Structure learning as a mechanism of overharvesting

Nora C. Harhen (nharhen@uci.edu)

Aaron M. Bornstein (aaron.bornstein@uci.edu)
Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92697 USA

Abstract

In patch leaving problems, foragers must decide between engaging with a currently available, but depleting, patch of resources or foregoing it to search for another, potentially better patch. Overharvesting, or staying in the patch longer than what is optimally prescribed, is widely observed in these problems. Most previous explanations for this phenomenon focus on how foragers' mis-estimations of the environment could produce overharvesting. They suggest that if the forager correctly learned the environment's quality, then they would behave according to Marginal Value Theorem (MVT). However, this proposal rests on the assumption that the forager has full knowledge of the environment's structure. Rarely does this occur in the real world. Instead, foragers must learn the structure of their environment. Here, we model foragers as pairing an optimal decision rule with an optimal learning procedure that allows for the possibility of heterogeneously-structured (i.e. multimodal) reward distributions. We then show that this model can appear to produce overharvesting, as measured by the common optimality criterion, when applied to the usual tasks, which employ homogeneous reward distributions. This model accounts for behavior in a previous serial stay/leave task, and generates novel predictions regarding sequential effects that agree with participant behavior. Taken together, these results are consistent with overharvesting reflecting optimality with respect to a different set of conditions than MVT and suggests that MVT's definition of optimality may need to be adjusted to account for behavior in more naturalistic contexts.

Keywords: foraging; structure learning; reinforcement learning; decision-making;

Introduction

Marginal Value Theorem (MVT; Charnov, 1976) provides the optimal decision rule for maximizing reward in patch-foraging tasks: leave the current patch of resources when the estimated reward rate drops below the average reward rate of the global environment. The rule sets aside the question of how the environment is learned - it is assumed that the forager has full knowledge of the environment: overall quality and any structure (states) to the distribution of rewards.

Foragers, including rodents, primates, and humans, however, demonstrate a consistent bias towards staying too long in the current patch, or "overharvesting" it. Several explanations have been proposed. Some accounts accept that the forager has full knowledge of the environment and attribute overharvesting to different biases and goals irrespective of the environment. These include sunk costs (Wikenheiser et al, 2013), impatient time preferences (Kane et al, 2019), and maximizing marginal utility over reward (Constantino & Daw, 2015). However, rarely is a forager fully certain of their environment. Given this broken assumption, previous work has focused on adapting MVT to include learning of the environment's quality (e.g. average richness) and explored errors in this learning as a potential mechanism of overharvesting. For instance, biased updating of beliefs can explain overharvesting in non-stationary environments (Garrett & Daw, 2020). Uncertainty over environment quality from insufficient experience can also explain patterns of over and under harvesting (Kilpatrick, Davidson, & El Hady, 2021). However, this work suggests that with sufficient experience deviations from MVT optimality should diminish.

In previously proposed mechanisms of overharvesting, less focus has been placed on how the environment's structure (i.e. distribution of rewards) is learned relative to how its quality has been learned. Most prior work assumes that the forager knows the different patch types within the environment (e.g. richer or poorer). However, in naturalistic settings, foragers are not given this information, they must infer it from experience alone. Here, we question this assumption and propose that the structure learning process can itself explain the appearance of overharvesting. We developed a computational model of how foragers might learn the structure of the environment's reward distribution (the number of modes). First, we show in simulation that the model can generate overharvesting in a single patch-type environment. Then, we examine if the model can explain behavior in a previous serial stay/leave decision task. Finally, we test a novel prediction of the model — that harvesting behavior depends on the order of shifts in volatility — and show that human behavior agrees with the model's predictions. Taken together, these results demonstrate a novel mechanism for overharvesting and, more broadly, brings into question whether MVT is the right optimum to compare behavior to as its assumptions fail to meet the conditions of natural environments.

Methods

Model

Structure learning model We apply rational models of categorization (Anderson, 1991; Sanborn, Griffiths, & Navarro, 2006) to capture how foragers learn the latent structure of the environment. The model (adapted from Harhen, Hartley, & Bornstein, 2021) allows for the possibility that patches belong to different categories varying in richness. The num-

ber of patch categories is not pre-specified and is instead inferred from experience. The forager begins with assumptions of how their observations were generated. They assume that: 1. Rewards exponentially deplete with each harvest; 2. Each patch belongs to a category; 3. Each category is characterized by a unique distribution over depletion rates; 4. A new patch is more likely to belong to a popular category (i.e. many category members); 5. There is some probability that a new patch will belong to a new, previously unobserved category.

Foragers combine these prior beliefs with observed data (depletion rates) to generate new beliefs. The forager then compares the expected reward from harvesting the current patch, v_{stay}, to a reference point, v_{leave}. V_{stay} is estimated as the last received reward multiplied by the expected depletion rate. Having categorized the patch, the forager can better predict the upcoming depletion rate.

MVT's reference point averages overall all previous patches as it assumes homogeneous reward distribution. Our model allows for the possibility that the environment is heterogeneous (e.g. has multiple patch types or multiple modes), so the reward rate of one patch may not be predictive of all other patches' reward rate. Consequently, our model's reference point uses patch experiences integrated over a much shorter time-scale. The reference point for the current patch depends only on the reward rate of the last encountered patch of a different type/category.

Generative model. At trial t, c_{t-1} reflects the assignment of all patches up until the current trial. Each new patch can be assigned to an existing category or a new category. The prior probability of it belonging to an existing category, k, is proportional to the number of patches already assigned to that category, N_k, at trial t. The prior probability of it belonging to a new category is proportional to the parameter α which reflects how densely distributed patches are across categories.

$$P(c_t = k|c_{t-1}) = \begin{cases} \frac{N_k}{t-1+\alpha} & \text{if k is an old cluster} \\ \frac{\alpha}{t-1+\alpha} & \text{if k is a new cluster} \end{cases} \quad (1)$$

Each category is associated with its own normal distribution over depletion rates, parameterized by μ_c and σ_c^2. When a patch is assigned to an existing category, depletion rates observed in that patch update the category-specific distribution.

Inference model. Given a set of observed depletion rates up to trial t, D_t, the forager's posterior beliefs over patch assignments to categories, c_t, are described by:

$$P(c_t|D_t) = \frac{P(D_t|c_t)P(c_t)}{p(D_t)} \quad (2)$$

Exact computation of the posterior is computationally intractable, so we use particle filtering as an approximate inference algorithm (Gershman, Niv, & Blei, 2010; Sanborn, Griffiths, & Navarro, 2006). The posterior is represented with a set of m particles. Each particle represents a possible assignment of patches to categories. Some particles will have the same category assignments. The posterior probability of a category assignment is proportional to the number of particles within the set which contain that assignment. To approximate the posterior distribution, we can average over the particles:

$$P(c_t|D_t) \approx \frac{1}{M}\sum_{m=1}^{M} \delta(c_t - c_t^{(m)}) \quad (3)$$

where $\delta(\cdot)$ is 1 when its input is 0, and 0 otherwise.

We can then approximate the prior distribution over category assignments for t+1 trials with

$$\begin{aligned} P(c_{t+1}|D_t) &= \sum_{c_t} P(c_{t+1}|c_t)P(c_t|D_t) \\ &\approx \sum_{c_t} P(c_{t+1}|c_t)\frac{1}{M}\sum_{m=1}^{M}\delta(c_t - c_t^{(m)}) \\ &= \frac{1}{M}\sum_{m=1}^{M} P(c_{t+1}|c_{t+1}^{(m)}) \end{aligned} \quad (4)$$

We can then approximate the posterior for trial $t+1$ with:

$$\begin{aligned} P(c_{t+1}|D_{t+1}) &\propto \sum_{c_t} P(d_{t+1}|c_{t+1}, D_t)P(c_{t+1}|D_t) \\ &\approx \frac{1}{M}\sum_{m=1}^{M} P(d_{t+1}|c_{t+1}, D_t)P(c_{t+1}|c_{t+1}^{(m)}) \end{aligned} \quad (5)$$

M samples are drawn from this distribution to create a new particle set. 50 particles were used for all simulations. An intermediate number of particles allows for fairly accurate prediction while being psychologically plausible and capable of capturing the variability and order sensitivity people display (Sanborn, Griffiths, & Navarro, 2006).

Prediction To predict how much the harvest will deplete next, possible depletion rates are sampled from the forager's inferred generative model of the environment, and these samples are averaged over. Depletion rates are sampled in the following way: first, a category is drawn with probability proportional to its posterior probability, and then, a depletion rate is drawn from the category-specific normal distribution over depletion rates. In our simulations, we used 1000 samples.

Single Patch Type Learning model Patches are assumed to all belong to the same category. This is equivalent to setting alpha to 0. This should generate behavior similar to what MVT would produce, with the additional ability to account for the variance of observed rewards.

Making a choice

To make a decision, the forager compares the value of staying with the value of leaving. The value of staying, v_{stay}, is the reward received from the last harvest multiplied by the predicted depletion rate. The value of leaving, v_{leave}, is calculated as the average reward rate in the last visited patch of a different category multiplied by the time that would be spent harvesting it. This serves as a more dynamic, shorter time scale reference point than MVT's.

Experiment 1: Simulating the structure learning model in single patch type environments

We propose that overharvesting may emerge from inferring more structure in the environment than what is actually present. In particular, inferring that the environment has multiple patch types when it is, in fact, a single highly variable patch type. Simulated agents visited patches to harvest for resources. They decided between harvesting the current exponentially depleting patch or spending more time to travel to a new, unharvested patch (harvest time = 3.5, travel time = 15.5 sec). Depletion rates were drawn from a Beta distribution parameterized by a = 1.5 and b = 1.5 . The mean depletion rate was 0.5 with a SD of 0.25.

Experiment 2: Reanalysis of Constantino & Daw (2015)

We fit our model to data from Constantino & Daw (2015). Participants harvested trees for apples. After each harvest, they could decide between harvesting again or traveling to a new tree and incurring a time delay. The number of apples gained per harvest depleted exponentially. Participants foraged in four different environments that varied in their mean depletion rate and travel time delay. These two features controlled the overall richness of the environment (i.e. higher depletion rate $\rightarrow$ richer, shorter travel time $\rightarrow$ richer). Critically, in this experiment, participants were told when (though not how) the environment changed.

Experiment 3: Novel task

Participants We recruited 82 participants from Amazon Mechanical Turk (ages 23 - 63, Mean= 38, SD=10). Participation was restricted to workers who had completed at least 100 prior studies with at least 99% approval rate. Participants were paid $6 as a base payment and a bonus contingent on performance ($0-$4). We excluded 7 participants for failing at least one catch trial or having average patch residence time 2 standard deviations above or below the group mean.

Procedure We adapted from Harhen et al (2021) a novel variant of the Constantino & Daw (2015). The task was framed as a space mining game where participants were told to collect as many space gems as possible. On each trial, participants had to decide if they wanted to continue digging for gems on the current planet or travel to a new planet. If they stayed and harvested, they watched a short animation of an astronaut digging and then the reward would be displayed. If they chose to travel to a new planet, they watched an animation of a flying rocket ship and then an image of a trial-unique alien was displayed for 5 seconds. Participants had 2 seconds to make their choice. If they did not make a decision, they had to wait another 2 seconds before making another choice. To ensure participants' reaction times did not affect their reward rate, the reaction time (RT) was subtracted from the ensuing dig or travel animation display time.

Participants completed 6 blocks lasting 5 minutes each. Blocks varied in the spread of depletion rates experienced. Depletion rates in highly volatile blocks were sampled from a Beta distribution with parameters a = 1 and b = 1. The mean depletion rate was 0.5 with a SD of 0.29. Depletion rates in the least volatile blocks were sampled from a Beta distribution with a = 20 and b = 20 (Mean = 0.5, SD = 0.078). In the medium volatility block, depletion rates were sampled from a Beta distribution parameterized by a = 4 and b = 4 (Mean = 0.5, SD = 0.17). Participants were told when a new block began, but were not told if and how it changed. Participants were placed in one of two conditions that differed in the order of blocks encountered. In the high early volatility condition the first two blocks were the most volatile, and the third and fourth blocks were the least volatile. In the other condition (low early volatility), the order of the blocks first four blocks were reversed. In both conditions, the last two blocks were of medium volatility. By matching the last two blocks on volatility, we were able to directly compare behavior between the conditions.

Model fitting procedure

The MVT learning model's free parameters were beta (softmax temperature), c (stay/leave bias), alpha (learning rate), and ρ_0 (initial global reward rate). For both the tasks, the free parameters for the structure learning model were the prior over cluster dispersion (alpha), and prior over environment richness. For the Constantino & Daw (2015) task, participant data was characterized by the mean patch residence time (PRT) in each of the blocks. For the novel variant of the task, participant data was characterized by the mean patch residence time (PRT) relative to MVT optimal in each of the blocks. We compared this to the same measures predicted by the model. The loss for a parameter set was calculated as the sum of squared error between the participant's data and the model's simulated data averaged across 10 simulations. 500 sets of parameters were sampled from a Sobol Sequence, and the set of parameters that produced the lowest sum of squared error was chosen. Generating candidate parameter sets from a Sobol Sequence rather than a grid, can provide superior fits, particularly, when there are more than two parameters (Bergstra & Bengio, 2012).

Model comparison

To compare models, we used cross validation. We held out one test block and then fit the model using the PRTs for the remaining blocks. The model error was then measured by taking the absolute difference between the model prediction for the held-out block and the participant's measure for that block. We repeated this procedure for every possible combination of fit blocks and test block and then averaged over the errors to compute the cross validation score.

Results

Experiment 1: Simulating the structure learning model in single patch type environments

We first simulated variants of the model which differed in whether they allowed for the possibility of multiple patch

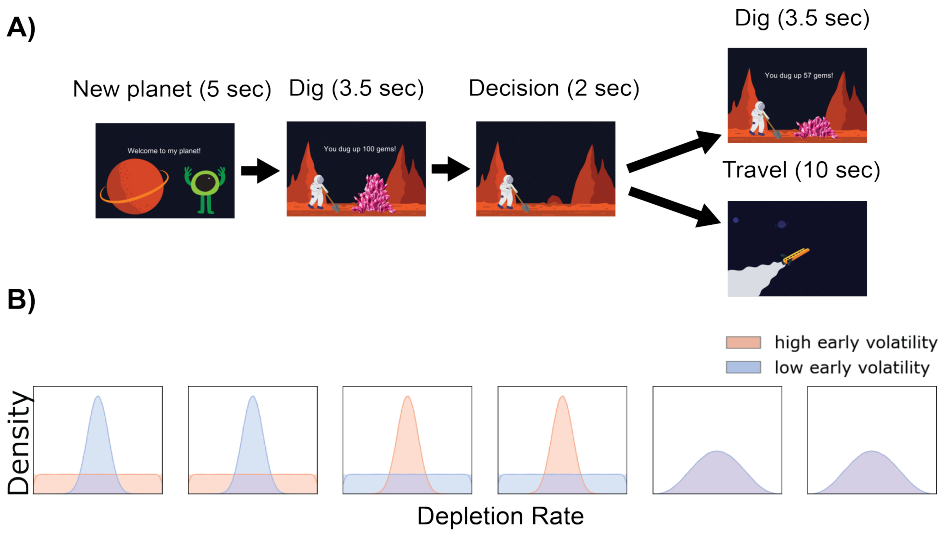

Figure 1: **Task Designs**. **A.** Participants sequentially decide whether to dig or travel to a new planet. **B. Novel task volatility structure** The experiment is broken up into six blocks. Blocks differ in the distribution from which depletion rates are sampled. Some have high variance, others have low variance, and some fall in between. The two conditions, high early volatility and low early volatility, differ only in the order of blocks.

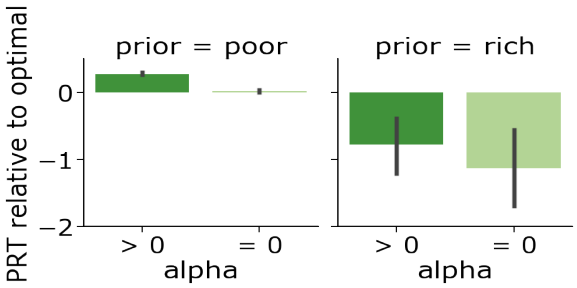

Figure 2: **Results from Experiment 1** Overharvesting and under harvesting behavior depends on both the prior over environment complexity, alpha, and the prior over environment richness. Error bars are 95% CI.

types in highly variable a single patch environment. When allowing for multiple patch types to be inferred (alpha > 0), simulated agents did so. Consistent with our prediction, rational behavior that began with this mismatch between environment type and assumptions resulted in overharvesting when the environment was initially believed to be poor (Figure 2). However, when the true environment structure was assumed (single patch type, alpha = 0), behavior was MVT-optimal. Underharvesting behavior emerged from an initial belief that the environment was rich regardless of assumptions about the environment's structure.

Experiment 2: Reanalysis of Constantino & Daw (2015)

Constantino & Daw found that Marginal Value Theorem (MVT) with an error-driven learning rule better explained participants' data than a temporal-difference learning model. The MVT learning model had four free parameters: learning rate (α), softmax temperature (β), initial global reward rate (ρ_0), and stay-leave bias (c). C captured an individual's bias to stay in the current patch. We reasoned that this bias parameter would be instrumental in capturing behavior that deviated from MVT optimality. To test this hypothesis, we fit the data with the MVT model with and without c. We found that c was indeed critical to capturing participant's overharvesting behavior (Figure 3, t(24) = -6.04, $p < 0.0001$). Given the importance of this parameter, how does this bias emerge?

When comparing both the MVT and the structure learning model with a stay/leave bias, neither was superior to the other (t(24) = -1.23, p = 0.23). However, when comparing the MVT and structure learning model without the stay/leave bias, the structure learning model was superior (Figure 3, Table 1, t(24) = 3.63, p = 0.001). Taken together, these results suggest that the (nonstandard) stay/leave bias added to the MVT model in Constantino & Daw (2015) — added to place it on par with the temporal-difference learning model used as an alternative hypothesis in that study — was a primary factor in the fit quality of that model, perhaps due to the fact that the long blocks in that experiment allowed learning to reach a steady state. Here, we show that the optimal structure learning procedure can account for much if not all of the variance that this parameter adds, while rooting the behavior in a principled, rational learning approach.

Experiment 3: Novel Task

We next tested whether human behavior reflected a novel prediction of the structure learning model, namely sensitivity to the order in which patch volatility is experienced (Figure 1b).

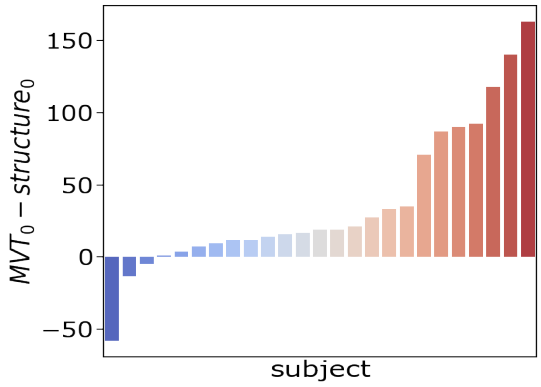

Figure 3: **Results from Experiment 2**. Each bar reflects the difference in cross-validation scores between the MVT learning model without c, the stay-leave bias parameter, and the structure learning model, also without a c, for an individual participant. Positive values indicate the structure learning model provides a better fit to the participant's data. Overall, 22 out of 25 participants were better fit by the structure learning model than by MVT.

Model predictions & participant behavior Our model predicts that the order of volatility shifts in the environment will affect how patch categories are inferred and consequently, stay/leave decisions. When prior beliefs about environment structure and/or richness do not align with experience in the environment, the model infers more patches than there really are, leading to overharvesting. The pattern of experience in an initially predictable environment discourages inferring multiple patch types such that there is less of an influence of a prior bias towards complexity on foraging behavior.

Across the population, participants in both conditions overharvested roughly equally (t(73)=0.21 p = 0.83). We next examined the fit parameters to identify heterogeneity the population. Most participants were better fit with alpha as a free parameter (Figure 4a, t (74) = 2.90, p = 0.004). Participants in both conditions had a similar range of fit alpha parameters (Figure 4b, t(73) = -0.73, p = 0.47). However, matching the simulation results in Experiment 1, the inferred prior over environment richness differed between conditions. Participants in the high early volatility condition had lower prior estimates of environment richness or quality (Figure 4c, t(73) = -3.30, p = 0.001). Participants were split into high and low parameter groups (alpha and prior over environment richness) based the median value of the parameter. We looked for differences in overharvesting/underharvesting behavior between these groups. There were no differences in behavior between the high and low alpha group in either condition (Figure 4d, high early volatility - t(36) = 0.32, p = 0.75; low early volatility - t(34) = 0.54, p = 0.59). However, when splitting by prior over environment richness, those in the low group overharvested more than those in the high group in both conditions (Figure 4e, high early volatility - t(35) = -3.82, $p < 0.001$; low early volatility - t(35) = -4.105, $p < 0.001$).

Discussion

We asked if the process of learning the environment's structure could explain overharvesting behavior in certain contexts. To address this question, we developed a computational model of how foragers could learn environment structure and leverage it during decision making. First, in simulation, we showed that allowing the possibility of inferring multiple patch types results in overharvesting in highly variable single patch type environments. Next, we showed that our structure learning model could capture behavior in a previously collected stay/leave task. In this prior work, a model with error-driven learning of environment quality and a MVT decision rule was found to replicate participant's behavior. However, its success in fitting the data critically depended on a stay-leave bias parameter to account for overharvesting. Our model, on the other hand, provided a superior fit to participants' overharvesting relative to the MVT model without a stay-leave bias parameter. A possibility is that some of the variance explained by the stay-leave bias parameter emerged from the learning process formalized in our model. Finally, we tested a novel prediction of the structure learning model. Namely, that participant responses should be sensitive to the order of shifts in volatility. Participant behavior was consistent with this prediction, providing further evidence in favor of the model.

Taken together, these results suggest that seemingly suboptimal behavior like overharvesting can be explained with statistically optimal learning of environment structure and a prior expectation of heterogeneous environments. This is consistent with previous work demonstrating that people will infer structure or observe non-existent patterns even when there is no incentive to do so (Yu & Cohen, 2009) and even when it's disadvantageous (Collins & Frank, 2013; Gaissmaier & Schooler, 2008). This prior bias towards structure possibly emerges from it being frequently incentivized in the real world.

Potentially, MVT's definition of optimality may need to be expanded. In particular, foraging has been suggested to provide a decision context that we were evolutionarily adapted to and consequently, likely to yield normative behavior. However, MVT assumes an environment that does not concord with naturalistic environments which tend to be heterogeneous, non-stationary, and exhibit multiple scales of spatiotemporal regularities. Prior work demonstrates that foragers do consider this multi-scale information in adapting their search strategies in naturalistic settings (Fagan et al., 2013). Future work could explore extending the model to include multiple scales of reference points — one integrating over a longer time scale like MVT and another integrating over a shorter time scale as presented here. The present work and potential future work could suggest optimality in foraging may need to be redefined to incorporate dealing with the mul-

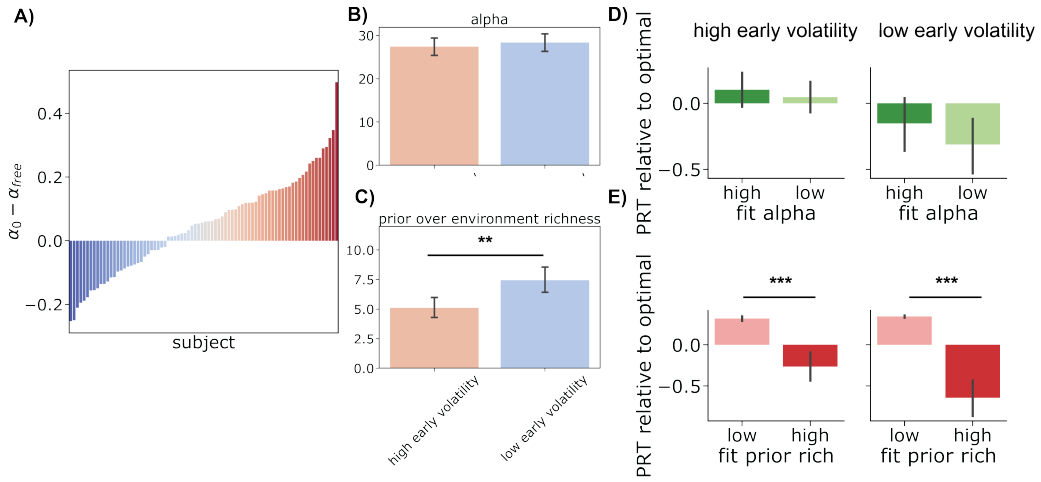

Figure 4: **Results from Experiment 3**. **A.** Each bar reflects the difference in cross-validation scores between the structure learning model with alpha fixed at 0 and the same model when alpha is a free parameter. Positive values indicate the structure learning model with free alpha provides a better fit to the participant's data. Overall, 49 out of 75 participants were better fit by the structure learning model with free alpha than the alpha fixed at 0 model. **B-C.**Participants' fit parameters for the structure learning model. **D-E.** Participants' overharvesting/underharvesting behavior separated by a median split on fit parameters from the structure learning model. Error bars are 95% CI.

tiple scales of uncertainty that natural environments present foragers with.

Acknowledgements

This work was supported by NIMH P50MH096889 to AMB. NCH was supported by a National Defense Science and Engineering Graduate fellowship. The authors thank Sara M Constantino for providing data for Experiment 2.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychol. Rev.*, *98*(3), 409–429.

Bergstra, J., & Bengio, Y. (2012). *Random search for hyper-parameter optimization.*

Charnov, E. L. (1976, April). Optimal foraging, the marginal value theorem. *Theor. Popul. Biol.*, *9*(2), 129–136.

Collins, A. G. E., & Frank, M. J. (2013, January). Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol. Rev.*, *120*(1), 190–229.

Constantino, S. M., & Daw, N. D. (2015, December). Learning the opportunity cost of time in a patch-foraging task. *Cogn. Affect. Behav. Neurosci.*, *15*(4), 837–853.

Fagan, W. F., Lewis, M. A., Auger-Méthé, M., Avgar, T., Benhamou, S., Breed, G., ... Mueller, T. (2013). Spatial memory and animal movement. *Ecol. Lett.*, *16*(10), 1316–1329.

Gaissmaier, W., & Schooler, L. J. (2008, December). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422.

Garrett, N., & Daw, N. D. (2020, July). Biased belief updating and suboptimal choice in foraging decisions. *Nat. Commun.*, *11*(1), 3417.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010, January). Context, learning, and extinction. *Psychol. Rev.*, *117*(1), 197–209.

Harhen, N. C., Hartley, C. A., & Bornstein, A. M. (2021). Model-based foraging using latent-cause inference. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*, to appear.

Kane, G. A., Bornstein, A. M., Shenhav, A., Wilson, R. C., Daw, N. D., & Cohen, J. D. (2019, September). Rats exhibit similar biases in foraging and intertemporal choice tasks. *Elife*, *8*.

Kilpatrick, Z. P., Davidson, J. D., & El Hady, A. (2021, April). *Uncertainty drives deviations in normative foraging decision strategies.*

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006, January). A more rational model of categorization.

Wikenheiser, A. M., Stephens, D. W., & Redish, A. D. (2013, May). Subjective costs drive overly patient foraging strategies in rats on an intertemporal foraging task. *Proc. Natl. Acad. Sci. U. S. A.*, *110*(20), 8308–8313.

Yu, A. J., & Cohen, J. D. (2008). Sequential effects: Superstition or rational behavior? *Adv. Neural Inf. Process. Syst.*, *21*, 1873–1880.