AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

ISSN: 2325-9965

2025, Vol. 12, No. 4, 334–355
https://doi.org/10.1037/dec0000267

Uncertainty-Driven Exploration During Planning

Haoxue Fan[1], Frederick Callaway[2, 3], and Samuel J. Gershman[3, 4, 5]

[1] Carney Institute for Brain Science, Brown University
[2] Department of Psychology, New York University
[3] Department of Psychology, Harvard University
[4] Center for Brain Science, Harvard University
[5] Center for Brains, Minds, and Machines, Massachusetts Institute of Technology

In complex environments, the space of possible plans is vast. Generating a good plan therefore requires judicious selection of which parts of the plan space to mentally explore. Drawing on past studies of human exploration, we propose that mental exploration might invoke similar mechanisms. In particular, we test the hypothesis that mental exploration during planning is uncertainty-driven, such that people will exhibit a tendency to explore parts of the plan space that have high epistemic uncertainty. We developed a route-planning task, displayed as a binary tree, where participants were instructed to collect as many treats (rewards) as possible by traversing the tree. By separating the planning and execution phases, we encouraged participants to externalize their planning process. We manipulated uncertainty by varying the number of potential future states available from each current state. Across two studies, the data suggest that people preferred to explore options with more successor states after controlling for value differences, supporting the uncertainty-driven planning hypothesis. We also found that uncertainty played a larger role during the planning phase than during the execution phase, consistent with the hypothesis that the uncertainty effect primarily reflects a property of human planning algorithms rather than an intrinsic preference for uncertainty.

Keywords: uncertainty, planning, exploration

Supplemental materials: https://doi.org/10.1037/dec0000267.supp

Tim Rakow served as action editor.

Haoxue Fan https://orcid.org/0000-0003-3967-6457
Frederick Callaway https://orcid.org/0000-0001-7687-5987
Samuel J. Gershman https://orcid.org/0000-0002-6546-3298

Deidentified behavioral data as well as analysis code have been made publicly available via the Open Science Framework and can be assessed at https://osf.io/t2cmr. The preregistration form for Experiments 1 and 2 are available at https://aspredicted.org/83Y_MN8 and https://aspredicted.org/YTG_4K2. The data and materials used in the present study have not been disseminated before.

The authors declare no conflicts of interest. This work was supported by Grant FA9550-20-1-0413 from the Air Force Office of Scientific Research grant to Samuel J. Gershman, the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University, and a Stimson research grant award to Haoxue Fan from the Harvard University Department of Psychology. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the article. The authors thank Yuqing Lei, Hayley Dorfman, Peng Qian, and Gershman Lab members for helpful discussions.

Haoxue Fan played a lead role in conceptualization, data curation, investigation, formal analysis, methodology, visualization, software, writing–original draft, and writing–review and editing and a supporting role in methodology and funding acquisition. Frederick Callaway played an equal role in writing–review and editing and a supporting role in conceptualization, formal analysis, methodology, software,

continued

Planning, the process of using a world model to inform behavior, is an important capacity for an organism to flexibly direct their behavior toward goals, especially when achieving a goal requires a sequence of many actions—that is, multistep planning (Miller & Venditto, 2021). Finding the optimal plan in a large environment is notoriously intractable, due to the combinatorial explosion of possible decision sequences. Yet planning is also ubiquitous: From spending a day in a new city to preparing a future career, planning underlies many real-life sequential decision problems. For this to be possible, the brain must use algorithms that intelligently search the space of decision sequences without brute-force enumeration.

For inspiration, we can look to planning algorithms that have been implemented in machines (LaValle, 2006). The fundamental object of study is the *decision tree*, where each node corresponds to a state and each edge corresponds to an action (Figure 1). The root node represents the agent's current state. Choosing an action in a particular state moves the agent along the corresponding edge to a new state. Classical planning algorithms either exhaustively enumerate the possible actions at a given level of the decision tree before choosing one and moving to the next level (breadth-first search) or exhaustively enumerate the possible actions along a single branch of the decision tree before moving to the next branch (depth-first search). Both approaches can fail when the state space is very large or the optimal plans are very long. More efficient algorithms selectively search along particular paths based on an evaluation function that ranks the actions at each state (best-first search). The basic challenge for these algorithms is to define a good evaluation function that can be easily computed. A good choice of evaluation function may be problem-specific, hindering the generic application of such algorithms.

An important insight into the design of efficient planning algorithms came from a connection with the exploration–exploitation dilemma in reinforcement learning (Sutton & Barto, 2018). An agent interacting with an environment faces the problem of simultaneously optimizing reward and gathering information. The agent can choose to exploit its current action value estimates, but this may yield suboptimal reward if the estimates are poor. The agent can improve the estimates by exploring the environment, but this runs the risk of incurring an opportunity cost if the explored states have low reward. Although the optimal algorithm for balancing exploration and exploitation is intractable, uncertainty-directed reinforcement learning algorithms have been highly successful (Auer, 2002; Ciosek et al., 2019; Dayan & Sejnowski, 1996; Srinivas et al., 2010). In particular, these algorithms add an "uncertainty bonus" to the action values based on the agent's ignorance about the true value. The Upper Confidence Bound algorithm (Auer, 2002), for example, defines the uncertainty bonus based on a confidence interval around the value estimate. By taking actions that have high upper confidence bounds, the agent focuses their exploration on actions whose value could be much higher than currently estimated.

It might seem that the exploration–exploitation dilemma does not apply in the case of planning: Exploration in reinforcement learning involves taking actions in an unknown environment, whereas planning involves thinking about actions in a known environment. However, both problem settings involve reducing uncertainty about state-action values by traversing the state space (Hunt et al., 2021). Another difference between reinforcement learning and planning is that there is no opportunity for true exploitation while planning, since the agent does not actually receive the rewards associated with simulated actions. Here, the analogy is less precise. However, note that the goal of planning is to find a high-value sequence of actions. If an action already has high estimated value, it is more likely to be part of such a sequence. The agent can thus "exploit" this knowledge to focus their search on more promising actions—indeed, this is precisely the idea behind best-first search.

and supervision. Samuel J. Gershman played a lead role in methodology, project administration, funding acquisition, and supervision, an equal role in writing–review and editing, and a supporting role in conceptualization and formal analysis.

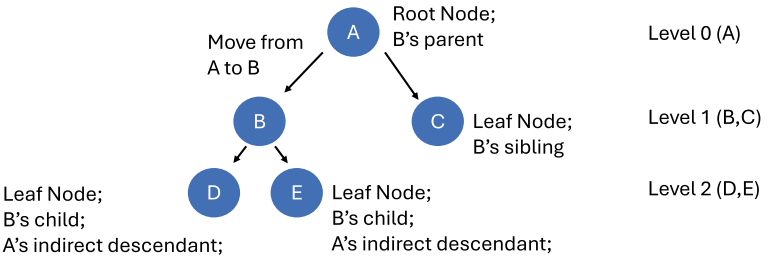
The data are available at https://osf.io/t2cmr.

The experimental materials are available at https://osf.io/t2cmr.

The preregistered design is available at https://aspredicted.org/83Y_MN8 and https://aspredicted.org/YTG_4K2.

Correspondence concerning this article should be addressed to Haoxue Fan, Carney Institute for Brain Science, Brown University, 164 Angell Street, Providence, RI 02906, United States. Email: haoxue_fan@brown.edu

Figure 1
Illustration of a Decision Tree



Note. This example decision tree consists of five nodes and three layers. Each node reflects a state and the root node (Node A) is the agent's current state. Edges (e.g., A → B) indicate choosing an action in a specific state (Node A) and moving to a new state (Node B). The node(s) that falls under another node is termed as the child node (e.g., D is B's child), and the current node is termed as the parent node (e.g., B is D's parent). Child node's child node is termed as the indirect descendant (e.g., D is A's indirect descendant). Nodes that do not have a child node are termed leaf nodes (D and E in the example). Nodes that share the same parent are sibling nodes (e.g., B and C are siblings). See the online article for the color version of this figure.

These arguments suggest that planning presents a form of exploration–exploitation dilemma, in which an agent must strike a balance between refining promising plans (exploitation) and seeking out new ones (exploration). This idea has been implemented in many different ways (e.g., Bellman, 1956; Sanner et al., 2009; Sutton, 1990). One notable example is upper confidence bounds applied to trees (Kocsis & Szepesvári, 2006), the core mechanism of modern Monte Carlo tree search algorithms (Browne et al., 2012). In upper confidence bounds applied to trees, the agent simulates several action sequences (or "rollouts"), applying a variant of Upper Confidence Bound algorithm (Auer, 2002) to decide which action to simulate at each step.

There is already considerable evidence that people use uncertainty-directed algorithms for reinforcement learning (Fan, Burke, et al., 2023; Fan, Gershman, & Phelps, 2023; Frank et al., 2009; Gershman, 2018a, 2019; Schulz et al., 2020; Speekenbrink & Konstantinidis, 2015; Wu et al., 2018, 2021) and that these algorithms also appear to be used in real-world environments such as food purchasing (Schulz et al., 2019). However, this work almost exclusively focuses on one-step decision making ("bandit" tasks), in which one's action only determines the immediate reward. In contrast, the present study focuses on multistep decision making, in which one's action additionally determines the next state (and therefore future rewards). Performing well on such problems generally requires constructing a plan before making the first choice. Accordingly, one must explore the available options within one's own mind rather than in the world. We hypothesize that people direct this internal exploration using the same kind of reinforcement learning strategies that direct their external exploration. Here, we explore whether people use uncertainty as a heuristic approximation to optimal exploration while planning.

In this article, we report two experiments designed to study the hypothesis that people use a similar uncertainty-directed algorithm for planning in a multistep decision-making setting.[1] The key idea is to create situations where participants have varying levels of epistemic uncertainty about different branches of the decision tree. We accomplish this by varying the number of children nodes (branching factor) at different nodes in the tree. All else being equal, there will be greater uncertainty in the values of nodes with more children because these nodes lead to a greater number of possible future rewards. The uncertainty about which of those rewards will actually be attained as well as

[1] Note that bandit problems also present a certain kind of multistep problem, in that one's current action provides information that may inform one's later actions. Formally, bandits can be modeled as a sequential problem where the states correspond to beliefs about the reward rate of each bandit (Gittins, 1979). Here, however, we focus on problems where the external environment has sequential structure (what one typically means by "multistep decision making").

uncertainty in the rewards themselves will both contribute to higher uncertainty in the node's value. Experiment 1 was designed to test whether people demonstrate a preference to visit the node with more children during planning. Experiment 2 aims at replicating Experiment 1's findings over a more diverse set of task structure and further examined whether people would also prefer to approach the node with more indirect descendants, which is a more farsighted indicator of uncertainty.

One methodological challenge in studying the dynamics of planning in real time is that planning is a mental process that is not directly observable. To address this, planning researchers often use process-tracing methods that make aspects of the planning process observable, for example, think aloud (De Groot, 1965; Newell & Simon, 1972), eye tracking (Callaway et al., 2024; Cristín et al., 2022; Kadner et al., 2023; van Opheusden et al., 2023; Zhu et al., 2022), and mouse tracking (Callaway, van Opheusden, et al., 2022; Eluchans et al., 2025). Here, we adopt a similar approach, providing participants with an explicit interface to perform rollouts before they commit to a plan. Forcing participants to structure their planning in this way carries two benefits. First, it prevents participants from using more flexible strategies, like best-first search, whose cognitive cost is artificially reduced by presenting the full environment in a visual display. Second, it allows us to directly compare participant data with rollout-based planning algorithms.

Method

The experiment design, sample size, exclusion criteria, and primary data analysis plan for Experiments 1 and 2 were preregistered at https://aspredicted.org/83Y_MN8 and https://aspredicted.org/YTG_4K2 (Fan et al., 2025b, 2025c). This study was approved by the Harvard University Committee on the Use of Human Subjects (IRB19-0789) and conformed to American Psychological Association ethical standards. Data and code to regenerate results are publicly available at https://osf.io/t2cmr (Fan et al., 2025a).

Participants

Participants were recruited via the Prolific platform, and informed consent was given prior to testing. Participants were excluded if they did not utilize the PLAN phase or used the PLAN phase for one trajectory in >50% of all trials. We also excluded participants who chose the more rewarding option <60% in the EXECUTE phase (see the Experiment Design section for more information on different phases of the experiment). The exclusion criteria were preregistered. We recruited 95 participants in total (Experiment 1: $N = 42$; Experiment 2: $N = 53$), and the final sample size is 39 for Experiment 1 (25 male, 14 female; age: $M = 36.8$, $SD = 10.3$, range 21–59) and 45 for Experiment 2 (20 male, 25 female; age: $M = 33.9$, $SD = 7.6$, range 19–56).

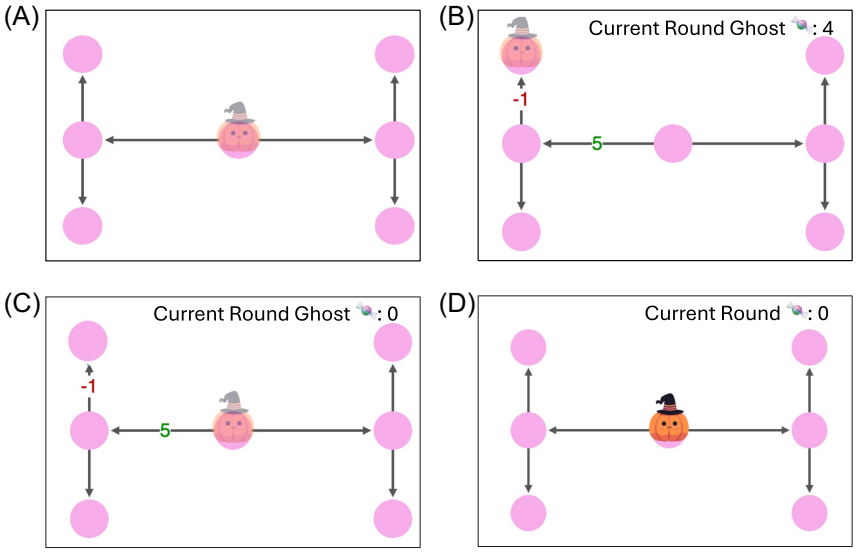
Experiment Design

Trick-or-Treat Game

Our game is a modified version of the Mouselab-MDP paradigm, which extends the approach of the Mouselab paradigm to a general class of planning tasks known as Markov decision processes and has been widely used for studying human planning (Callaway et al., 2017; Callaway, Jain, et al., 2022; He & Lieder, 2023; Jain et al., 2023). On each trial, participants were presented with a route-planning problem displayed as a binary tree (Figure 2). As part of the cover story, they were told that each node represents a house, and their goal was to collect as many treats (reward) as possible by visiting a series of interconnected houses (nodes). See Appendix A for full instructions. Participants moved between nodes using the arrow keys. Participants are only allowed to move in the direction specified by the arrows connecting the houses (Figure 2), and each node can at maximum be visited once in each route. Upon arriving at a node, they received a prespecified reward, drawn from a normal distribution, $N(\mu = 0, \sigma^2 = 2.25)$. Critically, the reward associated with each node was consistent within a trial but was randomly sampled on each new trial. Thus, participants could learn about the value of a path on the current trial but could not do any useful learning across trials. Participants are also explicitly instructed that the goodness of a route is independent of its location and direction to prevent them from making spurious generalization about the goodness of a specific path across trials.

There were two phases in the game: PLAN and EXECUTE. During the PLAN phase—referred to as "ghost mode" in the instructions—participants could not collect treats, but they could simulate possible action sequences as if

Figure 2
Task Schematic



Note. Panel A: Participants started in the PLAN phase, indicated by a transparent avatar. The treats (rewards) were briefly flashed at the beginning of each trial (not shown here). Panel B: Participants used arrow keys to traverse the tree. Reward along the same trajectory remained visible during the rollout. Panel C: During the PLAN phase, participants can choose to restart from the root node for additional rollouts. The previously revealed reward remained visible throughout the PLAN phase. Participants had up to 20 s to carry out rollouts. They could also choose to terminate the PLAN phase earlier if they were ready for the EXECUTE phase. Panel D: When entering EXECUTE phase, indicated by a solid avatar, the rewards revealed during PLAN phase disappeared. Participants committed to one route and received the accumulated reward on the specific route. See the online article for the color version of this figure.

they were actually executing them (i.e., perform rollouts). Similar to the EXECUTE phase, participants move between nodes in the PLAN phase using arrow keys in the direction specified by the arrows on the screen. Immediately before the PLAN phase, the rewards at all nodes were displayed for 500 ms. This provides participants with a rough idea of the reward function, allowing them to direct their rollouts toward higher rewards if they so desired. At each moment during the PLAN phase, participants could either move to an adjacent node and reveal its reward (arrow keys), jump back to the starting node (space), or end the PLAN phase (letter key t). Switching between phases is one-directional: participants cannot reinitiate PLAN mode after deciding to end it. These, respectively, correspond to continuing a rollout, cutting off a single rollout early, and terminating the planning process. Any reward revealed during the PLAN phase remained visible until the next phase began. The revealed reward is identical to the outcome of the node during in the EXECUTE phase—that is, participants are provided with the true payoff and do not need to sample repeatedly to form an estimate. The total reward of the current rollout is displayed on the screen to assist evaluating the goodness of the current path. After 20 s, the PLAN phase was automatically terminated (if the participant had not already done so). During the EXECUTE phase, participants committed to one route and received the total reward associated with all the nodes they visited. Participants were incentivized with a monetary bonus and were told that the bonus is proportional to the total number of treats they have collected during the EXECUTE phase throughout the experiment. They were also explicitly instructed that the treats collected during the PLAN phase do not count toward

the final monetary bonus. The treats-to-bonus conversion scale is 0.01 and is capped at $1 for Experiment 1 ($M$ = 0.91, SD = 0.26; base payment $5) and $2 for Experiment 2 ($M$ = 1.53, SD = 0.56; base payment $13).

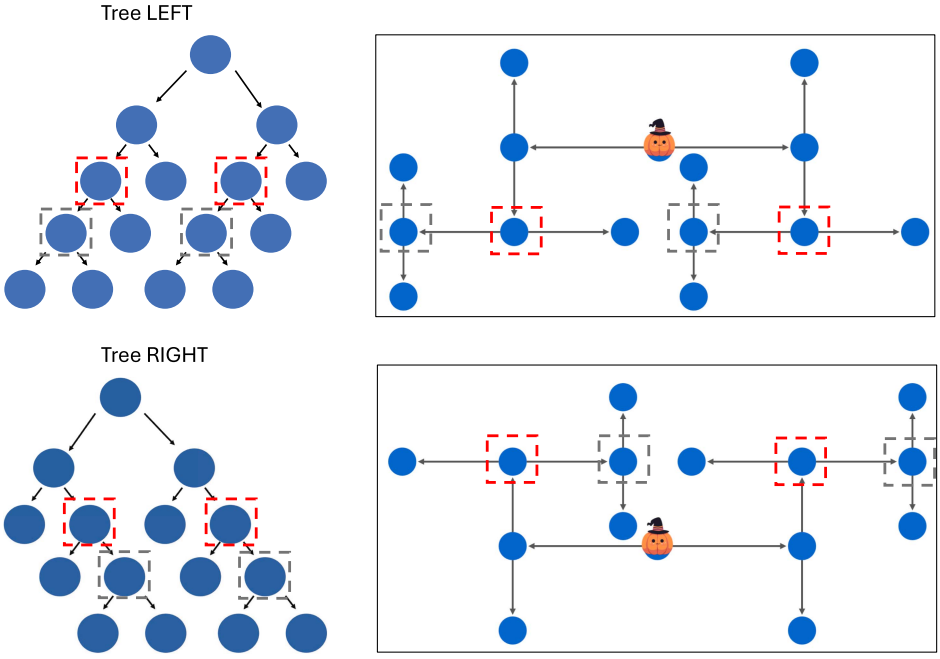
In Experiment 1, participants encountered two types of tree structure (LEFT/RIGHT) and completed eight trials each (Figure 3). Both tree structures have five levels, including the root node (Level 0). The root node has two children (Level 1), both of which have two children (Level 2) as well. The left (right) child node of nodes at Levels 2 and 3 in LEFT (RIGHT) tree has two children, and the right (left) child node of nodes at Levels 2 and 3 in LEFT (RIGHT) tree has zero children. Nodes at Level 4 are all leaf nodes. Trials with LEFT and RIGHT tree structures were intermixed.

Experiment 2 expanded the tree structure repertoire that participants could interact with. We generated all 68 tree structures that have eight leaf nodes, have at least one branch that reaches five levels, and each node has either zero or two child nodes. The LEFT and RIGHT tree structures used in Experiment 1 satisfied the above requirements and were included in the set of 68 unique tree structures. Participants encountered each tree structure once. In both experiments, the reward of each node is resampled at the beginning of a new trial—that is, each trial is associated with a unique reward map. An experiment demo can be found here at https://lu23taj5xe.cognition.run.

Planning-as-Exploration Model

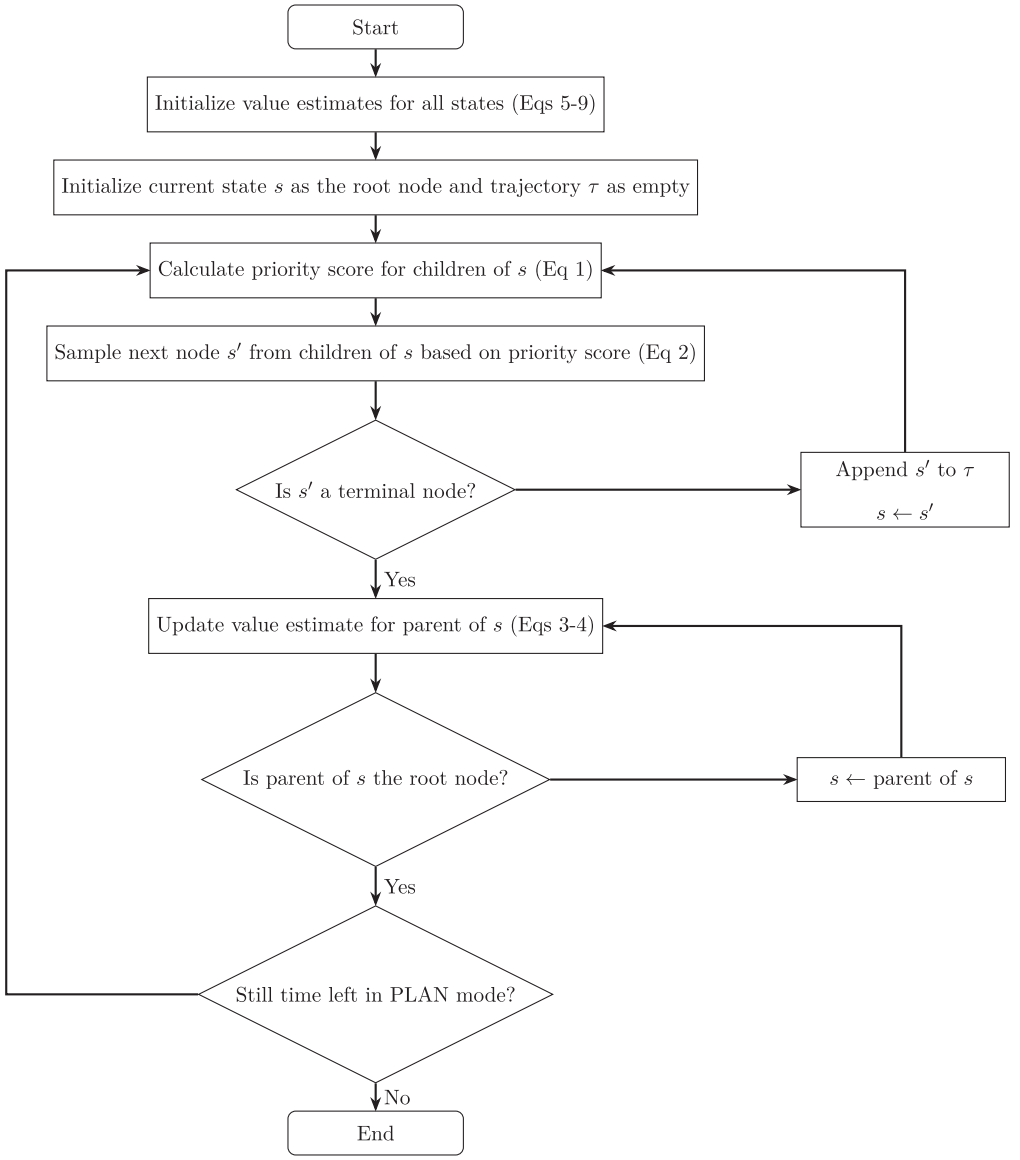
We hypothesized that people would use an exploration algorithm to determine which node to visit next when performing rollouts in the PLAN phase. Algorithm 1 provides the pseudocode for the model, and a flowchart of the model is shown in Figure 4. Specifically, our model uses a Bayesian variant of the Upper Confidence Bound reinforcement learning

Figure 3
LEFT and RIGHT Tree Illustrations



Note. For the LEFT (RIGHT) tree, nodes on the left (right) at Levels 2 and 3 (surrounded by red and gray rectangles, respectively) have two child nodes, while their siblings have zero child nodes. The panels on the right show the experiment interface for LEFT/RIGHT trees. See the online article for the color version of this figure.

Figure 4
Planning-as-Exploration Model Flowchart



Note. Eq = equation.

algorithm (Gershman, 2018a; Schulz & Gershman, 2019; Srinivas et al., 2010), similar to that used in Monte Carlo tree search (Liu & Tsuruoka, 2016). Adopting a Bayesian approach allows our reinforcement learning model to quantify epistemic uncertainty people hold in their belief about the environment and derive normative ways to update them (Bellemare et al., 2017). This Bayesian reinforcement learning model assumes that nodes are explored according to a priority score:

$$h_s = \hat{v}_s + w_\sigma \sigma_s. \quad (1)$$

That is, the priority, h, for exploring node s is a weighted sum of that node's estimated value, $\hat{v}$, and the uncertainty in that value, σ (derived using

a Bayesian approach, defined below). The balance between value and uncertainty is set by the w_σ parameter, with larger values indicating more uncertainty-directed.[2]

At each step of a rollout, the model selects the next node to visit by noisily maximizing over the priority for the two child nodes, *a* and *b*. We assume the noise is Gaussian with standard deviation 1/β, leading to choice probability:

$$P(s' = a) = \Phi(\beta(h_a - h_b)), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. β can be understood as a determinacy parameter. It is analogous to an inverse softmax temperature[3] used in other reinforcement learning models (Dayan & Daw, 2008) and reflects the propensity of the participant choosing the option with higher priority score. Higher β indicates that the participant's choices are more sensitive to the priority score difference between two options, just as a higher inverse temperature indicates increased sensitivity to the value difference among options.

Value Updates

Having defined the rollout policy, we now turn to how rollouts update the estimated values. The value of a node is the total amount of reward an agent can expect to accumulate over the future nodes, starting from that node. As mentioned above, this Bayesian reinforcement learning model tracks both the mean and variance (uncertainty) in the value estimate of each node. Given the underlying Gaussian distribution of the rewards delivered at each node, we assume that the posterior value estimate of node *s* is a Gaussian distribution $N(\hat{v}_s, \sigma_s^2)$.[4] Larger $\hat{v}_s$ reflects a higher point estimate of the value of node *s*, and larger σ_s^2 reflects a wider distribution—that is, higher uncertainty surrounding the current point estimate. The model updates these estimates during the rollouts, applying a Bayesian Bellman backup at each step. In our current setup, the value of a node is its reward plus the value of its best child (s^*), the one with maximal estimated value. However, if the value of that child node is itself uncertain, this uncertainty must be "backed up" into the value of its parent. In other words, the uncertainty over the accumulated future rewards (i.e., the value of the best child) should be reflected in the uncertainty of the value estimate of the parent. We further assume that this backup may be only partially applied (with a learning rate parameter $\eta \in [0, 1]$, reflecting how much the value estimate of the current node is updated based on new experience obtained during the rollout), and that participants may discount future rewards (with a discount parameter, $\gamma \in [0, 1]$, reflecting how much the value and uncertainty estimate of the child node is propagated to the corresponding statistics of its parent). This yields the following update equations:

$$\Delta\hat{v}_s = \eta(r_s + \gamma\hat{v}_{s^*} - \hat{v}_s), \tag{3}$$

$$\Delta\sigma_s^2 = \eta(\gamma^2\sigma_{s^*}^2 - \sigma_s^2). \tag{4}$$

These updates were applied backward along the entire trajectory at the end of each rollout during the PLAN phase.

Prior Value Sketch

To initialize the value estimates, we assume that participants form a rough "sketch" (gist memory) of the value function based on the 500 ms reward display at the beginning of each round. Because value is composed of immediate reward obtained at the current node and future long-term reward collected across a set of future nodes, we compute the value function sketch in two steps. First, we use a Bayesian approach to derive estimates of the individual rewards for each node by combining a reward prior, $N(\mu_r, \sigma_r^2)$, and a noisy observation of the true rewards, $\mathbf{r}_{\text{obs}} \sim N(\mathbf{r}_{\text{true}}, \sigma_{\text{obs}}^2)$. It is possible that participants may have different levels of memory precision for the reward of different nodes they have seen during the initial brief reward display. The observation noise σ_{obs}^2 is meant to capture the overall imperfect gist memory under

[2] Though the uncertainty-seeking component is incorporated in priority score calculation as a heuristic, previous theoretical work has shown that it can be derived from rational principles (Gittins, 1979; Sezener et al., 2019).

[3] Prior work has suggested that decision noise may itself be an exploration parameter (Fan, Burke, et al., 2023; Fan, Gershman, & Phelps, 2023; Gershman, 2018a, 2019; Lee et al., 2023; Wilson et al., 2014); however, our experiments were not designed to test this hypothesis.

[4] Note that the posterior is only Gaussian if we ignore uncertainty in the policy itself. Specifically, our belief update assumes that the policy deterministically selects the action with maximal estimated value at every state. This is a simplifying assumption; see Tesauro et al. (2012) and Sezener and Dayan (2020) for methods that incorporate policy-related uncertainty into value estimation.

limited time. For simplicity, we take a mean-field approximation of the observation noise, averaging over trial- and node-specific randomness due to, for example, differential attention. This results in a posterior reward estimate $N(\hat{\mathbf{r}}, \sigma^2_{\text{est}})$ with parameters

$$\sigma^2_{\text{est}} = \frac{1}{1/\sigma^2_{\text{obs}} + 1/\sigma^2_r}, \qquad \hat{\mathbf{r}} = \frac{\mathbf{r}_{\text{true}}/\sigma^2_{\text{obs}} + \mu_r/\sigma^2_r}{\sigma^2_{\text{est}}}. \tag{5}$$

In these equations, μ_r and σ^2_r are the true mean and variance of the distribution from which rewards are drawn (0 and 2.25, respectively). We arbitrarily set the observation noise as $\sigma^2_{\text{obs}} = 0.25$.

Given these reward estimates, we then compute a distributional value function based on the successor representation for a random walk policy (Dayan, 1993; Gershman, 2018b). Intuitively, we assume that participants know how likely they are (in general) to visit each node starting from any other node (e.g., they may want to visit the node where they have observed high reward during the initial reward display) at the start of the planning process, but that they do not initially account for how that probability depends on their future actions, which may depend on the rewards.

Formally, the random walk policy induces a transition function, T, defined as:

$$T_{s,s'} = \begin{cases} 1/|S_s| & s' \in S_s \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

where S_s is the set of states accessible from state s (the "children" of s). In our setup (Figure 2), S_s is readily displayed on the screen to the participant throughout the experiment; therefore, they have all the information needed to calculate T. The successor representation is defined as $\mathbf{M} = (\mathbf{I} - \gamma\mathbf{T})^{-1}$. The value function can then be computed as $\hat{\mathbf{v}} = \mathbf{M}\hat{\mathbf{r}}$ (Dayan, 1993), that is, the product of transition matrix and the reward estimates. However, this only provides a point estimate, whereas we are interested in identifying a distribution that captures uncertainty in the true value function. To do this, we begin by adopting a probabilistic interpretation of the successor representation (Carvalho et al., 2024; Eysenbach et al., 2021; Janner et al., 2021), treating $M_{s,s'}$ as the probability of reaching state s' starting from s.[5] We can then approximate the distribution over value as:

$$\tilde{v}_s \sim \sum_{s'} \text{Bernoulli}(M_{s,s'}) \cdot N(\hat{r}_{s'}, \sigma^2_{\text{est}}). \tag{7}$$

Finally, we initialize the value estimates using its mean and variance:

$$\hat{v}_s = E[\tilde{v}_s] = \sum_{s'} M_{s,s'}\hat{r}_{s'}, \tag{8}$$

$$\sigma^2_s = Var[\tilde{v}_s] = \sum_{s'} [M_{s,s'}\sigma^2_{\text{est}} + M_{s,s'}(1 - M_{s,s'})\hat{r}^2_{s'}]. \tag{9}$$

The variance term is of particular interest because it captures the initial uncertainty that we hypothesize to drive participant's early planning. To intuitively understand this expression, observe that the full variance is the sum of two terms. The first term captures uncertainty in the rewards themselves (σ^2_{est}), weighted by the probability that they will be attained ($M_{s,s'}$). The second term captures uncertainty in which of those rewards will be attained ($M_{s,s'}(1 - M_{s,s'})$) weighted by the estimated magnitude of those rewards ($\hat{r}^2_{s'}$). Critically, note that both of these terms will be larger for nodes that have more successors. Thus, a tendency to explore nodes with more successors (including both children and indirect descendants) is consistent with a preference for exploring nodes with higher uncertainty.

Intuitively, participants will have more uncertainty about the value of nodes with more descendants for two reasons (corresponding to the two terms in Equation 9): They mark the beginning of both longer paths and also more possible paths. Longer: all else equal, a node with more descendants will be more steps from the end. Each step comes with a reward of fixed value; however, the participant is uncertain about the value of this reward and this uncertainty compounds over each step of the trial. More: each step will typically involve making a choice,

[5] Note that we can directly interpret the entries of M as probabilities because the task environment is tree-structured, meaning each state can be visited at most once. However, M only represents the marginal probability of each state—it does not capture the dependencies between future states (e.g., you cannot visit two states at one depth). This is the sense in which Equation 7 is approximate.

leading to a "garden of forking paths." Initially, one does not know which way one will go; the uncertainty in one's own future actions translates into uncertainty about which rewards one will actually receive, that is, the value of node.

Model Fitting

We implemented our planning-as-exploration model in Stan (Stan Development Team, 2024) and fit it to participants' choice data from the PLAN phase. Similar to previous work (Ahn et al., 2017; Aylward et al., 2019; Lei & Solway, 2022), we use Stan only to infer distributions over model parameters given experiment data and the model. The Bayesian Bellman backup defined in Equation 4 is implemented analytically as part of the deterministic structure of the model. We fit participant-specific parameters for the learning rate η and the discount factor γ, and both group-level and participant-level parameters for coefficients β and w_σ (see Appendix B for parameter recovery details). For the i'th participant, we set the priors on participant-specific parameters as follows:

$$\eta_i = \varphi(\alpha), \quad \alpha \sim N(0,1), \tag{10}$$

$$\gamma_i = \varphi(\theta), \quad \theta \sim N(0,1), \tag{11}$$

$$\beta_i \sim N(\mu_1, \sigma_1), \tag{12}$$

$$w_{\sigma i} \sim N(\mu_2, \sigma_2). \tag{13}$$

The group-level parameters were given weakly informative hyper priors:

$$\mu_1 \sim N(0,1), \quad \mu_2 \sim N(0,1), \tag{14}$$

$$\sigma_1 \sim N_+(0,1), \quad \sigma_2 \sim N_+(0,1), \tag{15}$$

where $N_+(0, 1)$ denotes the half-normal distribution.

Algorithm 1 Planning-as-Exploration

Initialize value estimate $\hat{v}_s$ and uncertainty σ_s^2 for all $s \in S$
while time $< t_{\max}$ **do**
  $s \leftarrow$ root node
  $\tau \leftarrow$ empty trajectory
  while s has children **do**
    Compute priority score for each child c: $h_c = \hat{v}_c + w_\sigma \cdot \sigma_c$
    Sample next node s' from children of s: $P(s'|s) = \Phi(\beta \cdot (h_{s'} - h_{other}))$
    Append s' to τ
    $s \leftarrow s'$
  end while
  for all $s \in \tau$ (reverse order) **do**
    $s^* \leftarrow$ child of s with highest $\hat{v}_{s^*}$
    $\hat{v}_s \leftarrow \hat{v}_s + \alpha \cdot (r_s + \gamma \cdot \hat{v}_{s^*} - \hat{v}_s)$
    $\sigma_s^2 \leftarrow \sigma_s^2 + \alpha \cdot (\gamma^2 \cdot \sigma_{s^*}^2 - \sigma_s^2)$
  end for
end while

For each model, Markov Chain Monte Carlo was run with four chains. Each chain was run for 2,000 iterations, using the first 1,000 for warmup. The Gelman–Rubin $\hat{R}$ statistic was computed and ensured to be less than 1.1 for all variables. We report the median estimate of posterior distribution and the 95% highest posterior density intervals (HDIs). We treated an effect as statistically credible if the parameter's 95% HDI did not contain 0.

Model-Agnostic Analysis

To supplement this model-based analysis, we also analyzed participants' choices in a more model-agnostic way. Concretely, we modeled participants' choices as a function of the number of child nodes (n_{child}) and the number of indirect descendants (n_{ind}):

$$\begin{aligned} P(s' = a) = \Phi(&w_{\text{child}}(n_{\text{child},a} - n_{\text{child},b}) \\ &+ w_{\text{ind}}(n_{\text{ind},a} - n_{\text{ind},b}) \\ &+ w_{\text{val}}(\bar{v}_a - \bar{v}_b)), \end{aligned} \tag{16}$$

where $\bar{v}_a$ is computed as the average return received from node a prior to the current rollout, serving as an empirical proxy for the posterior mean estimate $\hat{v}_a$. It is also worth noting that since all tree structures are full binary trees (i.e., each node has zero or two child nodes), there exist three levels of relative number of child nodes: −2, 0, and 2. Given the LEFT/RIGHT tree design in Experiment 1, there exist three levels of relative number of indirect descendants as well: −2, 0, 2. In Experiment 2, more flexible tree structures lead

to more possible value for the relative number of indirect descendants, ranging from −10 to 10. Our model simulation results suggest that when the uncertainty weight w_σ is positive, w_{child} and w_{ind} are also positive (Figure 5A). In other words, a positive weight on uncertainty-driven exploration translates into a preference for visiting nodes with more successors in the model-agnostic analysis. Note that this preference for visiting nodes with more future states exists when the estimated value difference is zero ($x = 0$ in Figure 5C and 5D)—that is, when participants have no prior information about the reward distribution or when the value estimates for two nodes are nonzero but equivalent.

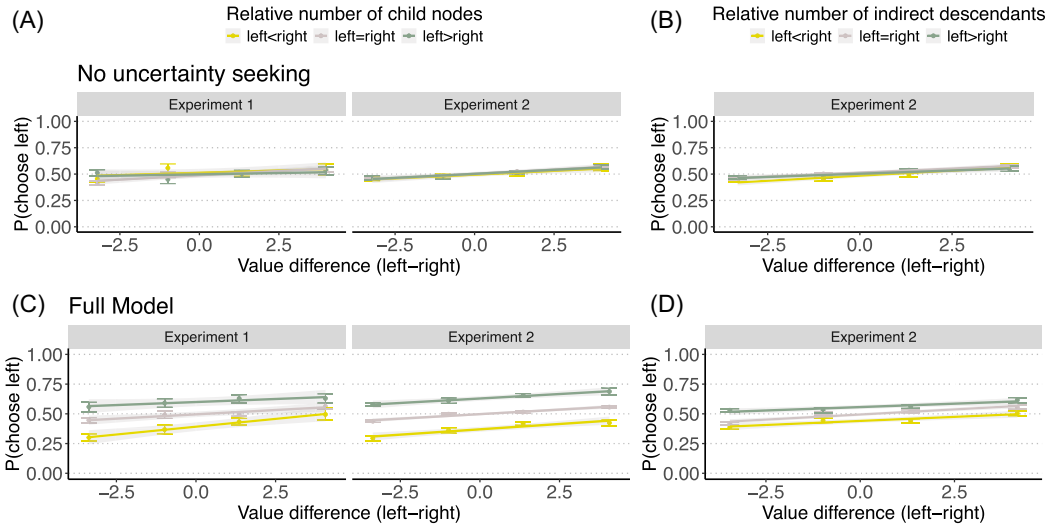
For this model-agnostic regression, we fit a Bayesian generalized mixed-effects model with a probit link function, the same link function used in Equation 2, and included fixed and random effects for all regressors (w_{child}, w_{ind}, and w_{val}). Regressors were standardized before entering the regression. The random effects were drawn from a multivariate Gaussian distribution with mean zero and unknown covariance matrix. We implemented the model with the *brms* R package (Bürkner, 2017), using the default prior for fixed effects (an improper flat prior over the reals). We used the same Markov Chain Monte Carlo sampling hyper-parameters and convergence criteria as for the main model.

Results

Overall, participants performed the task well, choosing the more rewarding node 73.3% of the time during EXECUTE phase. Though not required, people made use of the PLAN phase. On average, people carried out 4.52 rollouts per graph (Experiment 1: 4.87; Experiment 2: 4.51) and 91.2% of the rollouts reached a leaf node (Experiment 1: 90.5%; Experiment 2: 91.4%). People also primarily visited the nodes that they

Figure 5

Simulated Choice Probability Grouped by Number of Child Nodes and Indirect Descendants Using Different Parameter Combinations



Note. All data are simulated using Equations 1–9 with different parameter combinations. In (Panels A and B), the determinacy parameter β is fitted and uncertainty weight w_σ is fixed at 0 (i.e., no uncertainty seeking). In (Panels C and D) both determinacy parameter and uncertainty weight are fitted for the full model. We use fitted parameter values (see Figures 7D and 7E) as generative parameters. Because all tree structures are full binary trees, the yellow (green) line in (Panels A and C) corresponds to the situation where the node on the left (right) has zero child nodes while its sibling has two child nodes. The gray lines correspond to the situation where all nodes on the same level have either zero or two child nodes. For (Panels B and D), we only plot data where both nodes at the current layer has the same number of child nodes—that is, data contributing to the gray lines in (Panels A and C). The lines show the probit regression fit, and the points show binned means. The standard deviation of reward distribution is 1.5. Error bars indicate standard error. Gray ribbons indicate 95% confidence interval of the predicted values. See the online article for the color version of this figure.

inspected during the PLAN phase (Experiment 1: 92.0% trials; Experiment 2: 93.0%). Finally, people who planned more (i.e., carried out more rollouts), earned more rewards during the EXECUTE phase, Experiment 1: $r(37) = 0.66$; Experiment 2: $r(43) = 0.52$, both $p < .001$; Figure 6. These results suggest that the PLAN phase effectively probes planning.

Planning Reflect Both Value- and Uncertainty-Seeking

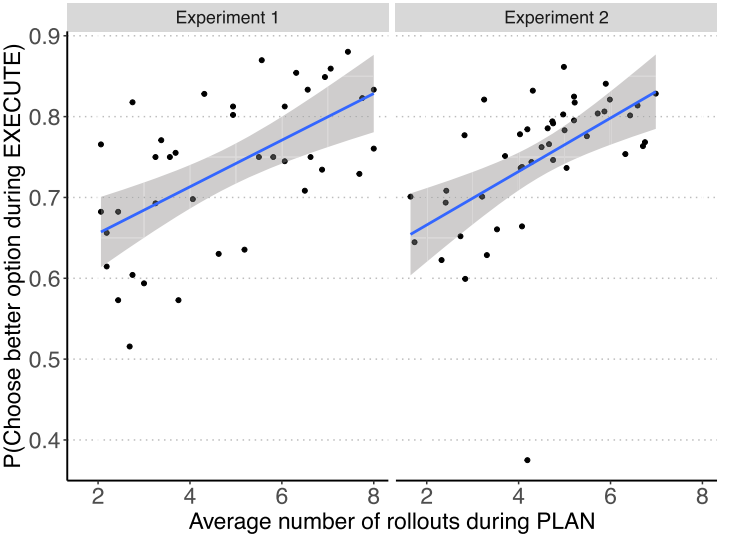
Our model proposes that human planners selectively explore states that have both high estimated value and high uncertainty in value (Equation 1). More concretely, simulations (Figure 5) revealed that the model predicts that people will be more likely to move to nodes from which more reward has been gained on average ($\bar{v}$), as well as nodes that have more children (n_{child}; our experimental manipulation). These two features act as empirical proxies of estimated value $\hat{v}$ and uncertainty σ in Equation 1, respectively. Intuitively, participants will have more uncertainty about the value of nodes with more descendants because there are more rewards one could potentially gain after visiting that node.

As illustrated in Figure 7A, our prediction was confirmed. Applying the Bayesian probit regression described above (Equation 16), we found that participant choices in the PLAN phase were positively related to both value (w_{val}; Experiment 1: $M = 0.16$, 95% HDI [0.11, 0.23]; Experiment 2: $M = 0.19$, 95% HDI [0.16, 0.22]) and number of children (w_{child}; Experiment 1: $M = 0.24$, 95% HDI [0.19, 0.29]; Experiment 2: $M = 0.20$, 95% HDI [0.17, 0.22]). See Figure 7C for posterior estimates of each coefficient.

We additionally found that participants were sensitive to a more farsighted indicator of uncertainty: the number of *indirect descendants*, that is, the number of nodes that could be reached in more than one step. Because the true value (cumulative future reward) of each child node depends on all its descendants, two nodes with the same number of child nodes but different numbers of indirect descendants will have different levels of uncertainty. This preference for nodes with more indirect descendants is also predicted by our model (Figure 5D). Indeed, only considering cases where the two immediately available

Figure 6

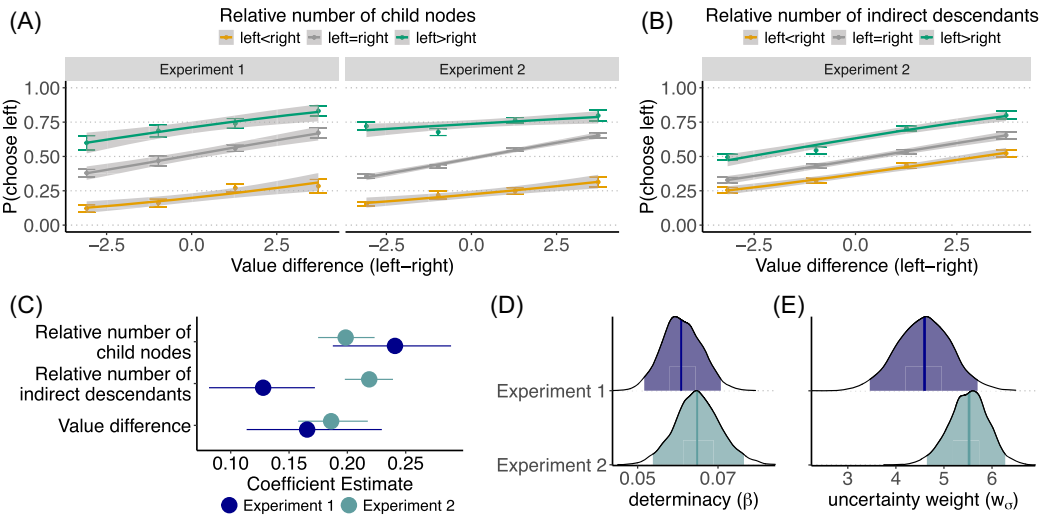
Relationship Between Average Number of Rollouts During the PLAN Phase and Performance During the EXECUTE Phase



Note. Each point represents one participant. Each line shows a linear fit. The shaded area represents the 95% confidence interval. See the online article for the color version of this figure.

Figure 7

Empirical Choice Probability (Panels A and B), Coefficient Estimates From Model-Agnostic Regression (Panel C) and Planning-as-Exploration Model (Panels D and E) During PLAN Phase



Note. Top panel shows empirical choice probability grouped by number of child nodes (Panel A) and indirect descendants (Panel B). For (Panel B), we only plot data where both nodes at the current layer has the same number of child nodes—that is, data contributing to gray lines in (Panel A). The standard deviation of the reward distribution is 1.5. The lines show the probit regression fit, and the points show binned means. Error bars indicate standard error. Gray ribbons indicate 95% confidence interval of the predicted values. Panel C: Shows coefficient estimates for value difference, relative number of child nodes, and relative number of indirect descendants during the PLAN phase for Experiments 1 and 2. Panels D and E: Show posterior estimates of determinacy β and uncertainty weight w_σ in Experiments 1 and 2. Vertical lines indicate posterior median estimate of the group-level effects. Shaded areas indicate 95% credible interval of the posterior coefficient estimates. See the online article for the color version of this figure.

nodes had the same number of children, participants were more likely to explore the node with more indirect descendants (w_{ind}; Experiment 1: $M = 0.13$, 95% HDI [0.08 0.17]; Experiment 2: $M = 0.22$, 95% HDI [0.20, 0.24]; Figure 7B).

Having confirmed our preregistered behavioral predictions, we then fit our theoretical model (Equation 2) directly to participants' planning behavior. This model implicitly captures all the effects described above, as well as more specific belief-updating dynamics that could yield more precise measures of value and uncertainty. The full model outperformed nested models where the coefficient for value or uncertainty was fixed to be 0 (Table 1). In addition, the full 95% credible intervals for the group estimates of β and w_σ were

Table 1

Model Comparison Results for Experiments 1 and 2 Using the LOOIC

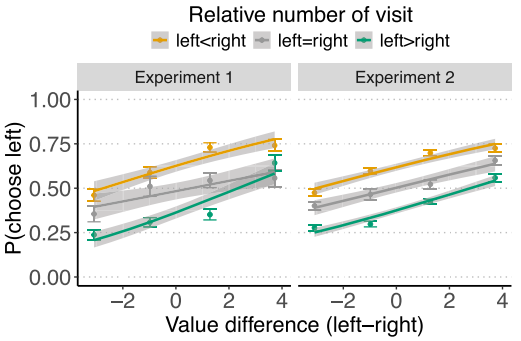
Model specification	Experiment 1	Experiment 2
Model 1 (no uncertainty seeking)	12538.15	56689.44
Model 2 (no value sensitivity)	12124.44	55144.66
Model 3 (full model)	**12085.62**	**54672.62**

Note. Leave-one-out information criterion (LOOIC) is computed as −2 × Expected Log Pointwise Predictive Density. It should be compared for the same experiment across models (i.e., column-wise). Smaller LOOIC indicates better model fit and the smallest LOOIC per column is bolded. For Model 1, the determinacy parameter β is fitted and the uncertainty weight w_σ is fixed at 0. For Model 2, the uncertainty weight is fitted and the coefficient for value in Equation 1 is fixed at 0. For Model 3, both the determinacy parameter and the uncertainty weight are fitted.

above 0 (Experiment 1 β: $M = 0.061$, 95% HDI [0.052, 0.071], w_σ: $M = 4.58$, 95% HDI [3.46, 4.60]; Experiment 2 β: $M = 0.065$, 95% HDI [0.054, 0.077], w_σ: $M = 5.51$, 95% HDI [4.64, 5.52]; Figure 7D and 7E).

The results above suggest that our participants were sensitive to value and uncertainty when choosing which action to simulate next during planning. However, it is possible that these results reflect an inherent preference for states that have more descendants (i.e., "keeping your options open"; Navarro et al., 2018). To address this, we conducted an exploratory analysis using a different measure of uncertainty: The number of times each node had been previously visited while planning. This metric has been applied in other planning algorithms such as upper confidence bounds applied to trees (Kocsis & Szepesvári, 2006) and Alpha Go. Intuitively, the more times that a node have been visited, the more rollouts have been performed, which in turn reduces the uncertainty through updates (Equation 3). As shown in Figure 8, participants were more likely to explore nodes that had been visited less often, controlling for both estimated value and the number of descendants (coefficient for previous visit; Experiment 1: $M = -0.44$, 95% HDI [−0.52, −0.37]; Experiment 2: $M = -0.38$, 95% HDI [−0.44, −0.33]).

Figure 8
Empirical Choice Probability Grouped by Number of Previous Visits



Note. Choice probability was modeled using probit regression. The yellow (green) lines correspond to the situation where the node on the left (right) has been visited more frequently than its sibling before the current choice. The gray lines correspond to the situation where all nodes on the same level have been visited the same number of times before the current choice. The lines show the probit regression fit, and the points show binned means. The standard deviation of reward distribution is 1.5. Error bars indicate standard error. Gray ribbons indicate 95% confidence interval of the predicted values. See the online article for the color version of this figure.

Uncertainty-Seeking Is Selective to Planning

Having confirmed an influence of uncertainty on participants' planning, we next considered a more subtle question. Is the exploratory behavior we observed specific to planning, or does it instead reflect a general influence of uncertainty that is not sensitive to the functional demands of planning versus acting? Intuitively, simulated actions should be more sensitive to uncertainty (and less sensitive to value) because these simulations can directly inform the upcoming choice (and do not incur real consequences). In contrast, real actions should be more sensitive to value (and less sensitive to uncertainty) because they have real consequences (and have a less immediate impact on future choice—in our task, one cannot learn any useful information while acting).

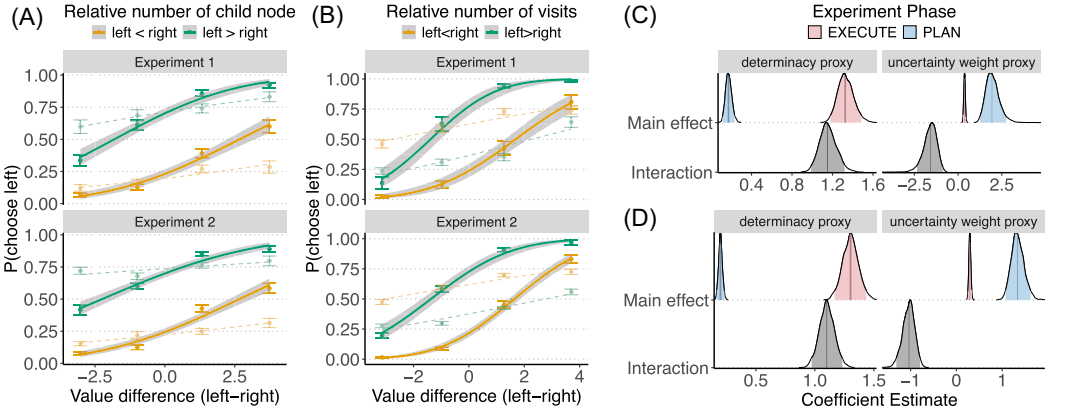
To test this intuitive prediction, we conducted an exploratory analysis comparing behavior in the two phases. As shown in Figure 9A, participants were sensitive to both value and number of children in both phases. However, while uncertainty dominated in the PLAN phase, value dominated in the EXECUTE phase. To statistically confirm this pattern, we regressed choices on relative value and relative number of child nodes, using the phase as an interaction term. Concretely, we constructed a regression analogous to Equations 1 and 2 to obtain parameters comparable to the determinacy β and uncertainty weight w_σ in the model-based analysis:

$$P(s' = a) = \beta^{\text{proxy}}((\bar{v}_a - \bar{v}_b) + w_\sigma^{\text{proxy}}(n_{\text{child},a} - n_{\text{child},b})). \quad (17)$$

We then collapsed the data across the PLAN and EXECUTE phases and include interactions between the proxy parameters and the experiment phase. Consistent with our prediction, we found that people put less weight on uncertainty in the EXECUTE phase (interaction between EXECUTE phase and w_σ^{proxy}; Experiment 1: $M = -1.15$, 95% HDI [−2.35, −0.87]; Experiment 2: $M = -1.04$, 95% HDI [−1.31, −0.80]; Figure 9D). People are also more deterministic during the EXECUTE phase (interaction between EXECUTE phase and β_{proxy}; Experiment 1: $M = 1.10$, 95% HDI

Figure 9

Empirical Choice Probability in the PLAN Phase (Panels A and B), Coefficient Estimates in the PLAN/EXECUTE Phases (Panels C and D)



Note. Empirical choice probability curves are grouped by number of child nodes (Panel A) and number of previous visits (Panel B). Panels C and D show coefficient estimates for determinacy proxy, uncertainty weight proxy, and their interactions. In (Panels A and B), dotted lines and translucent points are data from EXECUTE phase (Figures 7A and 8, left column), replotted here to aid visual comparison. The solid lines show the probit regression fit, and the points show binned means. Error bars indicate standard error of the mean. Gray ribbons indicate 95% confidence interval of the predicted values. In (Panels C and D), interaction terms are calculated using the PLAN phase as the reference level. Vertical lines indicate posterior median estimate of the group-level effects. Shaded areas indicate 95% credible interval of the posterior coefficient estimates. See the online article for the color version of this figure.

[0.97, 1.23]; Experiment 2: $M = 1.15$, 95% HDI [0.99, 1.32]; Figure 9C).

Conducting a similar analysis using the planning-as-exploration model (fit to each stage separately), we found that participants put a higher weight on model-derived uncertainty in the PLAN phase (PLAN–EXECUTE Δw_σ; Experiment 1: $M = 3.39$, 95% HDI [2.21, 4.61]; Experiment 2: $M = 4.35$, 95% HDI [3.48, 5.12]; Figures 9C and 9D) and showed overall more deterministic behavior in the EXECUTE phase (PLAN–EXECUTE $\Delta\beta$; Experiment 1: $M = -0.49$, 95% HDI [−0.57, −0.42]; Experiment 2: $M = -0.52$, 95% HDI [−0.59, −0.45]; Figures 9C and 9D).

The differential effect of uncertainty on decision during PLAN and EXECUTE is also reflected when we quantify uncertainty as the number of times the node had been previously visited during planning. By regressing choices with previous visit, estimated value, number of descendants and interactions between parameters and experiment phase, we find that the effect of previous visits is actually *opposite* in the two phases (Figure 9B): while people tend to visit nodes that they have visited less often before during PLAN phase, they tend to visit the option with more accumulated visits during EXECUTE phase (interaction between EXECUTE phase and coefficient for previous visit; Experiment 1: $M = 0.78$, 95% HDI [0.67, 0.88]; Experiment 2: $M = 0.82$, 95% HDI [0.77, 0.87]).

Discussion

In two experiments, we studied the guidance of planning by uncertainty. The central hypothesis was that mental exploration of decision trees may involve mechanisms similar to overt exploration of environments during reinforcement learning (Schulz & Gershman, 2019). In support of this hypothesis, we reported several converging measures showing that people tend to explore parts of the decision tree with greater uncertainty (i.e., nodes with more children and indirect descendants). Quantitative comparison of computational models confirmed that an uncertainty bonus (in addition to value) improves fit to human data. We also found that this uncertainty seeking was selective to planning, over and above a general tendency to take actions that lead to more future choices.

Our experimental paradigm has two important features. First, the planning phase has a time limit (a maximum of 20 s). The time limit makes it almost impossible to visit all nodes during planning, much like in naturalistic planning tasks. Therefore, this design reduces the possibility of using strategies whose objective is to visit all states (e.g., breath-first or depth-first search). Second, before the start of the planning phase, all the rewards were flashed quickly to the participant. This differs from previous work using the Mouselab-MDP paradigm, where participants need to click the node to reveal the reward for the first time (Callaway, van Opheusden, et al., 2022). Therefore, the act of visiting a node in the current setup is ostensibly not for revealing reward information but rather for reducing uncertainty in the value estimate.[6] In addition, both the rewards and the tree structure varies from trial to trial, especially in Experiment 2 where all tree structures are different. This design prevents people from memorizing fixed action sequences (Huys et al., 2015).

A diverse set of tree structures in Experiment 2 also allows a more nuanced manipulation on uncertainty. First, there exist situations in Experiment 2 where two nodes have different number of indirect descendants yet equal number of child nodes, which is not the case for LEFT and RIGHT tree structure used in Experiment 1. The positive coefficients for both child nodes and indirect descendants in Experiment 2, thus provide strong evidence that people are sensitive to both nearsighted and farsighted indicators of uncertainty. Second, the number of indirect descendants has a wider range in Experiment 2 compared to Experiment 1. As shown in Figure 2, the relative number of indirect descendants of a node has three possible values: 0 (nodes at Levels 3 and 4), 2 (nodes at Level 2 surrounded by red rectangles), and −2, while in Experiment 2, this farsighted uncertainty variable ranges from −10 to 10. The increased variability in the number of indirect descendants could potentially explain the main difference in results between two experiments: though in both experiments people show a clear preference for the node with more child nodes and indirect descendants, participants in Experiment 1 are more sensitive (i.e., larger coefficient estimate) to a change in the relative number of child nodes than to the change in the relative number of direct descendants, whereas the coefficients for child nodes and indirect descendants do not differ in Experiment 2.

Uncertainty-directed actions occurred more frequently during the planning phase compared to the execution phase, consistent with the interpretation that these actions primarily reflect exploration in support of planning. Intriguingly, we still found evidence for (weaker) uncertainty-directed exploration during execution, manifesting as a continued preference for the node with more successors (Figure 9A) and a positive uncertainty weight w_{σ}^{proxy} (Figures 9C and 9D). This raises a normative computational question: Why does uncertainty persist in influencing actions even during the execution phase? Presently, we can only offer speculation. One possibility is that participants continued to engage in planning even after the planning phase; this would indicate that our externalization protocol was not entirely successful. Another possibility is that participants have an intrinsic preference for information, separate from its instrumental role in planning (van Lieshout et al., 2020).

It is important to note that the majority of previous studies, where the planning phase is not externalized, tend to assume that people stick to their plan after its formation. Our finding that people still show preference for nodes with more successors after controlling for value estimates during EXECUTE phase challenges this assumption. In addition, we have shown that, although infrequent, people visit nodes that they have not inspected during the PLAN phase. This is in line with recent work extending uncertainty-driven exploration from a one-step multiarmed bandit problem to a temporally extended decision-making setting (Antonov & Dayan, 2023; Fox et al., 2023).

At first glance, the distinction between PLAN and EXECUTE phases in our study resembles the sampling paradigm that is typically used to study decisions from experience and information sampling, where people gather information about the outcome of available options prior to committing to a single choice (Clark et al., 2006;

[6] Arguably, the brief exposure to the rewards means the rewards are remembered imperfectly by participants, and therefore visiting the nodes during planning does in fact provide information. Future work will be needed to investigate the extent to which this occurs. In addition, note that uncertainty in the value estimate can be reduced even if the immediate reward was remembered perfectly by integrating information about the following rewards.

Gonzalez & Dutt, 2011; Hertwig & Erev, 2009; Hertwig et al., 2004; Hunt et al., 2016; Rakow & Newell, 2010). Despite the seemingly similar separation, the task structure of the sampling paradigm is very different from our experimental paradigm, which presumably leads to different feasible strategies and cognitive processes supporting these strategies. First, the outcomes observed in the sampling paradigm are usually outcomes drawn from a distribution—that is, participants do not have access to the true payoff. In our study, participants directly observe the ground truth reward without noise—that is, the payoff structure is deterministic. Therefore, it is unlikely that participants visit one node for the sake of approximating the reward distribution. Second, the sampling paradigm is mostly applied to study single-step decision making and focuses on the role of external risk, while we examine the algorithms people adopt during planning in a more complex, multistep scenario. Nonetheless, it would be interesting for future research to compare people's weight of uncertainty-driven exploration with people's uncertainty attitude under different decision-making scenarios (e.g., risk and ambiguity attitude) to see if they stem from the same underlying mental construct.

The present study has several limitations. First, the experiment only allows rollout-based planning. We chose this design in part because it allowed us to directly apply exploration strategies from reinforcement learning to model participant's planning. This comes at the cost of not being able to compare our model to search algorithms that are not rollout-based such as best-first search (van Opheusden et al., 2023). Importantly, however, our finding that people use value uncertainty to guide their search could be translated to such models, which typically prioritize based on estimated value alone. However, this restriction may also act as a countermeasure against an artifact of the task. Specifically, visually presenting the full state space makes it easy to jump between different parts of the tree (e.g., by making a saccade). When planning entirely in one's head (as one must typically do), these jumps would likely incur substantial cognitive/computational cost. Indeed, artificial planning agents often use rollout-based strategies when dealing with extremely large state spaces (Barto et al., 1995; Silver et al., 2016). Thus, by restricting participants to this class of strategies in our task, we may be better equipped to understand the way they would plan in a real-world problem. On the other hand, people and animals can employ other planning strategies, including ones that work backwards from desirable states (Afsardeir & Keramati, 2018; Newell & Simon, 1972; Sharp & Eldar, 2024). Understanding the role of uncertainty in these types of algorithms is an important direction for future work.

A second limitation is that, despite the fact that people cannot exhaustively traverse the state space in our task, there is a substantial gap between the complexity of our task and the complexity of real-world planning problems (van Opheusden & Ma, 2019; van Opheusden et al., 2023). More complicated planning scenarios could involve larger state spaces, dependency between states, and nonstationary rewards. These features impose challenges on appropriate uncertainty estimation (e.g., if rewards are interdependent among a set of states, people may update the uncertainty estimate of these states jointly instead of sequentially), which could interact with the usage of uncertainty-driven exploration during planning.

In summary, the present study examined the mental exploration process during planning with a task that allows us to externalize the planning process together with an experiment design that exogenously manipulates the value uncertainty in different states. Our results show that people have a preference for approaching options about which they are more uncertain during planning (after controlling for value differences), suggesting that the mental exploration process, similar to exploration during reinforcement learning, is guided by uncertainty.

References

Afsardeir, A., & Keramati, M. (2018). Behavioural signatures of backward planning in animals. *European Journal of Neuroscience*, *47*(5), 479–487. https://doi.org/10.1111/ejn.13851

Ahn, W.-Y., Haines, N., & Zhang, L. (2017). Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hBayesDM package. *Computational Psychiatry*, *1*, 24–57. https://doi.org/10.1162/cpsy_a_00002

Antonov, G., & Dayan, P. (2023). *Exploring replay*. bioRxiv. https://doi.org/10.1101/2023.01.27.525847

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, *3*, 397–422. https://www.jmlr.org/papers/volume3/auer02a/auer02a.pdf

Aylward, J., Valton, V., Ahn, W.-Y., Bond, R. L., Dayan, P., Roiser, J. P., & Robinson, O. J. (2019). Altered learning under uncertainty in unmedicated mood and anxiety disorders. *Nature Human Behaviour*, *3*(10), 1116–1123. https://doi.org/10.1038/s41562-019-0628-0

Barto, A. G., Bradtke, S. J., & Singh, S. P. (1995). Learning to act using real-time dynamic programming. *Artificial Intelligence*, *72*(1–2), 81–138. https://doi.org/10.1016/0004-3702(94)00011-O

Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. *International Conference on Machine Learning* (pp. 449–458). https://doi.org/10.48550/arXiv.1707.06887

Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics*, *16*(3–4), 221–229.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., & Colton, S. (2012). A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, *4*(1), 1–43. https://doi.org/10.1109/TCIAIG.2012.2186810

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Callaway, F., Jain, Y. R., van Opheusden, B., Das, P., Iwama, G., Gul, S., Krueger, P. M., Becker, F., Griffiths, T. L., & Lieder, F. (2022). Leveraging artificial intelligence to improve people's planning strategies. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(12), Article e2117432119. https://doi.org/10.1073/pnas.2117432119

Callaway, F., Lieder, F., Krueger, P. M., & Griffiths, T. L. (2017). *Mouselab-MDP: A new paradigm for tracing how people plan* [Conference session]. The 3rd Multidisciplinary Conference on Reinforcement Learning and Decision Making. https://doi.org/10.31219/osf.io/7wcya

Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., & Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, *6*(8), 1112–1125. https://doi.org/10.1038/s41562-022-01332-8

Callaway, F., Yu, M., & Mattar, M. G. (2024). Revealing human planning strategies with eye-tracking. *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. *46*, pp. 4343–4349).

Carvalho, W., Tomov, M. S., de Cothi, W., Barry, C., & Gershman, S. J. (2024). Predictive representations: Building blocks of intelligence. *Neural Computation*, *36*(11), 2225–2298. https://doi.org/10.1162/neco_a_01705

Ciosek, K., Vuong, Q., Loftin, R., & Hofmann, K. (2019). Better exploration with optimistic actor critic. *Advances in Neural Information Processing Systems*, *32*. https://doi.org/10.48550/arXiv.1910.12807

Clark, L., Robbins, T. W., Ersche, K. D., & Sahakian, B. J. (2006). Reflection impulsivity in current and former substance users. *Biological Psychiatry*, *60*(5), 515–522. https://doi.org/10.1016/j.biopsych.2005.11.007

Cristín, J., Méndez, V., & Campos, D. (2022). Informational entropy threshold as a physical mechanism for explaining tree-like decision making in humans. *Entropy*, *24*(12), Article 1819. https://doi.org/10.3390/e24121819

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624. https://doi.org/10.1162/neco.1993.5.4.613

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, *8*(4), 429–453. https://doi.org/10.3758/cabn.8.4.429

Dayan, P., & Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Machine Learning*, *25*(1–3), 5–22. https://doi.org/10.1007/BF00115298

De Groot, A. D. (1965). *Thought and choice in chess*. De Gruyter Mouton. https://doi.org/10.1515/9783110800647

Eluchans, M., Maselli, A., Lancia, G. L., & Pezzulo, G. (2025). Eye and hand coarticulation during problem-solving reveals hierarchically organized planning. *Journal of Neurophysiology*, *134*(3), 985–997. https://doi.org/10.1152/jn.00188.2025

Eysenbach, B., Salakhutdinov, R., & Levine, S. (2021). *C-learning: Learning to achieve goals via recursive classification* [Conference session]. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2011.08909

Fan, H., Burke, T., Sambrano, D. C., Dial, E., Phelps, E. A., & Gershman, S. J. (2023). Pupil size encodes uncertainty during exploration. *Journal of Cognitive Neuroscience*, *35*(9), 1508–1520. https://doi.org/10.1162/jocn_a_02025

Fan, H., Callaway, F., & Gershman, S. J. (2025a). *Uncertainty-driven exploration during planning*. Open Science Framework. https://osf.io/t2cmr/?view_only=345b3296b11a4c1d960300b843b6654a

Fan, H., Callaway, F., & Gershman, S. J. (2025b). *Uncertainty-driven exploration during planning—Exp1* [Preregistration]. https://aspredicted.org/83Y_MN8

Fan, H., Callaway, F., & Gershman, S. J. (2025c). *Uncertainty-driven exploration during planning—Exp2* [Preregistration]. https://aspredicted.org/YTG_4K2

Fan, H., Gershman, S. J., & Phelps, E. A. (2023). Trait somatic anxiety is associated with reduced directed exploration and underestimation of uncertainty. *Nature Human Behaviour*, *7*(1), 102–113. https://doi.org/10.1038/s41562-022-01455-y

Fox, L., Dan, O., & Loewenstein, Y. (2023). On the computational principles underlying human exploration. *eLife*, *12*, Article RP90684. https://doi.org/10.7554/eLife.90684.1

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. https://doi.org/10.1038/nn.2342

Gershman, S. J. (2018a). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

Gershman, S. J. (2018b). The successor representation: Its computational logic and neural substrates. *Journal of Neuroscience*, *38*(33), 7193–7200. https://doi.org/10.1523/JNEUROSCI.0151-18.2018

Gershman, S. J. (2019). Uncertainty and exploration. *Decision*, *6*(3), 277–286. https://doi.org/10.1037/dec0000101

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 148–164. https://doi.org/10.1111/j.2517-6161.1979.tb01068.x

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, *118*(4), 523–551. https://doi.org/10.1037/a0024558

He, R., & Lieder, F. (2023). What are the mechanisms underlying metacognitive learning in the context of planning? *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. *45*, pp. 3304–3311). https://doi.org/10.48550/arXiv.2302.04840

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*(8), 534–539. https://doi.org/10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517–523. https://doi.org/10.1002/bdm.598

Hunt, L. T., Daw, N. D., Kaanders, P., MacIver, M., Mugan, U., Procyk, E., Redish, A., Russo, E., Scholl, J., Stachenfeld, K., Wilson, C. R. E., & Kolling, N. (2021). Formalizing planning and information search in naturalistic decision-making. *Nature Neuroscience*, *24*(8), 1051–1064. https://doi.org/10.1038/s41593-021-00866-w

Hunt, L. T., Rutledge, R. B., Malalasekera, W. N., Kennerley, S. W., & Dolan, R. J. (2016). Approach-induced biases in human information sampling. *PLOS Biology*, *14*(11), Article e2000638. https://doi.org/10.1371/journal.pbio.2000638

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., Dayan, P., & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(10), 3098–3103. https://doi.org/10.1073/pnas.1414219112

Jain, Y. R., Callaway, F., Griffiths, T. L., Dayan, P., He, R., Krueger, P. M., & Lieder, F. (2023). A computational process-tracing method for measuring people's planning strategies and how they change over time. *Behavior Research Methods*, *55*(4), 2037–2079. https://doi.org/10.3758/s13428-022-01789-5

Janner, M., Mordatch, I., & Levine, S. (2021, November). *Generative temporal difference learning for infinite-horizon prediction*. arXiv. https://doi.org/10.48550/arXiv.2010.14496

Kadner, F., Willkomm, H., Ibs, I., & Rothkopf, C. (2023). Finding your way out: Planning strategies in human maze-solving behavior. *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. *45*, pp. 1660–1666).

Kocsis, L., & Szepesvári, C. (2006). Bandit based Monte-Carlo planning. *European Conference on Machine Learning* (pp. 282–293). Springer. https://doi.org/10.1007/11871842_29

LaValle, S. M. (2006). *Planning algorithms*. Cambridge University Press. https://doi.org/10.1017/CBO9780511546877

Lee, J. K., Rouault, M., & Wyart, V. (2023). Adaptive tuning of human learning and choice variability to unexpected uncertainty. *Science Advances*, *9*(13), Article eadd0501. https://doi.org/10.1126/sciadv.add0501

Lei, Y., & Solway, A. (2022). Conflict and competition between model-based and model-free control. *PLOS Computational Biology*, *18*(5), Article e1010047. https://doi.org/10.1371/journal.pcbi.1010047

Liu, Y.-C., & Tsuruoka, Y. (2016). Modification of improved upper confidence bounds for regulating exploration in Monte-Carlo tree search. *Theoretical Computer Science*, *644*, 92–105. https://doi.org/10.1016/j.tcs.2016.06.034

Miller, K. J., & Venditto, S. J. C. (2021). Multi-step planning in the brain. *Current Opinion in Behavioral Sciences*, *38*, 29–39. https://doi.org/10.1016/j.cobeha.2020.07.003

Navarro, D. J., Tran, P., & Baz, N. (2018). Aversion to option loss in a restless bandit task. *Computational Brain & Behavior*, *1*, 151–164. https://doi.org/10.1007/s42113-018-0010-8

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Prentice Hall.

Rakow, T., & Newell, B. R. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, *23*(1), 1–14. https://doi.org/10.1002/bdm.681

Sanner, S., Goetschalckx, R., Driessens, K., & Shani, G. (2009). Bayesian real-time dynamic programming. *Proceedings of the 21st International Joint Conference on Artificial Intelligence* (pp. 1784–1789). International Joint Conference on Artificial Intelligence.

Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(28), 13903–13908. https://doi.org/10.1073/pnas.1821028116

Schulz, E., Franklin, N. T., & Gershman, S. J. (2020). Finding structure in multi-armed bandits. *Cognitive Psychology*, *119*, Article 101261. https://doi.org/10.1016/j.cogpsych.2019.101261

Schulz, E., & Gershman, S. J. (2019). The algorithmic architecture of exploration in the human brain. *Current Opinion in Neurobiology*, *55*, 7–14. https://doi.org/10.1016/j.conb.2018.11.003

Sezener, E., & Dayan, P. (2020). Static and dynamic values of computation in MCTS. *Conference on Uncertainty in Artificial Intelligence* (pp. 31–40). Curran. https://doi.org/10.48550/arXiv.2002.04335

Sezener, E., Dezfouli, A., & Keramati, M. (2019). Optimizing the depth and the direction of prospective planning using information values. *PLOS Computational Biology*, *15*(3), Article e1006827. https://doi.org/10.1371/journal.pcbi.1006827

Sharp, P. B., & Eldar, E. (2024). Humans adaptively deploy forward and backward prediction. *Nature Human Behaviour*, *8*(9), 1726–1737. https://doi.org/10.1038/s41562-024-01930-8

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*(2), 351–367. https://doi.org/10.1111/tops.12145

Srinivas, N., Krause, A., Kakade, S., & Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. *Proceedings of the 27th International Conference on Machine Learning* (pp. 1015–1022). https://doi.org/10.48550/arXiv.0912.3995

Stan Development Team. (2024). *RStan: The R interface to Stan* (R package Version 2.32.6) [Computer software]. https://mc-stan.org/

Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In B. Porter & R. Mooney (Eds.), *Machine learning proceedings 1990* (pp. 216–224). Morgan Kaufmann. https://doi.org/10.1016/b978-1-55860-141-3.50030-4

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Tesauro, G., Rajan, V. T., & Segal, R. (2012). *Bayesian inference in Monte-Carlo tree search*. arXiv. https://doi.org/10.48550/arXiv.1203.3519

van Lieshout, L. L., de Lange, F. P., & Cools, R. (2020). Why so curious? Quantifying mechanisms of information seeking. *Current Opinion in Behavioral Sciences*, *35*, 112–117. https://doi.org/10.1016/j.cobeha.2020.08.005

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, *618*(7967), 1000–1005. https://doi.org/10.1038/s41586-023-06124-2

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

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wu, C. M., Schulz, E., & Gershman, S. J. (2021). Inference and search on graph-structured spaces. *Computational Brain & Behavior*, *4*, 125–147. https://doi.org/10.1007/s42113-020-00091-x

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924. https://doi.org/10.1038/s41562-018-0467-4

Zhu, S., Lakshminarasimhan, K. J., Arfaei, N., & Angelaki, D. E. (2022). Eye movements reveal spatiotemporal dynamics of visually-informed planning in navigation. *eLife*, *11*, Article e73097. https://doi.org/10.7554/elife.73097

(*Appendices follow*)

Appendix A

Task Instructions

The instructions for participants in Experiments 1 and 2 are as follows:

In this game, you—the pumpkin candy jar—are going to travel from house to house (colored circles) in a neighborhood and collect candies on your way. Your goal is to collect as many candies as possible.

You are going to use the arrow keys to move between houses via paths, represented as arrows. You can only travel in the direction specified by the arrows. On each path, it is possible to collect or drop candies. A path's goodness does not depend on its location and direction. You will see how many candies you have collect/drop after you walked through the path, indicated by positive and negative numbers, respectively. You will also see how many candies in total you have on the top right corner of the screen.

I am now going to introduce you to ghost mode! During ghost mode, you can visit different paths as a ghost to inspect how many candies you are going to collect/drop on each path, indicated by positive and negative numbers, respectively. Your avatar will be see through. These numbers will remain on the screen during the ghost mode, and the sum of the total ghost candy for the current path is going to show up on top right to help you track the goodness of the path.

In each game, you will start as a ghost. When you are ready to embark your trick-or-treat adventure, press t to enter trick-or-treat mode.

You will start from the initial location to collect treats! Note that if you have already entered the trick-or-treat mode, you can NOT be a ghost again in this round. So use the ghost mode wisely!

During ghost mode, if at any moment you want to return back to the starting point to explore a different path, press SPACE. You can return to the starting point at any moment when you are in ghost mode—that is, you can return to the starting point before you hit the end of the path. In this game, you will have the opportunity to earn up to $X bonus by collecting as many treats as possible. You will receive a bonus proportional to the number of treats you collect in the game. Note: Ghost treats are NOT real treats. Only treats you collect in trick-or-treat mode matter.

Let me introduce two more features of the game before we start the game.

1. Candy map: At the beginning of each round, the whole candy map—that is, how many treats you will collect/drop on each path—will briefly show up on the screen. During this period time, you can NOT move or change mode. After the candy map disappears, you can start exploring the neighborhood as what you have done in previous practices.
2. Time limit: There is a time limit on how long you can be as a ghost. You will have at maximum 20 s. You do not need to use up the 20 s—You can enter trick-or-treat mode at any time within 20 s when you feel ready to start collecting treats. Again, note that as long as you have entered trick-or-treat mode, you can NOT be a ghost any more for the current round. So use the ghost mode wisely! In addition, if you have used up 20 s as a ghost, you cannot be a ghost anymore and will be forced to enter trick-or-treat mode.

Appendix B

Parameter Recovery

We simulated the model (Equations 1–9) and compared the correlation between generated and fitted parameters to test for the parameter recoverability. For generated parameters, learning rate η and discount factor γ are sampled from a uniform distribution $U(0,1)$. Determinacy parameter β and uncertainty weight w_σ are sampled from a Gaussian distribution $N(0,1)$. For each simulation, we

(*Appendices continue*)

randomly select the set of trials (i.e., identical conditions and reward structure) encountered by one participant in either Experiment 1 or Experiment 2 and simulate the choice trajectory using generative parameters. The simulation process was repeated 100 times. The fitted weights highly correlated with their generative counterparts (β: $r = 0.95$; w_σ: $r = 0.93$; both $p < .001$).

Received February 3, 2025
Revision received September 12, 2025
Accepted September 18, 2025 ■