. (2021). Linear reinforcement learning in planning, grid fields, and cognitive control. Nat. Commun. *12*, 4942.

Raby, C.R., Alexis, D.M., Dickinson, A., and Clayton, N.S. (2007). Planning for the future by western scrub-jays. Nature *445*, 919–921.

Ragozzino, M.E., Ragozzino, K.E., Mizumori, S.J., and Kesner, R.P. (2002). Role of the dorsomedial striatum in behavioral flexibility for response and visual cue discrimination learning. Behav. Neurosci. *116*, 105–115.

Ramus, S.J., Davis, J.B., Donahue, R.J., Discenza, C.B., and Waite, A.A. (2007). Interactions between the orbitofrontal cortex and the hippocampal memory system during the storage of long-term memory. Ann. N. Y. Acad. Sci. *1121*, 216–231.

Redish, A.D. (2016). Vicarious trial and error. Nat. Rev. Neurosci. *17*, 147–159.

Rich, E.L., and Wallis, J.D. (2016). Decoding subjective decisions from orbitofrontal cortex. Nat. Neurosci. *19*, 973–980.

Roesch, M.R., and Olson, C.R. (2004). Neuronal activity related to reward value and motivation in primate frontal cortex. Science *304*, 307–310.

Rudebeck, P.H., and Murray, E.A. (2014). The orbitofrontal oracle: cortical mechanisms for the prediction and evaluation of specific behavioral outcomes. Neuron *84*, 1143–1156.

Rudebeck, P.H., Saunders, R.C., Prescott, A.T., Chau, L.S., and Murray, E.A. (2013). Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat. Neurosci. *16*, 1140–1145.

Ruediger, S., Spirig, D., Donato, F., and Caroni, P. (2012). Goal-oriented searching mediated by ventral hippocampus early in trial-and-error learning. Nat. Neurosci. *15*, 1563–1571.

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

Russell, S., and Norvig, P. (2010). Artificial Intelligence: A Modern Approach (Prentice Hall).

Rust, N.C., and Movshon, J.A. (2005). In praise of artifice. Nat. Neurosci. *8*, 1647–1650.

Sadacca, B.F., Wied, H.M., Lopatina, N., Saini, G.K., Nemirovsky, D., and Schoenbaum, G. (2018). Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. eLife *7*, e30373.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. Nature *588*, 604–609.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Schultz, W., Tremblay, L., and Hollerman, J.R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. Cereb. Cortex *10*, 272–284.

Scoville, W.B., and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. J. Neurol. Neurosurg. Psych. *20*, 11–21.

Shallice, T. (1982). Specific impairments of planning. Philos. Trans. R. Soc. Lond. B Biol. Sci. *298*, 199–209.

Shohamy, D., and Daw, N.D. (2015). Integrating memories to guide decisions. Curr. Opin. Behav. Sci. *5*, 85–90.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. Nature *529*, 484–489.

Silver, D., Hasselt, H., Hessel, M., Schaul, T., Guez, A., Harley, T., Dulac-Arnold, G., Reichert, D., Rabinowitz, N., Barreto, A., et al. (2017a). The predictron: end-to-end learning and planning. In Proceedings of the 34th International Conference on Machine Learning (PMLR), pp. 3191–3199.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017b). Mastering the game of Go without human knowledge. Nature *550*, 354–359.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science *362*, 1140–1144.

Simon, D.A., and Daw, N.D. (2011). Neural correlates of forward planning in a spatial decision task in humans. J. Neurosci. *31*, 5526–5539.

Singer, A.C., Carr, M.F., Karlsson, M.P., and Frank, L.M. (2013). Hippocampal SWR activity predicts correct decisions during the initial learning of an alternation task. Neuron *77*, 1163–1173.

Snider, J., Lee, D., Poizner, H., and Gepshtein, S. (2015). Prospective optimization with limited resources. PLoS Comput. Biol. *11*, e1004501.

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

Solway, A., and Botvinick, M.M. (2015). Evidence integration in model-based tree search. Proc. Natl. Acad. Sci. USA *112*, 11708–11713.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A.G., Niv, Y., and Botvinick, M.M. (2014). Optimal behavioral hierarchy. PLoS Comput. Biol. *10*, e1003779.

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

Suddendorf, T., and Corballis, M.C. (1997). Mental time travel and the evolution of the human mind. Genet. Soc. Gen. Psychol. Monogr. *123*, 133–167.

Sutton, R.S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. SIGART Bull. *2*, 160–163.

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

Sutton, R.S., McAllester, D.A., Singh, S.P., and Mansour, Y. (1999a). Policy gradient methods for reinforcement learning with function approximation. In Advances in Neural Information Processing Systems 12 (MIT Press), pp. 1057–1063.

Sutton, R.S., Precup, D., and Singh, S. (1999b). Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. Artif. Intell. *112*, 181–211.

Talvitie, E. (2017). Self-correcting models for model-based reinforcement learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI), pp. 2597–2603.

Tesauro, G., and Galperin, G. (1996). On-line policy improvement using Monte-Carlo search. In Advances in Neural Information Processing Systems 9 (MIT Press), pp. 1068–1074.

Todorov, E., and Jordan, M.I. (2002). Optimal feedback control as a theory of motor coordination. Nat. Neurosci. *5*, 1226–1235.

Tolman, E.C. (1938). The determiners of behavior at a choice point. Psychol. Rev. *45*, 1–41.

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

Toussaint, M., and Storkey, A. (2006). Probabilistic inference for solving discrete and continuous state Markov decision processes. In Proceedings of the 23rd International Conference on Machine Learning (ACM), pp. 945–952.

Tse, D., Langston, R.F., Kakeyama, M., Bethus, I., Spooner, P.A., Wood, E.R., Witter, M.P., and Morris, R.G.M. (2007). Schemas and memory consolidation. Science *316*, 76–82.

Unterrainer, J.M., and Owen, A.M. (2006). Planning and problem solving: from neuropsychology to functional neuroimaging. J. Physiol. Paris *99*, 308–317.

Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the neural substrates of goal-directed learning in the human brain. J. Neurosci. *27*, 4019–4026.

van Opheusden, B., Galbiati, G., Bnaya, Z., Li, Y., and Ma, W.J. (2017). A computational model for decision tree search. In Proceedings of the 39th Annual Meeting of the Cognitive Science Society (Cognitive Science Society), pp. 1254–1259.

Vikbladh, O.M., Meager, M.R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N., and Daw, N.D. (2019). Hippocampal contributions to model-based planning and spatial memory. Neuron *102*, 683–693, e4.

Wahlström, N., Schön, T.B., and Deisenroth, M.P. (2015). From pixels to torques: policy learning with deep dynamical models. arXiv preprint arXiv:1502.02251.

Wallis, J.D., and Miller, E.K. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. Eur. J. Neurosci. *18*, 2069–2081.

Wang, J.X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci. *21*, 860–868.

Whittington, J.C.R., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T.E.J. (2020). The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. Cell *183*, 1249–1263.

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

Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. Neuron *81*, 267–279.

Wong, K.-F., and Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. J. Neurosci. *26*, 1314–1328.

Wood, R.A., Bauza, M., Krupic, J., Burton, S., Delekate, A., Chan, D., and O'Keefe, J. (2018). The honeycomb maze provides a novel test to study hippocampal-dependent spatial navigation. Nature *554*, 102–105.

Wu, Z., Kwon, M., Daptardar, S., Schrater, P., and Pitkow, X. (2020). Rational thoughts in neural codes. Proc. Natl. Acad. Sci. USA *117*, 29311–29320.

Wunderlich, K., Dayan, P., and Dolan, R.J. (2012). Mapping value based planning and extensively trained choice in the human brain. Nat. Neurosci. *15*, 786–791.

Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2005a). Blockade of NMDA receptors in the dorsomedial striatum prevents action–outcome learning in instrumental conditioning. Eur. J. Neurosci. *22*, 505–512.

Yin, H.H., Ostlund, S.B., Knowlton, B.J., and Balleine, B.W. (2005b). The role of the dorsomedial striatum in instrumental conditioning. Eur. J. Neurosci. *22*, 513–523.

Zénon, A. (2019). Eye pupil signals information gain. Proc. R. Soc. Lond. B Biol. Sci. *286*, 20191593.