Cognition 267 (2026) 106367

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full Length Article

Model-based planning in structured foraging environments

Thea R. Zalabak *, Laura A. Bustamante, Wouter Kool

Department of Psychological & Brain Sciences, Washington University in St. Louis, 1 Brookings Drive CB 1125, St. Louis, MO 63130, USA

ARTICLE INFO

Keywords:
Foraging
Model-based control
Planning
Decision making
Cognitive control
Exploration

ABSTRACT

In order to maximize reward, humans need to balance engaging with currently available sources of reward and searching for better ones. Optimal foraging theory provides a formal but simple mathematical choice rule to make such stay/leave decisions, contrasting expected and experienced rewards. However, this rule (given by the Marginal Value Theorem; MVT) describes a strategy that does not consider the structure of the environment. In other words, it does not leave room for planning during foraging. Yet, the real world is replete with such opportunities. Therefore, we developed a new structured foraging task to study how people employ goal-directed planning during foraging. Specifically, we explore the extent to which participants incorporate an internal model of the task structure during stay/leave decisions. We find that behavior in this task follows the basic principles of the MVT, but that its structure invites people to also consider the value of alternative reward options when deciding to leave their current one. Importantly, this behavior is pronounced in more goal-directed participants. Computational modeling suggests that incorporating this alternative information is beneficial, but to an extent dictated by choice stochasticity. This study provides a novel method for studying decision making in structured environments, and has implications for understanding how foraging and planning interact.

1. Introduction

When humans make choices, they must consider how they affect both current and future outcomes. This balance between immediate and future rewards is a key feature of *foraging* decisions, during which agents decide whether to engage with (harvest) a current reward option or leave in search of a better one. This is a common problem when humans make decisions about, for example, what or where to eat, maintaining relationships, or searching for a job. A key feature of foraging decisions is that the longer one continues to engage with a current option, the more the utility of its rewards diminish. This makes it necessary to consider when to switch to a new resource. In the real world, people typically consider the structure of the environment when searching for a new option. However, most previous foraging research assumes a relative lack of structure. In this study, we show that people account for such structure during foraging, and that doing so changes their decision making.

Optimal foraging theory (Charnov, 1976; Stephens & Krebs, 1986) has become a popular tool for studying decision making in simple but naturalistic environments. This theory addresses a key subclass of 'patch-leaving' foraging problems, in which agents encounter 'patches' in a given environment at a fixed rate. The optimal policy in this scenario can be implemented using one simple variable: the average experienced reward rate. Specifically, the Marginal Value Theorem (MVT; Charnov, 1976) states that an optimal agent should leave a current option for a new one when the expected reward of their next harvest falls below a threshold defined by the experienced average reward rate. In most experimental patch-leaving scenarios, all patches share statistical regularities, so that a single stable leaving threshold is effective across them. Indeed, previous work has shown that humans and animals use a threshold policy consistent with the MVT in foraging environments (Constantino & Daw, 2015; Hayden, Pearson, & Platt, 2011; Kolling, Behrens, Mars, & Rushworth, 2012; Le Heron et al., 2020).

Even though the MVT provides a parsimonious decision rule for simple decision-making contexts, it does not capture the complexity of many real-world foraging tasks. Most pertinently, the MVT describes optimal foraging behavior in environments in which agents do not have control over what options they encounter when leaving a current option. Instead, it assumes that agents will encounter a randomly selected new option with the same reward regularities as previous patches (i.e., the same patch type).

However, the world is inherently structured, which allows agents to plan where to forage next (K. J. Miller & Venditto, 2021). Rideshare drivers, for example, can maximize their earnings by considering

* Corresponding author at: Washington University in St. Louis, 1 Brookings Drive, CB 1125, St. Louis, MO 63130, USA.
E-mail address: thea@wustl.edu (T.R. Zalabak).

https://doi.org/10.1016/j.cognition.2025.106367
Received 25 April 2025; Received in revised form 21 October 2025; Accepted 28 October 2025

patterns of customer demand across time and location. Rather than accepting rides at any given location, drivers can learn which areas offer higher fares—like airports or stadiums—and weigh whether it's worth relocating. Because travel takes time, they must balance the potential gain from switching locations against the consistent fares that can be obtained in their current area. This raises the question of how people forage in more complex, structured foraging environments, where they can rely on their knowledge of the task structure to decide when to leave, and plan which option to visit next.

An extensive body of research on reinforcement learning (RL) describes how people are motivated to learn about, and take into account, an environment's structure when making decisions. Here, planning (or goal-directed control) is formalized as model-based RL (Daw, Niv, & Dayan, 2005; Drummond & Niv, 2020; Sutton & Barto, 2018). Model-based learners select optimal actions by simulating their consequences in an internal causal model of their environment (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Doll, Duncan, Simon, Shohamy, & Daw, 2015; Karagoz, Reagh, & Kool, 2024; Kool, Cushman, & Gershman, 2016; Kool, Gershman, & Cushman, 2017; Pouncy, Tsividis, & Gershman, 2021). This strategy stands in contrast to model-free RL, which simply chooses actions that previously led to rewards (Thorndike, 1898).

Foraging, like other RL problems, requires agents to learn the value of actions in an environment to decide whether to exploit the current option or explore alternatives (Alejandro & Holroyd, 2024; Morimoto, 2019). The optimal choice rule proposed by the MVT can be thought of as model-free, because the threshold is a simple running tally of experienced rewards and does not require simulation of future consequences (Constantino & Daw, 2015; Harhen & Bornstein, 2023; Kolling & Akam, 2017). It has been suggested that model-based simulations can be used to predict future reward rates from the current option in foraging scenarios (Frankenhuis, Panchanathan, & Barto, 2019; Kolling & Akam, 2017). Such an algorithm would explain the encoding of future reward rate predictions in the anterior cingulate cortex observed by Wittmann et al. (2016). However, it does not address model-based planning towards alternative options in foraging contexts.

The MVT is powerful because it provides an efficient learning rule that achieves optimal foraging behavior. At the same time, the more complicated nature of real-world foraging problems demands a more comprehensive approach to understanding patch-leaving behavior. Rewards may become predictably more or less volatile (Behrens, Woolrich, Walton, & Rushworth, 2007; Stephens & Krebs, 1986), options may stop being available (Navarro, Tran, & Baz, 2018), or environmental conditions can affect the possibility of gaining reward (Stephens, 2008). Here, we approach this set of questions by asking how a foraging task that allows goal-directed planning changes participants' behavioral policy of making stay/leave decisions.

Two recent studies have investigated foraging in structured environments. Hall-McMaster, Dayan, and Schuck (2021) found that foraging decisions change when humans can control which patches they visit. Their task included three distinct patches, each of which had its rewards replenished at a different rate. When given the option to revisit patches, participants sought out the ones that replenished the fastest. This suggests that people relied on information about alternative reward options along with the average reward rate during foraging decisions. More recently, Harhen and Bornstein (2023) demonstrated that people can use model-based state inference during foraging. In their task, participants used reward outcomes to learn the reward structure of a multimodal environment, which allowed them to infer the quality of the current patch. Critically, they found that overharvesting in this task could be explained by the uncertainty that arises in this learning process. Together, these studies demonstrate that people can learn and use a model of their environment to make foraging decisions, and that they are sensitive to explicit representations of alternative rewards outside of the current option.

At the same time, they leave unanswered the question of how the opportunity to plan in structured environments changes foraging behavior. In the study by Harhen and Bornstein (2023), participants used model-based computations to perform latent-cause inference in order to identify the current state. However, their ability to plan was limited because they were not given control over where to travel after leaving a patch. In fact, the transitions in this task were set so that participants typically returned to the same patch type after leaving. In the study by Hall-McMaster et al. (2021), participants directly chose their next option directly, which prevented an assessment of planning behavior.

Here, we build on these studies to explore how model-based planning and foraging interact in structured environments. We developed a novel paradigm that allowed us to simultaneously investigate patch-leaving decisions, planning, and reward learning throughout a limited foraging period. In our task, which combines features of current foraging tasks (Constantino & Daw, 2015; Hall-McMaster et al., 2021) and model-based RL tasks (Daw et al., 2011; Kool et al., 2016), participants navigate a tropical pirate-themed world consisting of three foraging patches with dynamically and independently changing initial rewards. At each of these patches, represented as visually distinct islands with treasure chests, participants decided between earning reward from a depleting treasure chest or leaving for a new island. After participants decided to leave the current island, they chose between two boats that 'traveled' to the other two islands according to a probabilistic transition structure (participants could not return to their current island). Before the task started, participants learned that each boat was connected to the other islands according to a probabilistic transition structure. Specifically, each boat was more likely to travel to one island (through a common transition) than the other (rare transition). This structure allowed us to measure the degree to which participants used model-based planning when making travel decisions.

We found that the task's structure encouraged participants to rely on a strategy more sophisticated than simpler ones as the MVT: they incorporated the rewards available at alternative islands into their patch-leaving decisions. Specifically, participants left their current islands earlier if the expected rewards at the other islands were high, and later if they were low. This behavioral pattern was pronounced for participants who used more model-based control during travel decisions. Thus, a stronger reliance on model-based control amplifies the influence of alternatives on foraging behavior. Formal computational modeling further showed that this way of incorporating alternative rewards yields a benefit in earned reward, with the magnitude dependent on choice stochasticity. Together, these results introduce a novel method for studying how planning, exploration, and motivation interact, and provide new insights into how humans make foraging decisions in structured environments.

2. Methods

2.1. Subjects

We recruited a sample of 150 younger adults (19–34, mean = 28.1 ± 4.2 years, 59 female, 85 male, and 6 non-binary) for this study using Prolific, an online research crowdsourcing platform. Participants were required to be between 18 and 35 years old (so as to prevent influences from age-related decline in goal-directed control; Bolenz, Kool, Reiter, & Eppinger, 2019), fluent in English, and not colorblind. There were 23 participants (15.3 %) who were excluded from analysis for a variety of reasons. We excluded participants who admitted in the post-task survey that they wrote down the task structure that they were asked to memorize (10), participants who missed the response deadline on 15 % or more travel trials (9), participants whose mean reaction time for the travel trial response deadline was at least 2SD below or above the group mean (4), participants who missed the response deadline on 15 % or more harvests (7) and participants whose final score fell below 2SD below the group mean (7). These exclusion criteria served to remove

participants who were not engaging with the task. We reasoned that, if participants generally respond exceptionally fast when choosing between boats, this signifies that they made random (reward-unrelated) choices. Indeed, most of these participants also displayed erratic behavior during the foraging phase (either dramatically under- or over-harvesting).

All participants were compensated at a mean rate of $10.38/h for the task. They were also given a performance bonus of 1 cent for every 30 points earned in the task (mean = $1.38, SD = 0.22, range = $0.59–$1.75).

All participants gave informed consent, and procedures were approved by the Washington University in St. Louis Institutional Review Board. A replication of this study can be found in the *Supplemental Materials.*

2.2. Design

We developed a novel task to investigate the interaction between model-based planning and foraging decisions (using the jsPsych library; de Leeuw, 2015). This task combined features of classic foraging tasks (Bustamante et al., 2023; Constantino & Daw, 2015) with those of tasks that measure the deployment of model-based RL.

In classic patch foraging tasks, participants harvest from 'patches' to gain as many rewards as possible within a given time period. In each patch, the rewards returned from each harvest decrease exponentially over time. Therefore, the participant can, at any time, choose to travel to a new patch with replenished rewards. While each foraging decision forces participants to incur a small time cost for harvesting rewards, choosing to travel to a new patch forces the participant to incur a typically greater travel time cost. This introduces a tradeoff between harvesting rewards from a current patch and incurring the greater time cost of traveling to a new, replenished patch.

The Marginal Value Theorem (Charnov, 1976) formalizes these tradeoffs by dictating that the optimal strategy in classic foraging tasks is to leave a patch when the expected reward rate on the following harvest falls below a threshold based on the average reward rate of the environment. One key assumption of this framework is that the quality of each new patch is drawn from a random distribution. Thus, it applies to contexts that lack the inherent structure that exists in real-world foraging settings, and not to settings in which foragers can identify

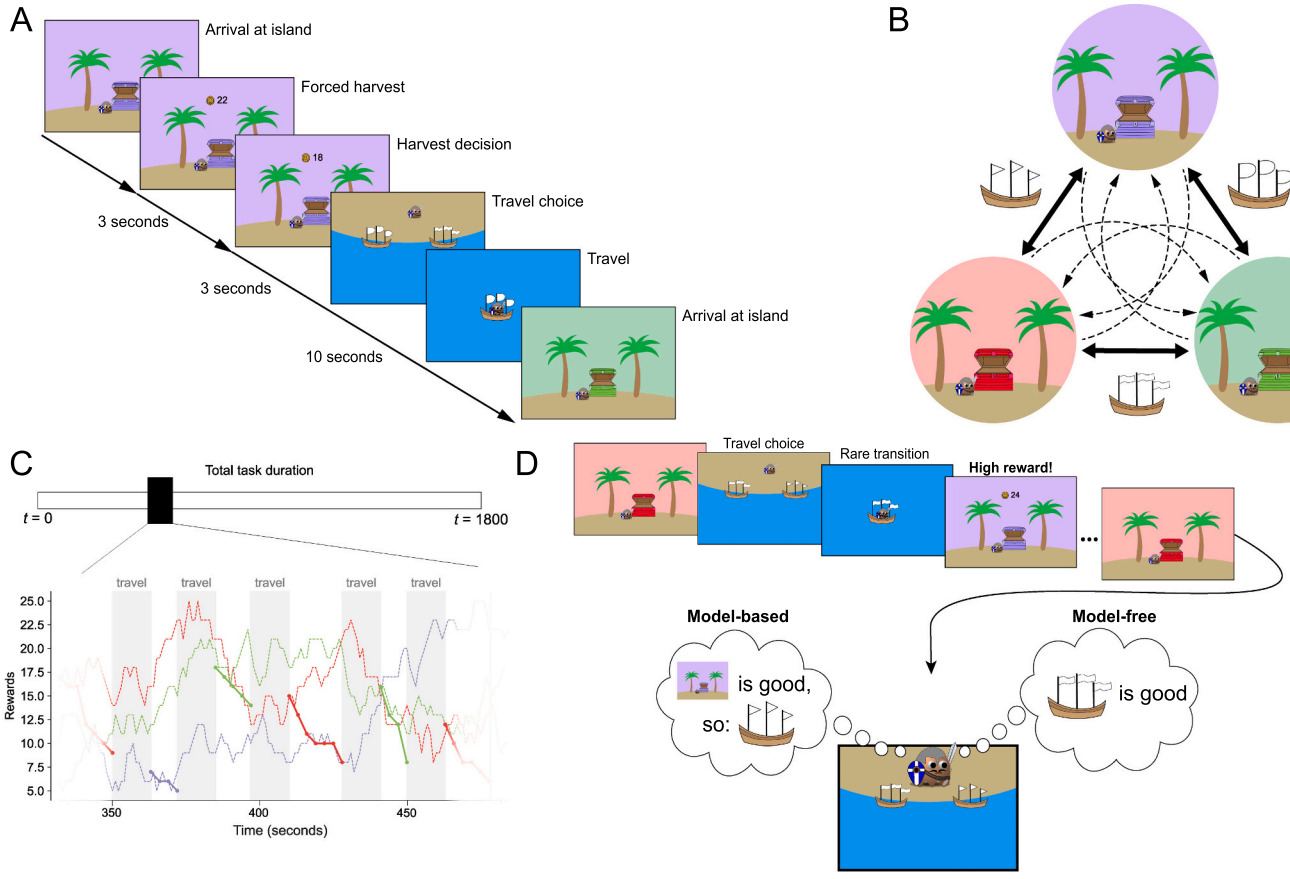

Fig. 1. Foraging task. A. Example task sequence. After participants arrived on an island, they chose between collecting reward and leaving (after an initial forced harvest). If participants chose to earn reward, they received feedback about the number of coins they harvested. Each instance of collecting reward lasted 3 s. Participants could continue to earn rewards until they chose to leave the island. Then, they chose between two boats, which then took them one of the other islands. Each instance of traveling (including the choice between boats) lasted 10s. **B.** Probabilistic transition structure. The task consisted of three distinct islands (purple, red, and green) that were connected by boats. Transitions between islands by boat were probabilistic, so that 80 % of the time, each boat would consistently sail between the same two islands ("common transition"), but the other 20 % of the time, a boat could be diverted to the third island ("rare" transition). **C.** Drifting initial reward dynamics. Three Gaussian random walks (bounds = [5, 25], σ = 1) determined the initial reward for each island at each second of the task (dotted lines). Each successive harvest would return a reward that was decaying from that initial reward (solid lines). Participants did not receive any reward while traveling between patches (grey sections). These example reward dynamics show a small subset of the task's duration (100 s out of 1800s). **D.** Example task sequence demonstrating different predictions of model-based vs. model-free behavior. After a rare transition that leads to a high-rewarding island, a model-based agent becomes less likely to repeat the same boat choice when returning to the original island. This is because their knowledge of the transition structure allows them to infer that selecting the other boat is more likely to lead them back to the previously rewarding island through a common transition. A model-free agent, however, becomes more likely to repeat the same boat choice because it previously yielded a positive prediction error. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and revisit individual patches.

While the MVT is unlikely to provide the optimal decision rule in tasks that allow planning towards future patches, it still highlights two key variables that people use during foraging: the global experienced reward rate and the expected reward of the current patch. We therefore use the general strategy outlined by the MVT as a springboard to investigate how people make decisions in our novel structured foraging task.

Our task (Fig. 1) was partially adapted from the patch-foraging task developed by Constantino and Daw (2015), as well as from aspects of the foraging tasks from Bustamante et al. (2023) and Hall-McMaster et al. (2021). Participants were given 30 min in a tropical pirate-themed foraging environment to collect as many coins as possible from treasure chests on three visually distinct islands. On each of these islands, participants experienced a 'harvesting phase' in which they decided whether to collect diminishing rewards from that island's treasure chest, or to leave. Importantly, once participants decided to leave the current island, they experienced a 'traveling phase' in which they chose between two boats that allowed them to travel to the two other islands. Participants alternated between foraging and travel phases, harvesting rewards and traveling between islands, for the entire task duration.

2.2.1. Foraging phase

At the start of each foraging phase, participants saw an animation of their avatar arriving on an island. Then, they were required to press the 'F' key to earn coins on an initial harvest. After this first forced harvest, participants were given the choice to continue harvesting by pressing the 'F' key again, or to leave the island by pressing the 'J' key.

Each harvest lasted exactly 3 s. This included a 2-s response window during which participants could make their choice (harvest or leave), and the duration of an animation displaying the harvest action and its results (Fig. 1A). If participants failed to respond within the 2-s deadline, they were shown a reminder to respond faster.

At the start of the task duration, the initial rewards for each island were drawn from a Uniform distribution between 5 and 25. From there, these initial rewards changed slowly and independently over time according to a Gaussian random walk (mean = 15, $\sigma = 1$, bounds = [5, 25]) pre-generated for every second of the task. Fig. 1C shows an example of how the initial rewards of the islands drift over time for a small interval (100 s) of the entire task duration (1800s). Because the initial rewards of the islands drifted independently, participants had to track the current value of each island throughout the experiment, in order to identify which one would yield the greatest return at any given time (Daw et al., 2011).

After the first harvest of each foraging phase, the rewards from each following harvest would exponentially decay. Specifically, each harvest would produce a reward equal to the product of the previous harvest reward and a depletion rate. This depletion rate was randomly sampled from a Beta distribution with a mean value of 0.88 ($\alpha = 14.909$, $\beta = 2.033$) on each harvest. Participants were told that each island's treasure chest would always be fully replenished once a participant chose to leave the current island. This way, only the first reward they observed upon arriving at an island needed to be used to determine its current value.

2.2.2. Travel phase

Each travel phase began when participants chose to leave their current island. Then, they saw an animation of their avatar leaving the current island and arriving on a shoreline. There, they were given 3 s to choose from one of the two boats, presented side-by-side, by pressing the 'F' key (for the left boat) or the 'J' key (for the right boat). The boats' positions were randomized on each travel phase. After selecting a boat, participants saw an animation of their avatar moving to the selected boat, leaving the shoreline, sailing across the ocean, and arriving on the destination island (Fig. 1A). Each travel phase lasted exactly 10 s (from the decision to leave to the onset of the stay/leave decision on the new island). If participants failed to select a boat within the 3-s deadline, a boat would automatically be selected at random.

Each pair of islands was connected by a single boat, identified by the shape of its sail (triangle, oval, rectangular). Thus, when arriving at the shoreline, participants would see the two boats that connected the current island to the other two islands. To capture model-based planning, these connections were stochastic (Fig. 1B). Specifically, during each travel phase, there was an 80 % probability that the chosen boat would travel to the connected island (a "common" transition) and a 20 % probability that "rough seas" would divert it to the third island (a "rare" transition). In order to ensure that participants had encoded a correct internal representation of the task structure (Feher da Silva & Hare, 2020), they were extensively trained on traveling between the islands and required to pass five sets of practice trials and comprehension tests that incrementally introduced components of the task before starting the main experiment (see *Procedure*). Note that by "task structure," we refer to knowledge about the existence of three distinct islands in the task and probabilistic transitions that connect them.

2.2.3. Task rationale

We used our task to explore the interaction between model-based planning and foraging decisions in structured environments. To accomplish this, we drew inspiration from previous two-stage RL tasks (Daw et al., 2011; Kool et al., 2016) by employing a probabilistic transition structure, allowing us to measure participants' use of model-based planning during travel decisions. We reasoned that after a rare transition and a high reward on the subsequent island, model-based, but not model-free, participants would be less likely to repeat that same boat choice on the original island (Fig. 1D).

This is because model-based agents use the learned transition structure to associate boats with islands and their expected values, making travel decisions based on which option will most likely bring them to the island with the highest expected reward. Thus, after a rare transition and a high reward on the subsequent island, their knowledge of the transition structure would lead them to realize that the alternative boat option would be more likely to bring them to the previously rewarding island (through a common transition).

Model-free agents, on the other hand, maintain reward expectations for each action based only on reward prediction errors. These agents increase the value of selecting actions (boats) that previously led them to more rewarding islands, regardless of whether these rewards were experienced after a common or rare transition. Thus, after a rare transition and a high reward, model-free agents increase the value of the action that produced it (the chosen boat), leading them to become more likely to choose it in the future.

These diverging predictions follow a similar logic to that of the two-step task, in which choices after rare transitions indicate their degree of planning (Daw et al., 2011). We rely and expand on this rationale in our following analysis of choice behavior.

2.2.4. Procedure

Participants were instructed that they would play the role of an explorer traveling between three distinct tropical islands to collect as much treasure as possible in a limited amount of time (see *Supplemental Materials*). Participants completed five phases of instructions and a set of practice trials before starting the main task. These phases introduced each component of the task in a step-by-step fashion.

First, participants learned how to collect coins from a treasure chest. They were informed about the slowly changing initial rewards on each island, the decaying nature of the treasure chests, and that the chests would become fully replenished after leaving the island. Second, they learned how to leave their current island. Third, participants were introduced to the boats, and practiced arriving on an island's shoreline and selecting a boat. Fourth, they were taught the transition structure, and completed a series of tests to ensure they had fully encoded it. In this phase, participants practiced traveling to each island separately. Then,

they practiced traveling to all three islands in a randomly determined order. To proceed to the next test, participants were required to choose the correct boats 6 times in a row when tested on individual islands, and 10 times in a row when tested on all three islands together. Participants were told that it was important for them to remember how all islands were connected, and were asked not to use any external tools (e.g., a piece of paper) to help them. Finally, participants were introduced to the probabilistic nature of the transition structure, and experienced rare transitions in a set of practice trials. As they were introduced to the rare transitions that were occasionally caused by the 'rough seas', they were encouraged to continue picking boats that would most likely take them to their intended destination island.

Then, the main task started. Participants were truthfully told that they were given 30 min to engage with the task.

After completing the task, participants were asked demographic questions, probed about whether they had used external tools to help them remember the task structure, and debriefed.

2.3. Analyses

We approached our analyses by separately evaluating data from the foraging and travel phases of the task, using RL to model travel decisions and generalized linear models to assess participants' patch-leaving decisions.

2.3.1. Dual-system RL model

In order to capture the degree of model-based planning during the travel phase, we fit a dual-system RL model to behavior (Daw et al., 2011; Doll et al., 2015; Kool et al., 2016, 2017). This model explains behavior as a mixture of a model-free and a model-based RL strategy. These strategies involve using predicted reward values and reward prediction errors, from sampling reward states in the task, to track and plan across each aspect of the task environment.

Each system consists of a function $Q(s, a)$ that associates each combination of state (island) and action (boats/treasure chests) with estimates of expected future reward. These pairs are maintained for both island-boat pairings encountered during the travel phase and for island-chest pairings encountered during the harvest phase (mirroring the two stages in the two-step task; Daw et al., 2011). Specifically, $s_{1,t}$ is the island on which participants made a boat choice (in trial phase 1) on trial t and $s_{2,t}$ is the island they subsequently visited (in trial phase 2). To make the distinction between phases even clearer, we use $Q_{MF_{boat}}(s_{1,t}, a_t)$ to represent the value of choosing boat a_t on island $s_{1,t}$, and $Q_{MF_{chest}}(s_{2,t})$ represents the value of the treasure chest on island $s_{2,t}$. Note that because there is only one treasure chest per island, we omit the action notation for Q-values representing the value of an island-chest combination.

Model-free system. The model-free system updates reward expectations based on reward prediction errors using the SARSA temporal difference learning rule (Sutton & Barto, 2018). Here, the first prediction error δ following a boat choice is calculated as the difference between the expected value given by the agent's boat choice on the original island and the initial reward expected from the treasure chest of the new island:

$$\delta_{1,t} = Q_{MF_{chest}}(s_{2,t}) - Q_{MF_{boat}}(s_{1,t}, a_t). \quad (1)$$

This prediction error is then used to update the value of the boat/island pairing, as

$$Q_{MF_{boat}}(s_{1,t}, a_{1,t}) = Q_{MF_{boat}}(s_{1,t}, a_{1,t}) + \alpha \cdot \delta_{1,t} \quad (2)$$

where α is a learning rate used to scale the prediction error δ following the boat choice. A second reward prediction error occurs after receiving the initial reward on the new island:

$$\delta_{2,t} = r_t - Q_{MF_{chest}}(s_{2,t}) \quad (3)$$

with r_t representing the initial reward and $Q_{MF_{chest}}(s_{2,t})$ the expected value of the treasure chest on the new island. The second prediction error is also used to update the value of the island/chest pairing:

$$Q_{MF_{chest}}(s_{2,t}) = Q_{MF_{chest}}(s_{2,t}) + \alpha \cdot \delta_{2,t} \quad (4)$$

It is also used to update the Q-value for the boat/island pair that was visited/selected at the start of the trial:

$$Q_{MF_{boat}}(s_{1,t}, a_{1,t}) = Q_{MF_{boat}}(s_{1,t}, a_{1,t}) + \lambda \cdot \alpha \cdot \delta_{2,t} \quad (5)$$

where λ represents an eligibility trace decay parameter representing how much a prediction error is used to update previous actions.

Model-based system. The model-based component calculates reward expectations for each boat by combining the probabilistic transition structure with the model-free estimates of the island-chest pairs. Thus, for each available boat, the model-based value is defined in terms of its expected value of the next island:

$$Q_{MB}(s_{1,t}, a_{1,t}) = \sum_{S_2'} P(s_2' | a_{1,t}, s_{1,t})\, Q_{MF_{chest}}(s_2') \quad (6)$$

where $P(s_2' | a_{1,t}, s_{1,t})$ is the probability of transitioning to state s_2' after choosing action $a_{1,t}$ in state $s_{1,t}$. The combination of transition probabilities and reward expectations provides an estimate for model-based reward expectations in the second phase of the trial. This computation approximates prospective simulation, or Monte Carlo Tree Search, and is the standard operationalization of model-based RL in the literature (Daw et al., 2011; Doll et al., 2015; Kool et al., 2016).

It should be noted that the model-based component of this model does not involve any learning of the environment dynamics (such as in Karagoz et al., 2024; Karagoz, Moran, Barch, Kool, & Reagh, 2025). Instead, the agent has encoded the full set of transition probabilities a priori. This reflects our task procedure in which participants are extensively trained on the transition probabilities and environment dynamics prior to the main task.

Choice rule. To connect these values to choices, the Q-values of both systems are mixed according to a model-based weighting parameter w:

$$Q_{net}(s_{1,t}, a_{1,t}) = w \cdot Q_{MB}(s_{1,t}, a_t) + (1 - w) \cdot Q_{MF}(s_{1,t}, a_t) \quad (7)$$

The model-based weighting parameter w is used in RL models to describe the degree of model-based control employed by the participant, where a value closer to 1 represents a high level of model-based control and a value closer to 0 represents a lower level of model-based control, or more model-free behavior. We then use the softmax function to convert these reward expectations to choice probabilities:

$$P(\mathrm{a} | s_{1,t}) = \frac{e^{\beta \cdot Q_{net}(s_{1,t}, \mathrm{a})}}{\sum_{a'} e^{\beta \cdot Q_{net}(s_{1,t}, a')}} \quad (8)$$

where β is the inverse temperature parameter which controls the exploitation-exploration trade-off between two choice options given their difference in value. This parameter can change the function from describing pure exploration (β closer to 0, insensitive to the value of actions) to pure exploitation (higher β value never exploring lower value options).

We used *maximum a-posteriori* (MAP) estimation to find best-fitting values for each of the four free parameters (α, β, λ, w) for each participant separately using the *optim* function in R. To avoid local maxima, we repeated this process ten times with random starting points, selecting the iteration with the highest log-likelihood (Gershman, 2016).

2.3.2. Return intentions

In order to provide a more intuitive description and statistical test of travel choice behavior in this task, we analyzed choice behavior during the travel phase as a function of previous reward history and transition

type. As introduced above (see *Task rationale)*, a model-based agent should change their willingness to return to an island only based on the prediction error they observed there, regardless of whether the prediction error was preceded by a rare or common transition. A model-free agent, however, will increase the likelihood of choosing a particular boat after choosing it results in a positive prediction error, even after a rare transition (Daw et al., 2011; Doll et al., 2015).

To make this more concrete, we frame this analysis in terms of "return intentions." We define a return intention as a boat choice that is most likely to bring the participant back to the island they visited before their current island. Thus, if participants first visited the red island, and then the green island, then upon leaving the green island, choosing the boat that most likely returned to the red island would be taken as a return intention.

We reasoned that model-based agents would be more likely to return to an island where they previously experienced a positive prediction error, regardless of the type of transition they experienced (Fig. 3A). We reasoned that while a model-free agent would be likely to return to an island where they previously experienced a positive prediction error after a common transition, this would not be the case if the positive prediction error occurred after a rare transition. This is because a model-free agent increases the value of the boat that brought them to the rewarding island, leading them to choose this boat again on a future trial instead of selecting the alternative boat that is more likely to bring them back to the rewarding island (see Fig. 3B).

We performed these analyses in two related, but distinct ways. First, we approached this analysis without relying on model-derived regressors, simply marking each trial with an initial reward lower than 15 (the middle point between the reward bounds) as a loss, and those with an initial reward higher than 15 as a win. Second, we followed up on this model-agnostic analysis by using each participant's best-fit learning rate from the dual-system RL model to generate prediction errors for each trial. Because this method uses the full history of reward outcomes, it provides a more subtle label of trial-specific outcomes as either positive or negative. To see this, note that an initial reward of 12 can be experienced as positive (e.g., if one expects 5) or negative (if one expects 20).

2.3.3. Hierarchical mixed effects models

We modeled participants' stay/leave decisions with a series of hierarchical mixed effects logistic regressions (following Hall-McMaster et al., 2021). These models explained participants' stay/leave decisions as a combination of expected reward rate, estimated average reward rate, and available alternative rewards, as well as an interaction between available alternative rewards and participants' degree of model-based behavior.

The goal of this set of analyses was to better understand how participants' knowledge of the task structure impacted their foraging decisions. To do this, we fit three models. The first, baseline, model explained stay/leave decisions as a combination of expected reward rate and average reward rate. The second model additionally included the available rewards at the alternative islands. The third model included all previous regressors, but added the interaction between available alternative rewards and participants' best-fit degree of model-based behavior (see *Dual-system RL model* above). We fit these models with hierarchical logistic regression using the *glmer* function from the *lme4* package in R.

For these models, the average reward rate was calculated for each participant separately, using a threshold updating rule outlined in Constantino and Daw (2015). After every trial, we calculated a prediction error:

$$\Delta = \frac{r_t}{\tau_t} - \rho \tag{9}$$

where r_t is the reward gained from the previous action, ρ is the average reward rate, and τ_t is the length of the previous action. This was then used to update the average reward rate as:

$$\rho = \rho + (1 - (1 - \alpha_\rho)^{\tau_t}) \bullet \Delta \tag{10}$$

where α_ρ represents each participant's learning rate. We also calculated an estimate of participants' expected reward rate for each harvest decision:

$$ER = \frac{r_{t-1} \cdot \kappa}{\tau_h} \tag{11}$$

where κ is the average depletion rate of a patch after each harvest, set to the mean value of 0.88 of the Beta distribution, and τ_h is the time cost of a harvest (3 s).

For each of the three models, we tested three different methods of calculating the available alternative rewards using the initial richness values from participants' most recent experience on each island. One of these represented the alternative rewards as the mean of the two values for the other islands, the second represented them as the maximum of these values, and the final one represented them as a combination weighted by the transition probabilities (assuming people would prefer to visit the most rewarding state).

Because this approach requires fitting individual learning rates to generate the reward rate predictors, we first fit these for each participant separately using individual logistic regressions (without the model-based control parameter). Specifically, for each participant, we used the *nloptr* function in R to find the learning rate and a starting value for the average reward rate that maximized the loglikelihood of the regression model. To avoid local maxima, we repeated this process for 35 iterations with random starting points and extracted the solution with the highest loglikelihood. We did this for each of the three versions of alternative reward (mean, max, weighted), resulting in three sets of reward-related regressors for each participant.

Next, we used these to run the hierarchical models described above. We then compared the AIC values from each model to determine which best explained performance.

3. Results

Participants performed a foraging task that was embedded in a task structure that allowed planning towards goals. The task alternated between a foraging phase, in which participants decided whether to continue harvesting a depleting patch (treasure chest) from their current island, and a travel phase, in which they chose between boats that took them to another island. Here, we will first analyze choice behavior from the travel phase to demonstrate that people used model-based control to plan towards island selection.

Second, we will analyze choice behavior from the foraging phase to determine how knowledge about task structure affects foraging behavior. While our task departs in fundamental ways from the context in which the MVT was derived, we use this influential rule as a mathematical foundation to show that foragers continue to be influenced by both average reward rate and expected reward. Then, we adopt this rule to demonstrate that the task's structure invites a systematically different strategy, with the extent of this difference depending on their degree of model-based control. Specifically, we show that people's inclination to plan (their use of model-based control) during the travel phase predicts how much they let alternative reward information (in addition to average reward rate) influence their stay/leave decisions during the foraging phase. In other words, participants' use of model-based control strengthened the influence of alternative rewards on their decision about when to leave their current island.

3.1. Travel decisions

3.1.1. Model-agnostic return intentions

We first analyzed choice probabilities as a function of the reward earned and the transition type experienced on the previous trial. We do

this in terms of "return intentions", the tendency to revisit an island on which the participants previously experienced a rewarding outcome.

The rationale for this analysis is that after a boat choice and a common transition, positive outcomes should influence model-free and model-based agents in the same way when they encounter the same boat again. Following common transitions, both types of agents become more likely to repeat their previous choice after a positive outcome (registering as an intent to return) and less likely to repeat their previous choice after a negative outcome. For the case of a rare transition, however, positive outcomes drive model-free agents to repeat their choice, and model-based agents to switch (a return intention).

To visualize these behavioral patterns, we plot return intentions as a function of previous outcome, which could be positive (more than 15 initial points) or negative (less than 15 initial points) and previous transition type (common vs. rare). Fig. 2 shows the results of this analysis. We found that participants are more likely to return to an island after a positive compared to a negative outcome, but only after common transitions ($t = 4.69, p < 0.001$, Cohen's $d = 0.56$) and not after rare transitions ($t = 0.08$, $p = 0.93$, Cohen's $d = 0.01$). The effect of previous reward on intention to return was significantly different between transition types ($t = 3.95$, $p < 0.001$, Cohen's $d = 0.35$).

These results indicate that participants were not entirely model-based when making travel choices. Instead, the pattern of participants' return intentions reflects a mixture of model-free and model-based contributions, for which the effect of previous outcome is robust following common transitions, but not following rare transitions.

3.1.2. Dual-system RL model

Next, we used a dual-system RL model to estimate the degree to which people used model-based control to plan towards islands. As mentioned above, this is possible because this task dissociates model-free from model-based control by exploiting low-probability pairings between behavior and reward. Model-free agents are sensitive to such reward, because they rely on the direct experience of action-reward pairings. Model-based agents, however, discount these experiences using an explicit causal model of the task structure (see *Dual-system RL model* and *Task rationale*).

The model-based weighting parameter w in our dual-system RL model reflects the degree to which participants used the task's transition structure to plan towards an island. As is often observed in the two-stage RL tasks that inspired the current design, participants' behavior reflected a mixture of model-based and model-free strategies as indicated by a mean w of 0.52 (SD = 0.18). Table 1 reports the descriptive statistics for all estimated parameters. The full distribution of each fitted parameter can be found in the Supplemental Materials.

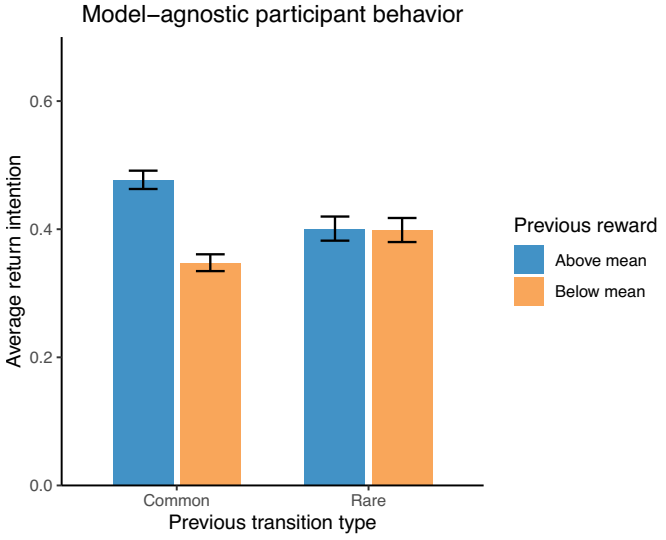

Fig. 2. Average return intentions by transition type and previous reward outcome for participant behavior. Previous reward outcome is determined using a model-agnostic method, in which a positive outcome is defined as an initial reward of more than 15 coins, and a negative outcome is defined as an initial reward of less than 15 points. Error bars indicate within-subject standard error of the mean.

3.1.3. Return intentions with model-derived PEs

Then, we analyzed participants' return intentions using the model-derived PEs calculated using participants' individual parameter fits. As detailed in the *Methods*, this analysis provides a more sensitive measure of the subjective experience of rewards in terms of negative and positive prediction errors.

First, for validation, we simulated behavior for fully model-free ($w = 0$) and fully model-based ($w = 1$) agents (Figure 3AB). These patterns were consistent with our hypotheses about behavior in this task.

Next, mirroring the model-agnostic analyses above, we found that participants were more likely to return to an island after receiving a positive reward prediction error compared to a negative reward prediction error, but only after common transitions ($t = 4.69$, $p < 0.001$, Cohen's $d = 0.42$) and not after rare transitions ($t = -0.11$, $p = 0.91$, $d = 0.01$). Importantly, the effect of previous reward on the intention to return was significantly different between transition types ($t = 2.55$, $p = 0.012$, $d = 0.23$). These results (Fig. 3C) provide convergent evidence that participants' behavior reflected a mixture of model-free and model-based influences.

3.2. Stay/leave decisions

Having confirmed that people use a mixture of model-free and model-based control in the travel phase of our task, we next investigated whether their propensity towards these systems influenced their foraging decisions.

Based on the foundation of the MVT, we predicted that participants would be less likely to stay on their current island if (a) the average reward rate was higher and (b) the expected reward for the next harvest decision was lower. Based on prior work (Hall-McMaster et al., 2021; Harhen & Bornstein, 2023), we also predicted that participants would be less likely to stay with increasing 'alternative' rewards. That is, if the rewards on the other two islands are higher, participants may consider leaving their current island sooner. Finally, we investigated whether this tendency was modulated by the degree of participants' model-basedness, reasoning that goal-directed participants may be more inclined to let task structure guide their foraging decisions.

Before modeling participants' data in a hierarchical regression, we fit individual learning rates and starting values for the average reward rate (Hall-McMaster et al., 2021). We did this to reduce the complexity of our full model and for computational tractability. These models explained stay/leave decisions as a function of expected reward, average reward rate, and available alternative rewards.

Next, we used these values to calculate individual regressors representing average reward rates and alternative reward values for each stay/leave decision. We used these to run three different, increasingly complex, hierarchical mixed effects models. The baseline model used only the average reward rate across all islands and the expected reward rate of the next harvest to explain foraging decisions. Consistent with the foundational principles of the MVT, we found that people were more likely to continue harvesting rewards on an island with an increasing expected reward rate for the next forage ($\beta = 5.62$, SE = 0.33, $p < 2e\text{-}16$) and when the average reward rate decreased ($\beta = -2.11$, SE = 0.296, $p = 1.06e\text{-}12$).

We then investigated whether the structure of the task led participants to implement strategies that are not fully captured by the MVT. As a first step, we added the alternative rewards as a regressor to model how they influenced stay/leave decisions. Here, we again saw the main effects of expected reward ($\beta = 5.70$, SE = 0.34, $p < 2e\text{-}16$) and average

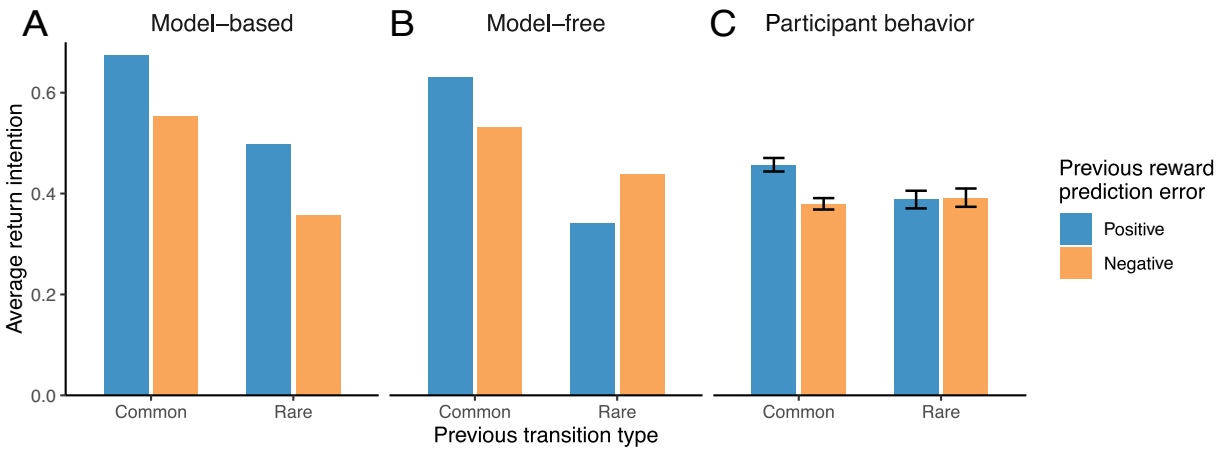

Fig. 3. Average return intentions by transition type and previous reward prediction error. The leftmost plot (A) displays simulated behavior for a fully model-based agent, and the center plot (B) displays simulated behavior for a fully model-free agent. Participant behavior is shown in the rightmost plot (C). Prediction errors were calculated using model-derived learning rates fit for each participant. Error bars indicate within-subject standard error of the mean.

Table 1
Travel choice RL model parameter fits. A dual-system RL model was individually fit to each participant's travel choices. The model included four parameters (learning rate, inverse temperature, eligibility trace decay, and model-based behavior weighting parameter), each of which was fit for each participant. Mean and SD of all participants' parameter fits are shown here.

Parameter	Mean	SD
learning rate: α	0.46	0.23
inverse temperature: β	0.20	0.07
eligibility trace decay: λ	0.48	0.11
model-based behavior: w	0.52	0.18

reward rate ($\beta = -2.05$, SE = 0.30, p = 1.6e-11). Most importantly, we also found a main effect of average[1] available alternative rewards ($\beta = -0.15$, SE = 0.03, p = 5.78e-8), indicating that participants left patches later when alternative rewards were low, and earlier when they were high. This replicates the finding of Hall-McMaster et al. (2021) in a task with a different structure.

This finding allowed us to test whether participants with a stronger reliance on alternative rewards during stay/leave decisions also used a more model-based strategy during travel decisions. Therefore, our final model added the participants' degree of model-based behavior, as measured by the RL weighting parameter w (see *Dual-system RL model*), and its interaction with alternative rewards, as regressors. As before, we again found main effects of expected reward ($\beta = 5.71$, SE = 0.34, $p <$ 2e-16), average reward rate ($\beta = -2.05$, SE = 0.30, p = 8.94e-12), and mean available alternative rewards ($\beta = -0.15$, SE = 0.03, p = 2.66e-8). However, this latter effect was qualified by an interaction effect with model-basedness ($\beta = -0.08$, SE = 0.03, p = 0.002). These findings were replicated in our second set of participants (*see Supplemental Materials)*.

We compared these three models using AICs (see Table 2), and found that the third model, that modeled stay/leave decisions as a function of (i) expected reward rate, (ii) average reward rate, and (iii) an interaction between the mean of available alternative rewards and participants' degree of model-based behavior was the best fitting model for our participants' data.

[1] As outlined in the *Methods* section, we tested three different methods of representing alternative rewards; the mean of the two alternative island reward values, the maximum of the two values, and a combination weighted by the transition probabilities. Here, we report the results of the model that uses the mean of the alternative reward values, as the results of these models do not qualitatively change between the three different methods of representing alternative rewards. We interpret this in the General Discussion.

Table 2
Hierarchical mixed-effects model comparison. Column 1: Predictors; Lists each regressor used in the models to predict stay/leave decisions. Columns 2–3: Baseline (MVT); Lists beta coefficients and p-values for each regressor in our baseline model, which uses expected reward rate of the next harvest and average reward rate across all islands to explain foraging decisions. Columns 4–5: Baseline + Alt rewards; Lists beta coefficients and p-values for each regressor in our second model, which predicts foraging decisions using a combination of expected reward rate, average reward rate, and average available alternative rewards. Columns 6–7: Baseline + Alt rewards*w; Lists beta coefficients and p-values for each regressor in our final model, which predicts foraging decisions using a combination of expected reward rate, average reward rate, average available alternative rewards, and an interaction between alternative rewards and w (degree of model-based behavior). The AIC scores for each model are denoted in the final row of the table.

	Baseline (MVT)		Baseline + Alt rewards		Baseline + Alt rewards*w	
Predictors	*Statistic*	*p*	*Statistic*	*p*	*Statistic*	*p*
(Intercept)	9.90	**<0.001**	9.85	**<0.001**	10.01	**<0.001**
Expected reward rate	16.87	**<0.001**	16.61	**<0.001**	16.71	**<0.001**
Average reward rate	−7.12	**<0.001**	−6.74	**<0.001**	−6.82	**<0.001**
Alternative rewards			−5.43	**<0.001**	−5.56	**<0.001**
w					−1.18	0.237
Alternative rewards × w					−3.06	**0.002**
N	127_{subid}		127_{subid}		127_{subid}	
Observations	47,273		47,273		47,273	
Marginal R^2 / Conditional R^2	0.368 / 0.959		0.367 / 0.960		0.368 / 0.960	
AIC	24,842.138		24,738.245		24,731.356	

These findings indicate that participants who used the task structure to make travel decisions also used it to make foraging decisions. Specifically, more model-based participants used the task structure more strongly when deciding whether to continue foraging. This result is striking, because it involves a cross-phase link between the degree of model-based planning from the travel phase and stay/leave decisions in the foraging phase here. This independence strongly suggests that goal-directed control played a key role during foraging decisions. It indicates that increased reliance on task structure enabled participants to more robustly retrieve the value of the alternative islands, incorporating this into the decision rule.

To visualize this finding, Fig. 4 depicts the probability of staying on

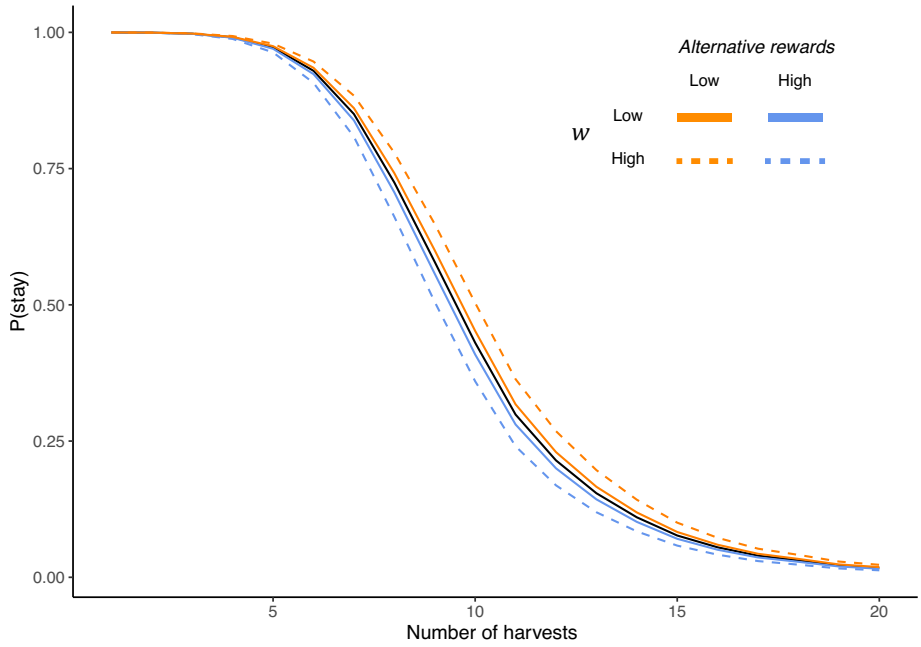

Fig. 4. Participants' stay probability as a function of time spent on each island, whether alternative rewards are low (10th percentile) or high (90th percentile) and whether participants' degree of model-based control is low or higher (+/− 1SD). Higher (lower) alternative rewards led more (less) model-based participants to leave their current island sooner than less (more) model-based participants. Alternative rewards are centered subject-wise.

the current island as a function of the time spent in the patch, the amount of alternative rewards, and participants' level of model-based control. To generate this graph, we first computed average levels of expected and average reward across a range of numbers of forage decisions. Next, these values were combined with the group-level regression coefficients and the logistic function to compute the predicted stay probabilities, assuming average alternative rewards and model-based control. Then, we calculated similar curves, but now for trials with low and high levels of alternative rewards (10th and 90th percentile across all data) and participants with low and high levels of model-based control (+/− 1SD).

This graph demonstrates that participants become less likely to stay with each foraging decision, because the expected reward continues to decline (black line). More importantly, the graph depicts the effects of the size of the alternative rewards and model-based control on stay/leave decisions. We see that participants leave islands faster when their expected alternative rewards are high (blue lines) compared to when they are low (orange lines). Most importantly, the degree of model-based control exaggerates this effect of alternative rewards. That is, when alternative rewards are high, model-based participants (dashed blue lines) leave their islands faster than model-free participants (solid blue lines). Likewise, when alternative rewards are low, model-based participants stay longer than model-free participants. In other words, model-based control does not predict simply how soon people leave their patch. Instead, it dictates how sensitive participants are to alternative rewards.

The results so far invite two different interpretations. It is possible that both relatively model-based and relatively model-free participants incorporate alternative rewards into their choice rule, but to different extents. Alternatively, the model-free participants may simply not consider the alternative rewards. Both scenarios would yield the interaction effect that we observed, but with different mechanistic interpretations.

To test these two hypotheses, we performed model comparison separately for a group of relatively model-based participants ($ws > 0.5$) and for a group of relatively model-free participants ($ws < 0.5$). For both groups, we compared the individual fits of the simplest model (with only expected and average reward rate) with those of the model that also incorporated the alternative rewards. Interestingly, we found that, at the group level, the more complicated model fit better for the model-based participants ($\Delta AIC = -101$), whereas the simpler model fit better for the model-free participants ($\Delta AIC = 24$). This suggests that model-free participants, in contrast to model-based participants, implement a foraging policy that does not consider the structure of the environment.

4. Simulations

Our findings indicate that knowledge about task structure not only influences decisions about where to travel next but also about whether to continue harvesting a current patch. Thus, structured foraging environments lead people to use more elaborate decision rules than those described by the MVT. As we have noted before, this is not unexpected, since our task violates a core assumption of this framework: agents have control over which patches they visit. However, it remains unclear whether the decision to let alternative rewards influence foraging decisions improves task performance.

We investigated this using a comprehensive generative model of behavior on the task, combining our existing models to simulate harvesting behavior using average reward rates and travel decisions using dual-system RL.

Based on our behavioral results, this model differed from standard computational models of foraging in two ways. First, we allowed stay/leave decisions to be influenced by not just the average reward rate across all choices, but also by the estimates of the reward rates accrued at individual islands. Thus, after each reward, the model updates the average reward rate as described above, but it also updates an island-specific reward rate on current island s_t:

$$\rho_{s_t} = \rho_{s_t} + (1 - (1 - \alpha_\rho)^{\tau_t}) \bullet \Delta \quad (12)$$

where ρ_s is the island-specific reward rate.

Based on our results, we investigated how task performance changes as a function of both model-basedness (w) and the extent to which alternative reward information is incorporated into stay/leave decisions. The latter was implemented by an additional weighting parameter, z, that governed the extent to which the agent relies on their

knowledge of available alternative rewards in making stay/leave decisions. Specifically, the agent calculates a "mixed" reward rate as:

$$\rho_{mixed} = (1 - z)\cdot\rho + z\cdot(\rho_{alternative}) \quad (13)$$

where $\rho_{alternative}$ is the average reward rate experienced at the other islands. They then use this to decide whether to stay in the current patch by comparing the expected rewards against this mixed reward rate using a logistic decision rule:

$$p_{stay} = \frac{1}{\left(1 + exp\left(-\beta_s \cdot \left(\frac{r_{t-1}\cdot\kappa}{\tau_h} - \rho_{mixed}\right)\right)\right)} \quad (14)$$

with an inverse temperature β_s specific to stay/leave decisions. Note that for values of z close to 0, the agent implements a similar strategy to the MVT, only comparing the expected reward to the average reward rate. For values of z closer to 1, however, the agent strongly relies on the rewards experienced on the islands that it is currently not visiting.

In other words, this parameter determines the degree to which the agent uses prospective simulation, calculating the value of the alternative (potentially upcoming) reward options, to make stay/leave decisions. This lets us model behavior we observed in the foraging phase of the study reported above.

During the travel phase, the agent is modeled using the dual-system RL model to make travel decisions. Here, the agent computes the model-based estimates of the available boats' action values using a one-step rollout or Monte Carlo Tree Search.

We used simulations of this model to better understand how average reward rate in the task changed as a function of the w and z parameters. Additionally, we varied the inverse temperature parameter, β_s, which governs choice stochasticity or the tradeoff between exploration and exploitation (Addicott, Pearson, Sweitzer, Barack, & Platt, 2017; Cohen, McClure, & Yu, 2007), during stay/leave decisions. The inverse temperature parameter for travel decisions and the learning rates for this model were fixed to values consistent with our prior modeling ($\beta = 0.36$, $\alpha_{RL} = 0.5, \alpha_\rho = 0.1$).

In short, we discovered that agents can increase their average reward by relying on information about the task structure during stay/leave decisions in the foraging phase. First, following prior work (Kool et al., 2016), we found that model-based agents gain a small increase in reward

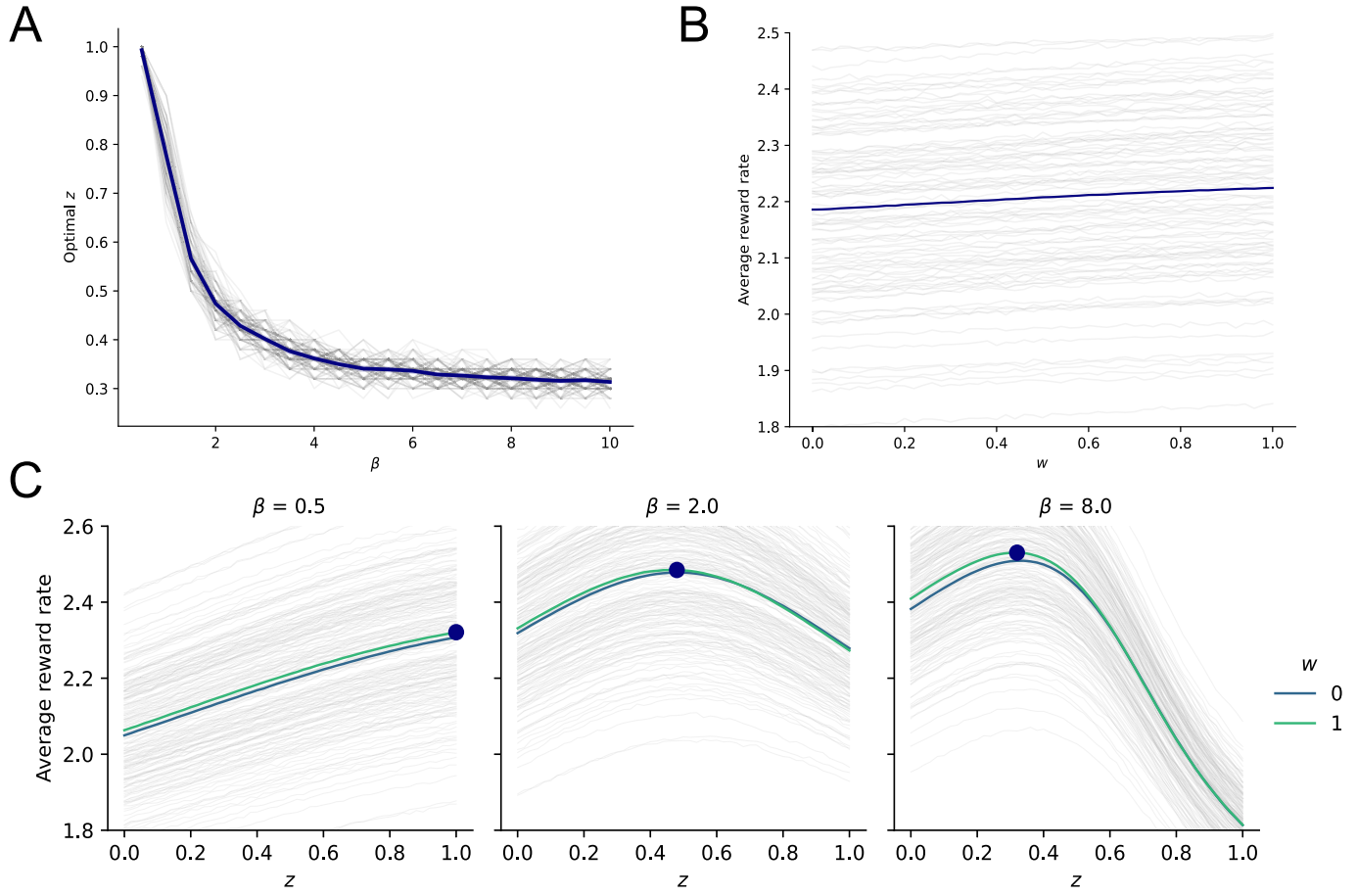

Fig. 5. Simulations. A. The relationship between β and reward-maximizing z. The simulations used to generate these values used a w value of 0.5, but the relationship between β and z remains qualitatively consistent across w values. Simulated data was generated for 1020 combinations of possible β and z values. We performed 1000 simulations of each combination of β and z for 100 different sets of generated rewards. Each point on the main (dark blue) line of this plot indicates the z value that led to the highest average reward rate for each β value (indicated by the x-axis). Each faint grey line shows the variability of reward-maximizing z within one of the 100 sets of randomly generated rewards. **B.** The relationship between w and reward rate. The simulations used to generate these values used a z value of 0.5 and a β of 0.5, though the average reward rate of a fully model-based agent ($w = 1$) is always somewhat greater than that of a fully model-free agent ($w = 0$) regardless of β. Each point represents the average of 1000 simulations of the task using that w value (indicated by the x-axis) for 100 different sets of generated rewards. The faint grey lines indicate the variability in the change of reward rate in one of the 100 sets of generated rewards—in each set of simulations, the fully model-based agents always achieve a slightly higher reward rate than fully model-free agents. **C.** Relationship between z, reward rate, and β. In each of the three plots, each point represents the average of 1000 simulations using that z value (indicated by the x-axis) and β (indicated on the plot) for 100 different sets of generated rewards. This relationship is plotted one for each of the 100 different sets of generated rewards (faint grey lines), as well as their average (thicker lines). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

compared to more model-free agents. This effect is not surprising because model-based agents make more accurate decisions about which islands to visit. This way, they can choose islands that carry higher initial reward values.

The difference in average reward rate between more model-free agents and more model-based agents is distinctly small, but consistent (see Fig. 5B). In short, factors such as the stochasticity of the drifting rewards and the transition structure impact the efficiency of model-based control in RL tasks (Kool et al., 2016). Notably, in the original two-step task developed by Daw et al. (2011), model-based control does not pay off (Kool et al., 2016). Yet, people employ model-based control in these tasks. We will discuss these results in more detail in the *Discussion*.

Second, and most pertinent to the goal of this analysis, we found that reliance on alternative rewards during stay/leave decisions also increased average reward. However, the degree to which the agent should rely on alternative rewards depends on their stay/leave choice stochasticity (β_s). We see that no matter how explorative a foraging agent's behavior is, it is always beneficial to incorporate alternative rewards to a certain extent.

Figure 5A shows the result of this analysis. When agents are very exploratory (lower values of β_s), they accrue maximal reward when fully abandoning the average reward rate and only using the rewards at the alternative islands. However, the benefit of incorporating the alternative rewards tapers off when agents become less exploratory. This is highlighted by the plateau in reward-maximizing z around 0.3, indicating that agents should weigh the alternative rewards and the average global reward rate in an approximately 3:7 ratio when deciding whether to continue harvesting.

We provide a more detailed look into the relationship between z, w, β_s, and reward in Fig. 5C. Reliance on model-based control yields a small but reliable increase in reward. However, the extent to which the agent incorporates alternative rewards during stay/leave decisions has a more dramatic effect on total reward. For agents with low to moderate degrees of exploitation, non-zero values of z only lead to increases in average reward rate. For more exploitative agents, however, intermediate reliance on alternative rewards improves task performance, but strong reliance is detrimental. These analyses highlight that our task demands a delicate balance: agents should incorporate alternative reward information into their decisions, but not let them dictate choice behavior entirely. We interpret these results in the *Discussion*.

5. Discussion

In naturalistic foraging environments, we must plan over the structure of the environment to maximize reward. However, most extant foraging tasks (Constantino & Daw, 2015; Hayden et al., 2011; Kane et al., 2017, 2019; Kolling et al., 2012) and computational frameworks (Charnov, 1976; Stephens & Krebs, 1986) do not consider task structure nor the role of planning during foraging (Hall-McMaster & Luyckx, 2019). Here, we addressed this gap using a structured foraging task in which participants alternated between deciding whether to continue harvesting on the current island or to visit another one.

We found that participants not only consulted the average reward rate and expected reward when making stay/leave decisions (akin to the MVT), but also incorporated the reward available on the alternative options into their decision threshold. This result, which replicates that of Hall-McMaster et al. (2021) in a different task, suggests that people use the task's structure to retrieve the identities and values of the next possible islands, which they then use to decide whether or not to leave the current one. This is a form of model-based control, through prospective simulation, during foraging. Moreover, because our task included a stochastic transition structure, we were also able to measure the degree to which people used model-based control to plan their island visits (Daw et al., 2011; Doll et al., 2015; Kool et al., 2016; Pouncy et al., 2021).

This revealed two key influences of environmental structure on foraging behavior. First, participants displayed a mixture of model-based and model-free control in their travel decisions. Participants were more likely to return to a patch after receiving a positive reward prediction error there, but only after common transitions. This aligns with the finding from Hall-McMaster et al. (2021) that participants were more likely to return to patches with faster reward-replenishment rates. After rare transitions, however, there was no group-level effect of prediction error sign. Together, these findings suggest that participants aimed to maximize reward rate, although they used multiple (sometimes conflicting) strategies to do so.

Second, we found that the degree to which people made model-based travel decisions predicted their reliance on alternative rewards when deciding whether to stay on or leave the current island. Thus, when structure is introduced into a foraging task, people adopt strategies more elaborate than those proposed by the MVT. The foraging literature contains many examples of departures from the MVT. For example, it has been suggested that people are prone to overstay their current option because they are sensitive to sunk costs (Wikenheiser, Stephens, & Redish, 2013), discount future rewards (Blanchard & Hayden, 2014), prefer short-term rewards (Kane et al., 2019), or are risk-averse (Eisenreich, Hayden, & Zimmermann, 2019). In contrast, our participants' choice not to solely rely on the expected and average reward rates was adaptive: our computational model simulations demonstrated that incorporating alternative rewards during stay/leave decisions improved performance.

Although the size of this effect was relatively modest, it should be noted that it represents a cross-phase link: model-based control in the travel phase predicted the influence of alternative reward options in the foraging phase. Moreover, we found this effect across two experiments and regardless of how we modeled participants' representation of the alternatives (maximum value, simple average, or probability-weighted average). In each case, more model-based participants let this information guide their decisions more strongly.

One limitation to our analyses is that we often use latent variables, such as expected rewards and expected alternative rewards. These variables cannot be measured directly and therefore need to be inferred using computational modeling. One notable exception is the "model-agnostic" analysis on return intentions, which found analogous behavioral patterns when labeling below- and above-median rewards as losses and gains. This similarity assuages this concern, but more direct measurements would have been ideal.

At first glance, the lack of unique support for the probability-weighted average case suggests participants were not fully leveraging the task structure. Normatively, this calculation (using the true transition probabilities) should be the most accurate. However, participants may have relied on a simpler heuristic for two reasons. First, participants were required to make this computation under time pressure, at the same time as deciding whether to continue harvesting. Second, because island values drift stochastically, the added precision of weighting by transition probabilities may not translate into appreciably better stay/leave choices, making a simpler approximation more appealing.

Crucially, any use of alternative rewards implies sensitivity to task structure. To incorporate them, participants must recognize their current island, recall the other two islands, and track their values. Consistent with this interpretation, we found that participants who displayed greater model-based control in the travel phase also used this knowledge more in the foraging phase. In fact, when we split participants into groups by model-basedness, only the more model-based group showed improved model fit from including alternative rewards, whereas more model-free participants behaved according to a rule closer to the MVT.

Recent work by Harhen and Bornstein (2023) has similarly demonstrated that a structured task environment leads to optimal changes to the decision rule posited by the MVT. In their task, participants were not shown explicit cues associated with each patch (c.f., our island-specific colors). Instead, they needed to use reward information to infer in which

of three possible patch types (low, medium or high reward) they were currently located. Participants in this task displayed substantial overharvesting, the tendency to stay with an option too long in terms of the MVT's threshold policy. The authors, however, argued that this behavioral pattern was the result of uncertainty about the environment. We designed our task to avoid such uncertainty, as participants completed a thorough training procedure to ensure that they encoded the task's structure and its transitions (Feher da Silva & Hare, 2020; Feher da Silva, Lombardi, Edelson, & Hare, 2023). This suggests that participants' strategies were driven by their knowledge of the task structure, and not despite it. At the same time, it would be interesting to investigate whether the main effects in this paper are dependent on the resolution of participants' internal representation of the task structure (which can be captured from behavioral assessments; Karagoz et al., 2024).

We found that participants' patch-leaving decisions were impacted by their knowledge of alternative reward options. This result, which suggests that participants considered the task structure when deciding whether to stick with their current patch, replicates the key findings first reported by Hall-McMaster et al. (2021). In their task, participants deterministically chose which patch to visit after each leave decision. In our task, this choice was less direct. Participants chose between boats that then transitioned them to the next island according to a stochastic transition structure. This allowed us to determine that people used a mixture of model-free and model-based control in this task. Most importantly, this measure predicted the degree to which participants used alternative rewards to decide when to leave. When alternative rewards were higher, our relatively model-based participants left their current island sooner than relatively model-free participants, reflecting the influence of the task structure. Because the study design from Hall-McMaster et al. (2021) does not assess goal-directed planning, it remains ambiguous whether our effects are directly related. However, they invite fascinating follow-up research to test this hypothesis. For example, one could investigate whether alternative rewards affect patch-leaving decisions differently between more direct 'island-selection' and less direct 'boat-selection' blocks of trials (i.e., varying the stochasticity of travel decisions).

Our computational model simulates patch-leaving and travel decisions in the same framework (Hall-McMaster et al., 2021; Kolling & Akam, 2017). These simulations revealed that it is always beneficial to incorporate alternative rewards to some extent when making patch-leaving decisions. However, they also showed that the degree to which this increases reward depends on the stochasticity of those foraging decisions. Particularly exploratory (i.e., random) agents benefit from mostly relying on alternative rewards, but more exploitative (i.e., greedy) agents should mix these values with average reward rate roughly in a 3:7 ratio. These results suggest that when structure is introduced, it becomes beneficial to consider alternative reward options in one's environment when making reward-maximizing patch-leaving decisions, rather than only contrasting expected rewards against overall average rewards.

One possible explanation for this benefit is that participants in our task are unable to immediately revisit the option from which they are currently harvesting rewards. However, because those rewards strongly affect the time-discounted average reward rate, this measure is not an accurate reflection of the reward rate that is to be expected in the near future (at one of the other two options). This is corrected for by incorporating an average of the alternative rewards into the decision threshold. This line of reasoning also explains why it is ideal for very exploratory agents to mostly rely on alternative rewards. In reward-rich environments, exploratory agents are more likely to leave than exploitative agents. Letting that decision be primarily dictated by alternative rewards compensates them for using a suboptimal explore/exploit tradeoff (note that the exploitative agents in Fig. 5C have higher maximal reward rates).

Our results add nuance to the role of explore/exploit tradeoffs in foraging (Addicott et al., 2017; Lloyd, McKay, Sebastian, & Balsters, 2021; van Dooren, de Kleijn, Hommel, & Sjoerds, 2021). They also invite us to consider how people determine their behavioral policy in terms of how much exploration and consideration of alternative rewards they should deploy. One possibility is that one of these parameters is relatively stable (e.g., a trait-level characteristic; Fridhandler, 1986), while the other is more flexible. Alternatively, both parameters may be flexibly adjustable. Can people learn reward-maximizing control policies using RL (Bustamante, Lieder, Musslick, Shenhav, & Cohen, 2021; Lieder, Shenhav, Musslick, & Griffiths, 2018) or through exhaustive evaluation of parameter settings (Shenhav, Botvinick, & Cohen, 2013)? Or does their inclination to follow the MVT (Constantino & Daw, 2015; Turrin, Fagan, Dal Monte, & Chang, 2017; Zhang, Gong, Fougnie, & Wolfe, 2017) hurt them in a task where the reward-maximizing decision rule requires the consideration of task structure? Implementing successful control strategies necessitates successful 'metacontrol' between them (Boureau, Sokol-Hessner, & Daw, 2015), and our results suggest this is a pertinent topic for foraging as well.

Simulation of behavior on our task revealed that while using model-based control to make boat choices increased reward, this effect was relatively modest. Why then, one might ask, did participants plan during the travel phase, especially if it is costly (Kool et al., 2017; Kool, Gershman, & Cushman, 2018)? Interestingly, it is not uncommon to observe planning in tasks that do not incentivize it. Indeed, behavior on the original Daw two-step task famously shows signs of model-based control (Daw et al., 2011), but others (Akam, Costa, & Dayan, 2015; Kool et al., 2016) have shown that it does not produce increased rewards. Thus, even in the absence of explicit rewards, model-based control emerges. One reason is that people may learn that, in the real world, planning is generally useful. Therefore, this assumption may invite them to also engage in goal-directed control in our task. Second, the training of the task structure during the instructions may have encouraged people to assume that it would be useful in the task. Third, our simulations demonstrate that considering the task structure is clearly beneficial during the foraging phase. These factors, plus the high amount of stochasticity, may make it difficult for participants to learn that planning during the travel phase is not particularly helpful for maximizing reward. Moreover, people may simply place value on having control over the task (i.e., knowing where they will go), even in the absence of explicit reward (Deci & Ryan, 1985; Leotti, Iyengar, & Ochsner, 2010).

A body of research suggests that the exertion of model-based planning requires some form of top-down cognitive control (Gershman, Markman, & Otto, 2014; Kool, Cushman, & Gershman, 2018; Kool, Gershman, & Cushman, 2018; Otto, Gershman, Markman, & Daw, 2013; Otto, Skatova, Madlon-Kay, & Daw, 2015; Schad et al., 2014; Shenhav, Rand, & Greene, 2017; Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013). This set of cognitive functions, dependent on computations in the frontal cortex, allows humans to execute novel and effortful tasks by reconfiguring information processing (E. K. Miller & Cohen, 2001). We predict that similar computations underlie the use of model-based planning in our foraging task, either through prospective simulation of future consequences (Doll et al., 2015), or retrospective credit assignment using the task structure (Gershman et al., 2014; Sharp & Eldar, 2024). Research on strategy arbitration in human RL provides several ways in which people's reliance on model-based control can be altered. When people are under cognitive load (Otto et al., 2013) or time pressure (Keramati, Smittenaar, Dolan, & Dayan, 2016), they use less model-based control. When potential rewards are amplified, they use more model-based control (Kool et al., 2017; Patzelt, Kool, Millner, & Gershman, 2019a). To assess the nature of model-based control in the current task, it would be interesting to see whether these factors also modulate people's reliance on task structure during foraging.

Over the last decade, decision-making research has started to uncover links between mental health and foraging strategies (Barack et al., 2024; Bustamante et al., 2024; Raio et al., 2022), as well as model-based planning (Gillan, Kosinski, Whelan, Phelps, & Daw, 2016; Patzelt et al.,

2019a; Ramakrishnan et al., 2025; Wyckmans et al., 2019). Our paradigm, which bridges these two theoretical concepts, is well-positioned to contribute to this research on how various mental health phenotypes affect planning and foraging. Specifically, the task allows us to transdiagnostically explore the interaction between model-based planning and patch-leaving decisions within the same framework. It may be that some transdiagnostic individual differences could predict participants' tendency to use the task structure (e.g., schizotypy), whereas others may modulate the perceived reward rate in the MVT (e.g., depressed mood). For example, if the exertion of model-based control carries an intrinsic effort cost (Kool & Botvinick, 2018), then this may lead to a reduced reliance on task structure during both travel and stay/leave decisions in individuals experiencing reduced motivation (i.e., apathy, anhedonia), or they may guide the adaptation of other choice strategies. Similarly, individuals with depression, which is associated with differences in motivational processes (Bustamante et al., 2024; Grahek, Shenhav, Musslick, Krebs, & Koster, 2019; Treadway, Bossaller, Shelton, & Zald, 2012; Yang et al., 2014), may also exhibit this behavior. Additionally, it has been argued that the process of forming internal representations of the environment, or cognitive maps, is different in individuals with and without schizophrenia (Karagoz et al., 2025; Musa, Khan, Mujahid, & El-Gaby, 2022; Nour, Liu, El-Gaby, McCutcheon, & Dolan, 2025). In the context of our current task, this would make the intriguing prediction that people with schizophrenia would more closely follow the MVT rather than rely on an internal representation of task structure. Persons with schizophrenia also display differences in reward learning, anticipation/prediction, and translating reward information into action plans (Barch, Pagliaccio, & Luking, 2016; Barch & Treadway, 2014; Culbreth, Moran, Kandala, Westbrook, & Barch, 2020; Green, Horan, Barch, & Gold, 2015; Moran, Prevost, Culbreth, & Barch, 2023). We predict that these differences could be reflected in our task through decreased reward outcomes, differences in return intentions, and a reduced reliance on the task structure. Recent research on foraging in individuals with attention-deficit/hyperactivity disorder reflects an increased tendency to explore while foraging (Barack et al., 2024). Based on our simulations of the task, this could translate to a stronger reliance on alternative reward information when making stay/leave decisions in these individuals. Generally, examining these behaviors through a transdiagnostic lens (Gillan et al., 2016; Patzelt et al., 2019a; Patzelt, Kool, Millner, & Gershman, 2019b) could enable enhancing the specificity of interventions for individuals with differences in planning or information processing behaviors.

In summary, we have shown that people rely on alternative reward information in structured environments, and that this tendency is adaptive and can be thought of as a form of model-based control. Though conventional foraging tasks already appeal to notions of ecological validity, we believe that our task goes further by incorporating structure. Of course, not all real-world foraging environments should be treated as structured (either because they are not or because people do not know they are), but the world is inherently structured. If agents in those environments try to maximize reward, then our analyses suggest they should be sensitive to this.

CRediT authorship contribution statement

Thea R. Zalabak: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Laura A. Bustamante:** Writing – review & editing, Methodology, Conceptualization. **Wouter Kool:** Writing – review & editing, Supervision, Resources, Funding acquisition, Conceptualization.

Funding

This work was supported by a Multi-University Research Initiative grant (ONR/DoD N00014–23-1-2792) to WK. Computations were performed using the facilities of the Washington University Research Computing and Informatics Facility (RCIF). The RCIF has received funding from NIH S10 program grants: 1S10OD025200-01A1 and 1S10OD030477–01.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Wouter Kool reports financial support was provided by Office of Naval Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank the members of the Control and Decision Making Lab for their advice and assistance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2025.106367.

Data availability

Data, task, and analysis code are publicly available at the Open Science Framework repository https://osf.io/xumy6/.

References

Addicott, M. A., Pearson, J. M., Sweitzer, M. M., Barack, D. L., & Platt, M. L. (2017). A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacology, 42*(10), 1931–1939. https://doi.org/10.1038/npp.2017.108

Akam, T., Costa, R., & Dayan, P. (2015). Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS Computational Biology, 11*(12), Article e1004648. https://doi.org/10.1371/journal.pcbi.1004648

Alejandro, R. J., & Holroyd, C. B. (2024). Hierarchical control over foraging behavior by anterior cingulate cortex. *Neuroscience & Biobehavioral Reviews, 160*, Article 105623. https://doi.org/10.1016/j.neubiorev.2024.105623

Barack, D. L., Ludwig, V. U., Parodi, F., Ahmed, N., Brannon, E. M., Ramakrishnan, A., & Platt, M. L. (2024). Attention deficits linked with proclivity to explore while foraging. *Proceedings of the Royal Society B: Biological Sciences, 291*(2017), 20222584. https://doi.org/10.1098/rspb.2022.2584

Barch, D. M., Pagliaccio, D., & Luking, K. (2016). Mechanisms underlying motivational deficits in psychopathology: Similarities and differences in depression and schizophrenia. In E. H. Simpson, & P. D. Balsam (Eds.), *Behavioral neuroscience of motivation* (pp. 411–449). Springer International Publishing. https://doi.org/10.1007/7854_2015_376.

Barch, D. M., & Treadway, M. T. (2014). Effort, anhedonia, and function in schizophrenia: Reduced effort allocation predicts amotivation and functional impairment. *Journal of Abnormal Psychology, 123*(2), 387–397. https://doi.org/10.1037/a0036299

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Blanchard, T. C., & Hayden, B. Y. (2014). Neurons in dorsal anterior cingulate cortex signal postdecisional variables in a foraging task. *Journal of Neuroscience, 34*(2), 646–655. https://doi.org/10.1523/JNEUROSCI.3151-13.2014

Bolenz, F., Kool, W., Reiter, A. M., & Eppinger, B. (2019). Metacontrol of decision-making strategies in human aging. *eLife, 8*, Article e49154. https://doi.org/10.7554/eLife.49154

Boureau, Y.-L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding how to decide: Self-control and Meta-decision making. *Trends in Cognitive Sciences, 19*(11), 700–710. https://doi.org/10.1016/j.tics.2015.08.013

Bustamante, L. A., Barch, D. M., Solis, J., Oshinowo, T., Grahek, I., Konova, A. B., … Cohen, J. D. (2024). Major depression symptom severity associations with willingness to exert effort and patch foraging strategy. *Psychological Medicine, 54*(15), 4396–4407. https://doi.org/10.1017/S0033291724002691

Bustamante, L. A., Lieder, F., Musslick, S., Shenhav, A., & Cohen, J. (2021). Learning to overexert cognitive control in a Stroop task. *Cognitive, Affective, & Behavioral Neuroscience, 21*(3), 453–471. https://doi.org/10.3758/s13415-020-00845-x

Bustamante, L. A., Oshinowo, T., Lee, J. R., Tong, E., Burton, A. R., Shenhav, A., … Daw, N. D. (2023). Effort foraging task reveals positive correlation between individual differences in the cost of cognitive and physical effort in humans.

Proceedings of the National Academy of Sciences, 120(50), Article e2221510120. https://doi.org/10.1073/pnas.2221510120
Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology, 9*(2), 129–136. https://doi.org/10.1016/0040-5809(76)90040-X
Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098
Constantino, S. M., & Daw, N. D. (2015). Learning the opportunity cost of time in a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience, 15*(4), 837–853. https://doi.org/10.3758/s13415-015-0350-y
Culbreth, A. J., Moran, E. K., Kandala, S., Westbrook, A., & Barch, D. M. (2020). Effort, Avolition, and motivational experience in schizophrenia: Analysis of behavioral and neuroimaging data with relationships to daily motivational experience. *Clinical Psychological Science, 8*(3), 555–568. https://doi.org/10.1177/2167702620901558
Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027
Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711. https://doi.org/10.1038/nn1560
Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer US. https://doi.org/10.1007/978-1-4899-2271-7
Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience, 18*(5), 767–772. https://doi.org/10.1038/nn.3981
van Dooren, R., de Kleijn, R., Hommel, B., & Sjoerds, Z. (2021). The exploration-exploitation trade-off in a foraging task is affected by mood-related arousal and valence. *Cognitive, Affective, & Behavioral Neuroscience, 21*(3), 549–560. https://doi.org/10.3758/s13415-021-00917-6
Drummond, N., & Niv, Y. (2020). Model-based decision making and model-free learning. *Current Biology, 30*(15), R860–R865. https://doi.org/10.1016/j.cub.2020.06.051
Eisenreich, B. R., Hayden, B. Y., & Zimmermann, J. (2019). Macaques are risk-averse in a freely moving foraging task. *Scientific Reports, 9*(1), 15091. https://doi.org/10.1038/s41598-019-51442-z
Feher da Silva, C., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour, 4*(10), 1053–1066. https://doi.org/10.1038/s41562-020-0905-y
Feher da Silva, C., Lombardi, G., Edelson, M., & Hare, T. A. (2023). Rethinking model-based and model-free influences on mental effort and striatal prediction errors. *Nature Human Behaviour, 7*(6), 956–969. https://doi.org/10.1038/s41562-023-01573-1
Frankenhuis, W. E., Panchanathan, K., & Barto, A. G. (2019). Enriching behavioral ecology with reinforcement learning methods. *Behavioural Processes, 161*, 94–100. https://doi.org/10.1016/j.beproc.2018.01.008
Fridhandler, B. M. (1986). Conceptual note on state, trait, and the state-trait distinction. *Journal of Personality and Social Psychology, 50*(1), 169–174.
Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology, 71*, 1–6. https://doi.org/10.1016/j.jmp.2016.01.006
Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General, 143*(1), 182–194. https://doi.org/10.1037/a0030844
Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife, 5*, Article e11305. https://doi.org/10.7554/eLife.11305
Grahek, I., Shenhav, A., Musslick, S., Krebs, R. M., & Koster, E. H. W. (2019). Motivation and cognitive control in depression. *Neuroscience & Biobehavioral Reviews, 102*, 371–381. https://doi.org/10.1016/j.neubiorev.2019.04.011
Green, M. F., Horan, W. P., Barch, D. M., & Gold, J. M. (2015). Effort-based decision making: A novel approach for assessing motivation in schizophrenia. *Schizophrenia Bulletin, 41*(5), 1035–1044. https://doi.org/10.1093/schbul/sbv071
Hall-McMaster, S., Dayan, P., & Schuck, N. W. (2021). Control over patch encounters changes foraging behavior. *iScience, 24*(9), Article 103005. https://doi.org/10.1016/j.isci.2021.103005
Hall-McMaster, S., & Luyckx, F. (2019). Revisiting foraging approaches in neuroscience. *Cognitive, Affective, & Behavioral Neuroscience, 19*(2), 225–230. https://doi.org/10.3758/s13415-018-00682-z
Harhen, N. C., & Bornstein, A. M. (2023). Overharvesting in human patch foraging reflects rational structure learning and adaptive planning. *Proceedings of the National Academy of Sciences, 120*(13), Article e2216524120. https://doi.org/10.1073/pnas.2216524120
Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience, 14*(7), 933–939. https://doi.org/10.1038/nn.2856
Kane, G. A., Bornstein, A. M., Shenhav, A., Wilson, R. C., Daw, N. D., & Cohen, J. D. (2019). Rats exhibit similar biases in foraging and intertemporal choice tasks. *eLife, 8*, Article e48429. https://doi.org/10.7554/eLife.48429
Kane, G. A., Vazey, E. M., Wilson, R. C., Shenhav, A., Daw, N. D., Aston-Jones, G., & Cohen, J. D. (2017). Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task. *Cognitive, Affective, & Behavioral Neuroscience, 17*(6), 1073–1083. https://doi.org/10.3758/s13415-017-0531-y
Karagoz, A. B., Moran, E. K., Barch, D. M., Kool, W., & Reagh, Z. M. (2025). Evidence for shallow cognitive maps in schizophrenia. *Cognitive, Affective, & Behavioral Neuroscience*. https://doi.org/10.3758/s13415-025-01283-3
Karagoz, A. B., Reagh, Z. M., & Kool, W. (2024). The construction and use of cognitive maps in model-based control. *Journal of Experimental Psychology: General, 153*(2), 372–385. https://doi.org/10.1037/xge0001491
Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences, 113*(45), 12868–12873. https://doi.org/10.1073/pnas.1609094113
Kolling, N., & Akam, T. (2017). (Reinforcement?) Learning to forage optimally. *Current Opinion in Neurobiology, 46*, 162–169. https://doi.org/10.1016/j.conb.2017.08.008
Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science, 336*(6077), 95–98. https://doi.org/10.1126/science.1216930
Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour, 2*(12), 899–908. https://doi.org/10.1038/s41562-018-0401-9
Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLoS Computational Biology, 12*(8), Article e1005090. https://doi.org/10.1371/journal.pcbi.1005090
Kool, W., Cushman, F. A., & Gershman, S. J. (2018). Chapter 7—Competition and cooperation between multiple reinforcement learning systems. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-directed decision making* (pp. 153–178). Academic Press. https://doi.org/10.1016/B978-0-12-812098-9.00007-3.
Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological Science, 28*(9), 1321–1333. https://doi.org/10.1177/0956797617708288
Kool, W., Gershman, S. J., & Cushman, F. A. (2018). Planning complexity registers as a cost in metacontrol. *Journal of Cognitive Neuroscience, 30*(10), 1391–1404. https://doi.org/10.1162/jocn_a_01263
Le Heron, C., Kolling, N., Plant, O., Kienast, A., Janska, R., Ang, Y.-S., … Apps, M. A. J. (2020). Dopamine modulates dynamic decision-making during foraging. *The Journal of Neuroscience, 40*(27), 5273–5282. https://doi.org/10.1523/JNEUROSCI.2586-19.2020
de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods, 47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y
Leotti, L. A., Iyengar, S. S., & Ochsner, K. N. (2010). Born to choose: The origins and value of the need for control. *Trends in Cognitive Sciences, 14*(10), 457–463. https://doi.org/10.1016/j.tics.2010.08.001
Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology, 14*(4), Article e1006043. https://doi.org/10.1371/journal.pcbi.1006043
Lloyd, A., McKay, R., Sebastian, C. L., & Balsters, J. H. (2021). Are adolescents more optimal decision-makers in novel environments? Examining the benefits of heightened exploration in a patch foraging paradigm. *Developmental Science, 24*(4), Article e13075. https://doi.org/10.1111/desc.13075
Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167
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
Moran, E. K., Prevost, C., Culbreth, A. J., & Barch, D. M. (2023). Effort-cost decision-making in psychotic and mood disorders. *Journal of Psychopathology and Clinical Science.*. https://doi.org/10.1037/abn0000822
Morimoto, J. (2019). Foraging decisions as multi-armed bandit problems: Applying reinforcement learning algorithms to foraging data. *Journal of Theoretical Biology, 467*, 48–56. https://doi.org/10.1016/j.jtbi.2019.02.002
Musa, A., Khan, S., Mujahid, M., & El-Gaby, M. (2022). The shallow cognitive map hypothesis: A hippocampal framework for thought disorder in schizophrenia. *Schizophrenia, 8*(1), 1–11. https://doi.org/10.1038/s41537-022-00247-7
Navarro, D. J., Tran, P., & Baz, N. (2018). Aversion to option loss in a restless bandit task. *Computational Brain & Behavior, 1*(2), 151–164. https://doi.org/10.1007/s42113-018-0010-8
Nour, M. M., Liu, Y., El-Gaby, M., McCutcheon, R. A., & Dolan, R. J. (2025). Cognitive maps and schizophrenia. *Trends in Cognitive Sciences, 29*(2), 184–200. https://doi.org/10.1016/j.tics.2024.09.011
Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science, 24*(5), 751–761. https://doi.org/10.1177/0956797612463080
Otto, A. R., Skatova, A., Madlon-Kay, S., & Daw, N. D. (2015). Cognitive control predicts use of model-based reinforcement learning. *Journal of Cognitive Neuroscience, 27*(2), 319–333. https://doi.org/10.1162/jocn_a_00709
Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019a). Incentives boost model-based control across a range of severity on several psychiatric constructs. *Biological Psychiatry, 85*(5), 425–433. https://doi.org/10.1016/j.biopsych.2018.06.018
Patzelt, E. H., Kool, W., Millner, A. J., & Gershman, S. J. (2019b). The transdiagnostic structure of mental effort avoidance. *Scientific Reports, 9*(1), 1689. https://doi.org/10.1038/s41598-018-37802-1
Pouncy, T., Tsividis, P., & Gershman, S. J. (2021). What is the model in model-based planning? *Cognitive Science, 45*(1), Article e12928. https://doi.org/10.1111/cogs.12928
Raio, C. M., Biernacki, K., Kapoor, A., Wengler, K., Bonagura, D., Xue, J., … Konova, A. B. (2022). Suboptimal foraging decisions and involvement of the ventral tegmental area in human opioid addiction (p. 2022.03.24.485654). *bioRxiv*. https://doi.org/10.1101/2022.03.24.485654
Ramakrishnan, S. A., Shaik, R. B., Kanagamani, T., Neppala, G., Chen, J., Fiore, V. G., … Parvaz, M. A. (2025). Impaired arbitration between reward-related decision-making strategies in alcohol users compared to alcohol non-users: A computational modeling

study. *NPP—Digital Psychiatry and Neuroscience, 3*(1), 1. https://doi.org/10.1038/s44277-024-00023-8

Schad, D. J., Jünger, E., Sebold, M., Garbusow, M., Bernhardt, N., Javadi, A.-H., … Huys, Q. J. M. (2014). Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Frontiers in Psychology, 5*. https://doi.org/10.3389/fpsyg.2014.01450

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

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron, 79*(2), 217–240. https://doi.org/10.1016/j.neuron.2013.07.007

Shenhav, A., Rand, D. G., & Greene, J. D. (2017). The relationship between intertemporal choice and following the path of least resistance across choices, preferences, and beliefs. *Judgment and Decision making, 12*(1), 1–18. https://doi.org/10.1017/S1930297500005209

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron, 80*(4), 914–919. https://doi.org/10.1016/j.neuron.2013.08.009

Stephens, D. W. (2008). Decision ecology: Foraging and the ecology of animal decision making. *Cognitive, Affective, & Behavioral Neuroscience, 8*(4), 475–484. https://doi.org/10.3758/CABN.8.4.475

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton University Press.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements, 2*(4), i–109. https://doi.org/10.1037/h0092987

Treadway, M. T., Bossaller, N. A., Shelton, R. C., & Zald, D. H. (2012). Effort-based decision-making in major depressive disorder: A translational model of motivational anhedonia. *Journal of Abnormal Psychology, 121*(3), 553–558. https://doi.org/10.1037/a0028813

Turrin, C., Fagan, N. A., Dal Monte, O., & Chang, S. W. C. (2017). Social resource foraging is guided by the principles of the marginal value theorem. *Scientific Reports, 7*(1), 11274. https://doi.org/10.1038/s41598-017-11763-3

Wikenheiser, A. M., Stephens, D. W., & Redish, A. D. (2013). Subjective costs drive overly patient foraging strategies in rats on an intertemporal foraging task. *Proceedings of the National Academy of Sciences, 110*(20), 8308–8313. https://doi.org/10.1073/pnas.1220738110

Wittmann, M. K., Kolling, N., Akaishi, R., Chau, B. K. H., Brown, J. W., Nelissen, N., & Rushworth, M. F. S. (2016). Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex. *Nature Communications, 7*(1), Article 12327. https://doi.org/10.1038/ncomms12327

Wyckmans, F., Otto, A. R., Sebold, M., Daw, N., Bechara, A., Saeremans, M., … Noël, X. (2019). Reduced model-based decision-making in gambling disorder. *Scientific Reports, 9*(1), 19625. https://doi.org/10.1038/s41598-019-56161-z

Yang, X.-H., Huang, J., Zhu, C.-Y., Wang, Y.-F., Cheung, E. F. C., Chan, R. C. K., & Xie, G.-R. (2014). Motivational deficits in effort-based decision making in individuals with subsyndromal depression, first-episode and remitted depression patients. *Psychiatry Research, 220*(3), 874–882. https://doi.org/10.1016/j.psychres.2014.08.056

Zhang, J., Gong, X., Fougnie, D., & Wolfe, J. M. (2017). How humans react to changing rewards during visual foraging. *Attention, Perception, & Psychophysics, 79*(8), 2299–2309. https://doi.org/10.3758/s13414-017-1411-9