Model-based foraging using latent-cause inference

Nora C. Harhen (nharhen@uci.edu)
Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92697 USA

Catherine A. Hartley (cate@nyu.edu)
Department of Psychology, New York University, New York, NY 10003 USA

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

Abstract

Foraging has been suggested to provide a more ecologically-valid context for studying decision-making. However, the environments used in human foraging tasks fail to capture the structure of real world environments which contain multiple levels of spatio-temporal regularities. We ask if foragers detect these statistical regularities and use them to construct a model of the environment that guides their patch-leaving decisions. We propose a model of how foragers might accomplish this, and test its predictions in a foraging task with a structured environment that includes patches of varying quality and predictable transitions. Here, we show that human foraging decisions reflect ongoing, statistically-optimal structure learning. Participants modulated decisions based on the current and potential future context. From model fits to behavior, we can identify an individual's structure learning ability and relate it to decision strategy. These findings demonstrate the utility of leveraging model-based reinforcement learning to understand foraging behavior.

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

Introduction

Often, we have to choose between accepting a currently available option or expending effort in search of a potentially better alternative. Humans encounter serial stay-or-leave problems across many domains from searching for a job to searching for a romantic partner. These decisions, referred to as patch-leaving problems, are the same ones animals encounter when scouring their environment for resources.

Foraging tasks have been proposed to provide a more ecologically valid decision context than standard human decision-making tasks for understanding phenomena like intertemporal choice and the explore-exploit trade-off (Blanchard & Hayden, 2015; Mobbs et al., 2018). Marginal Value Theorem (MVT; Charnov, 1976) prescribes the optimal decision strategy, under certain assumptions, for patch leaving problems — leave the current patch once its reward rate drops below the global environment's reward rate. One simplifying assumption is that patches of symmetrically-varying quality are encountered at random. Meeting this assumption, most foraging tasks tested in humans have considered environments with patches of highly consistent quality. However, most natural environments are richly structured with multiple levels of spatio-temporal correlation which influence foragers' search strategies (McNamara & Houston, 1985; Sparrow, 1999; Kareiva & Tilman, 1998; Fagan et al., 2013).

One way to extend the human foraging literature is to develop tasks with structured environments to investigate how foragers leverage their knowledge of the environment to guide their decisions (Hall-McMaster & Luyckx, 2019). In these environments, MVT may not provide the best decision rule. Instead, model-based Reinforcement Learning (RL) may provide decision rules that better maximize reward (Kolling & Akam, 2017). In particular, it may be worthwhile to use a model of the environment to flexibly estimate the value of staying in the current patch versus the value of leaving in search of another. Because real-world environments are highly structured, foraging problems presented in contexts like these may be the ones that we are evolutionary adapted to solve and consequently, this set of decision rules may approximate more closely the evaluative processes that humans actually use.

Here, we ask: Do humans exploit the structure of the environment while foraging, and how do they learn this structure? We apply a model that borrows from rational models of categorization (Anderson, 1991; Sanborn, Griffiths, & Navarro, 2006, 2010) and latent cause theory (Gershman, Niv, & Blei, 2010) to explain how foragers learn a model of the environment, and we extend these models to explain how that knowledge is used to make patch-leaving decisions. To test the model's predictions, we developed a novel serial stay/leave task with multiple patch types of differing quality and a predictable transition structure between patch types.

Methods

Model

Multimodal Bayesian model We apply rational models of categorization (Anderson, 1991; Sanborn, Griffiths, & Navarro, 2006, 2010) to capture how foragers learn the latent structure of the environment. In this task, learning the environment's structure is equivalent to inferring which category a patch belongs to, while being uncertain of how many possible categories there are, and learning the transition probabilities between categories. The forager can leverage the model they've learned to decide when to leave the current patch.

We provide some intuitions for how the model works before formally describing it. The forager learns through combining their prior beliefs with the observed data. The forager's prior beliefs are structured around a set of assump-

tions about how the data they have observed is generated — namely, that rewards decay exponentially with each harvest, each patch belongs to one category, each category is characterized by a unique distribution over decay rates, a new patch is more likely to belong to a common category (i.e. relatively more category members), and there is some small probability that a new patch belongs to a new category. Bayes' rule stipulates how prior beliefs and observed data are combined to give a posterior distribution over possible assignments, or groupings, of patches. This posterior is used to predict the decay rate in the current patch if the forager were to stay and harvest again. To make the decision to stay or leave, the forager compares the reward they would expect to receive by harvesting this patch with the expected reward from traveling to and harvesting another patch.

Generative model. The prior probability of a patch belonging to a category, k, at time t is given by

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

Where N_k is the number of patches already assigned to that category and α is the prior over category dispersion (i.e. how densely or sparsely distributed are patches over categories). This formally instantiates the assumptions that a patch is more likely to belong to a "popular" category and there is some probability that a patch will belong to a new, previously unobserved category.

Each category is associated with its own normal distribution over decay rates, parameterized by μ_k and σ_k^2. When a new patch is assigned to a cluster, the decay rates observed in that patch update the cluster-specific distribution.

We can derive from Equation 1 the prior probability of a particular assignment of patches to categories at time t, c_t. This is with the assumption that categories are assigned one after another.

$$P(c_t) = \frac{\alpha^s}{\prod_{t=0}^{T-1}[\alpha+t]} \prod_{k=1}^{s}(N_k-1)! \quad (2)$$

Inference model. A set of observed decay rates at time t, D_t, can then be combined with the prior probability specified in Equation 2 to generate a posterior distribution over category assignments

$$P(c_t|D_t) = \frac{P(D_t|c_t)P(c_t)}{p(D_t)} \quad (3)$$

Exact computation of this posterior is computationally demanding, so we use particle filtering as an approximate inference algorithm (Gershman, Niv, & Blei, 2010; Sanborn, Griffiths, & Navarro, 2006, 2010). The posterior is approximated with a set of m particles. Each particle, $c_t^{(l)}$ represents a particular grouping of patches into categories from the first t trials. A grouping is represented in the set approximately proportional to its posterior probability. Summing over the particles gives an approximation to the posterior distribution

$$P(c_t|D_t) \approx \frac{1}{m}\sum_{l=1}^{m}\delta(c_t - c_t^{(l)}) \quad (4)$$

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

We can approximate the prior distribution over groupings for the first $t+1$ trials with

$$\begin{aligned} P(c_{t+1}|D_t) &= \sum_{c_t} P(c_{t+1}|c_t)P(c_t|D_t) \\ &\approx \sum_{c_t} P(c_{t+1}|c_t)\frac{1}{m}\sum_{l=1}^{m}\delta(c_t - c_t^{(l)}) \\ &= \frac{1}{m}\sum_{l=1}^{m} P(c_{t+1}|c_{t+1}^{(l)}) \end{aligned} \quad (5)$$

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

$$\begin{aligned} P(c_{t+1}|D_{t+1}) &\propto \sum_{c_t} P(d_{t+1}|c_{t+1}, D_t)P(c_{t+1}|D_t) \\ &\approx \frac{1}{m}\sum_{l=1}^{m} P(d_{t+1}|c_{t+1}, D_t)P(c_{t+1}|c_{t+1}^{(l)}) \end{aligned} \quad (6)$$

We draw m samples from this distribution to create a new set of particles.

5 particles were used for all simulations. A smaller number of particles allows psychological plausibility and can capture the variability and order sensitivity people display (Sanborn, Griffiths, & Navarro, 2006, 2010).

Prediction To generate the agent's prediction of the upcoming decay rate, data is sampled in the way it is assumed to be generated. A category is drawn with probability proportional to the posterior and then a decay rate is drawn from that category's characteristic distribution. This process is repeated 1000 times and the samples are averaged over.

Unimodal Bayesian model Under this model, the forager assumes that all patches belong to the same cluster, and all decay rates are sampled from a unimodal normal distribution parameterized by μ and σ^2. This is equivalent to the multimodal model when $\alpha = 0$.

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 decay 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 timescale reference point than MVT's.

The forager compares v_{stay} and v_{leave} and always takes the higher value action. In the unimodal model, s will always be 1. Thus, the point estimate of the leave threshold, v_{leave}, will be equivalent to MVT's leave threshold, the total average reward rate across the environment.

Model fitting

We fit individual participant's data to the models with α, the category dispersion, and prior over environment richness as free parameters for the multimodal model and prior over richness as a free parameter for the unimodal model. 500 sets of parameters were sampled from a Sobol Sequence. 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). To pick amongst these candidate parameter sets, we compared the participant's PRT relative to the MVT-optimal policy for each planet type to the same measures generated by the model. The best fitting parameters were those that minimized mean squared error (MSE) between the participant's data and the model simulation's. Error was computed as Euclidean distance.

To compare models, we used cross validation. We held out one test block and then fit the model using the average 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.

Task

Participants We recruited 40 participants from Amazon Mechanical Turk (ages 19-75, Mean=42, SD=13.5). Participation was restricted to workers who had completed at least 100 prior studies and had at least a 99% approval rate. Participants were paid $6 as a base payment and could earn a bonus contingent on performance ($0-$4). We excluded 3 participants for having average patch residence times 2 standard deviations above or below the group mean.

Procedure We investigated how humans learn in a serial stay/switch foraging task (Constantino & Daw, 2015) in which participants decide between staying to harvest a depleting patch of resources or incurring a time delay to switch to a replenished patch (Figure 1a). The task was framed as a space mining game in which participants were told to collect as many space gems as possible because gems would be converted into their bonus payment at the end of the experiment. They were given approximately 0.001 cent per gem. However, they were not told this exchange rate, and the total amount of gems collected was not displayed. On each trial, participants visited a planet and had to decide, via keypress, if they wanted to dig on the current planet ('A') or travel to a new planet ('L'). If they decided to dig, they would be shown the number of gems collected after watching a short animation of an astronaut digging for 2 seconds. If they instead decided to travel, they watched a longer animation of a flying rocket ship for 10 seconds followed by an image of an alien for 5 seconds prior to arriving at the new planet. If they did not make a decision within 2 seconds, a warning was displayed and they had to wait 2 seconds before making another response. To ensure participants' reaction times did not affect their reward rate, the reaction time (RT) was subtracted from the following dig or travel animation display time.

Participants completed 5 blocks lasting 6 minutes each. The foraging environment consisted of three planet types characterized by quality (e.g. poor, neutral, & rich). Planet quality was determined by the distribution its decay rates were sampled from (Figure 1b). Rewards depleted the slowest on rich planets — decay rates were sampled from a Beta distribution with parameters a = 52 and b = 22. These parameters were chosen such that the mean decay rate was 0.8 with a SD of 0.05. Rewards depleted quicker on neutral planets. Decay rates were sampled from a Beta distribution with parameters a = 44, b = 44 (Mean = 0.5, SD = 0.05) and the rewards on poor planets depleted the quickest with parameters a = 22 , b = 51 (Mean = 0.2, SD = 0.05). Planets of the same type were encountered in "clusters" or "galaxies" (Figure 1c). When the participant left the current planet, there was an 80% probability they would travel to a planet of the same quality. If the participant traveled to a planet of a different quality, it was equally likely to be one of the two remaining planet types (10% of going to different planet type 1 and 10% of going to different planet type 2). Participants were not told that planets varied in quality, nor when transitions to new galaxies would occur, requiring them to infer this information from experience alone.

Model Predictions

Our model simulations predict distinct patterns of foraging behavior dependent on α, the prior over category dispersion, and the prior over environment richness (Figure 2). The unimodal learner's prior belief about environment richness does not affect foraging behavior. They overharvest relative to MVT-optimal on poor and neutral planets, but underharvest on rich planets. Because they are averaging over rewards received from all previous planets, the influence of prior beliefs quickly wanes as more data (rewards) are observed (Figure 3a). In contrast, the multimodal learner's prior belief about environment richness does affect their behavior. Inferring multiple modes acts as regularization towards the prior because the prior is only updated with observations assigned to the current category or mode (Figure 3b). With a prior belief that rewards will be sparse in the environment (poor), the multimodal learner overharvests on all 3 planet types, particularly on the rich planets. With a prior belief that rewards will be abundant in the environment (rich), the learner is less likely to overexploit relative to MVT. They only overharvest on poor planets and act MVT-optimally on neutral and rich planets. The unimodal learner's pattern of over- and underharvesting can be explained by the averaging over all observed decay rates. This results in an average that falls in between the true average decay rates on the neutral and rich planets (Figure 3c). The multimodal learner allows for the possibility of multiple patch types in the environment and, thus, with this added flexibility, generates more accurate decay rate predictions (Figure 3d).

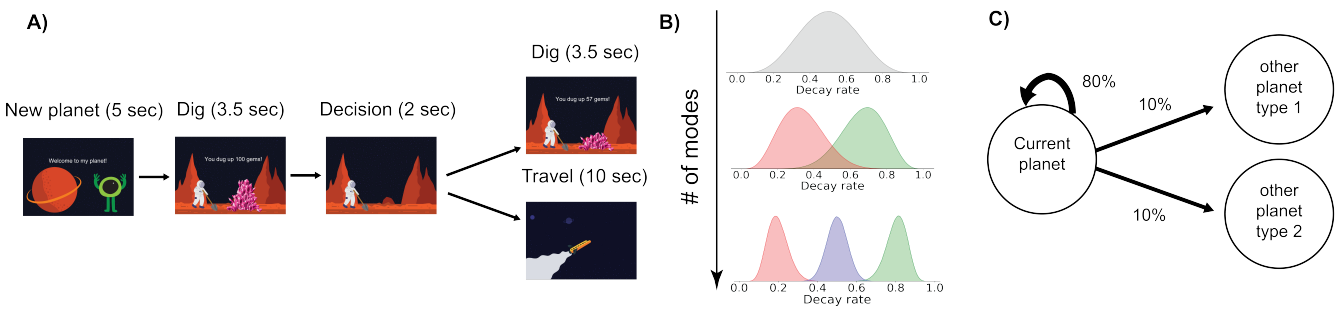

Figure 1: **A.** Task structure. Participants sequentially decide whether to dig or travel to a new planet. **B.** Example decay rate distributions the participant could infer. The bottom distribution is the true tri-modal decay rate distribution. **C.** Transition probabilities between planets. Participants were most likely to travel to a planet belonging to the same category as the current one. There was an equal probability of traveling to patch belonging to one of the remaining two categories.

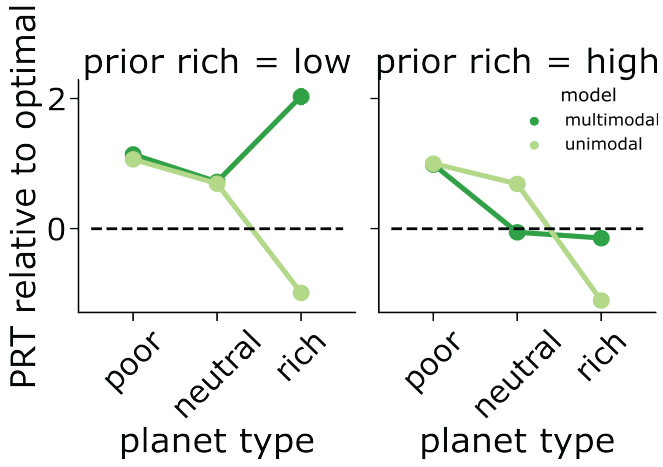

Figure 2: Model predictions. Overharvesting and underharvesting behavior depends on whether the learner assumes there is a single patch type in the environment (unimodal) or allows for the possibility of multiple (multimodal) and their prior belief over the density of rewards in the environment (environment richness).

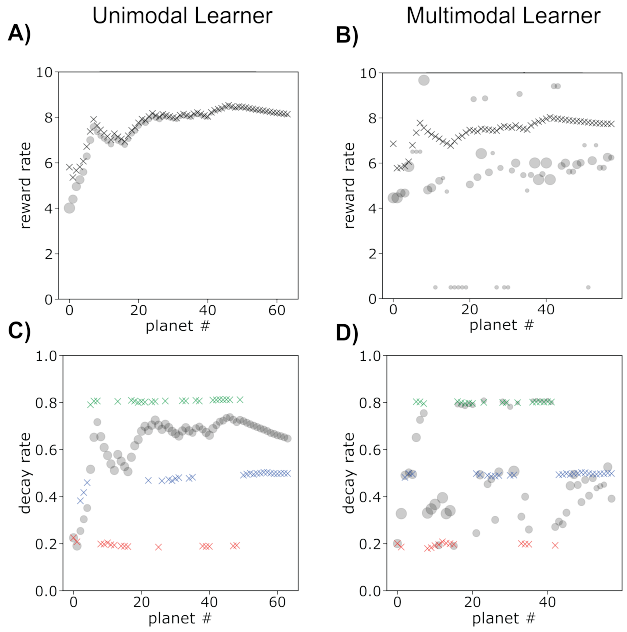

Figure 3: Simulated learner's (multimodal and unimodal) predicted decay rates on the current planet and predicted reward rate of the next planet (grey circles, size is proportional to variance of estimate). The multimodal learner's predictions more closely align with true decay rate (colored Xs; red - poor planet, blue - neutral, green - rich) of the current planet relative to the unimodal learner's. The unimodal learner's reference point for leaving the current patch more closely follows the MVT-optimal reference point (black Xs) compared to the multimodal learner

Results

Behavioral We compared participants' planet residence times (PRT) to the policy of an MVT-optimal agent. The optimal agent knows which type of planet they are currently on, and the parameters of each planet type's true decay rate distribution. They take the expected value of the true decay rate distribution to predict how much reward will be earned on the next dig. The optimal agent leaves if the predicted reward falls below the opportunity cost of the time spent harvesting it. Averaging over all participants and all planet types, participants overharvested, or stayed too long, relative to optimal (Figure 4a; $t(36) = 4.61$, $p < 0.0001$). However, when separating out PRTs by planet type, participants only overharvested on poor and neutral planets (Figure 4b; poor - $t(36) = 24.29$, $p < 0.0001$; neutral - $t(36) = 9.78$, $p < 0.0001$; rich - $t(36) = 0.58$, $p = 0.57$).

To asses the effect of potential transitions to other galaxies on current stay/leave decisions, each subject's experienced transition probabilities between planet types and their PRT was correlated (Spearman rank) (Figure 5). For statistical tests, each subject's correlation coefficient, ρ, was transformed with the inverse hyperbolic tangent function (arctanh) to make the data normally distributed. One sample t-tests were performed on the transformed correlation coefficients. Participants who observed a greater probability of transitioning from a poor to rich planet left poor planets earlier ($t(36) = -2.20$, $p = 0.03$).If they previously observed a greater probability of transitioning to another neutral planet, they left the current neutral planet earlier ($t(36) = -2.04$, $p = 0.049$). In contrast, on rich planets, participants who observed a higher probability of transitioning to another rich planet stayed longer ($t(36) = 3.05$, $p = 0.004$) and those who observed a higher probability of transitioning to a neutral planet left sooner ($t(36) = -5.21$, $p < 0.0001$). Taken together, these results are consistent with participants learning to distinguish between patches of varying quality, tracking the transition structure between them, and using this the knowledge to inform their decisions.

Individual differences in model fits On rich planets, participants displayed considerable variability in their PRTs relative to MVT-optimal. Our model predicts distinct patterns of behavior arising from an agent's prior over category dispersion, and the prior over richness of the overall environment (Figure 2).As a group, the multimodal learner model produced a similar fit to participants' data than the unimodal learner (Figure 6a; one sample t-test on difference, $t(36) = -1.05$, $p = 0.30$). Examining individual subjects, there were some who were much better fit by the multimodal model, others much better fit by the unimodal model, while the majority were slightly better fit by the unimodal model (Figure 6b). Whether a participant is better fit by a unimodal or a multimodal learner could be used as an individual difference metric reflecting either the participant's prior over environment complexity or their structure learning ability. The interpretation of this hinges on whether or not participants' better fit model varies based on the true structure of the environment. Of course, allowing the possibility of inferring multiple modes will be useful in multimodal environments like in the present task, but may be less so in unimodal environments. Regardless of the interpretation, our results confirm that the persistent observation of overharvesting, relative to MVT, obscures strategic decision variability that can be attributed to statistically optimal learning of environmental structure.

Discussion

We asked whether humans could learn the latent structure of an environment and exploit it while foraging. We found that that participants' stay/leave decisions were sensitive to the quality of the current planet, suggesting that they learned a multi-state representation and used it to guide their decisions. However, our model fitting results revealed that participants' learned representations varied, with some learning multiple planet types (modes) and others only learning one.

An appropriate representation of the task's environment includes not only the different planet types but the transitions between them. So, we asked if they adjusted their decisions based on potential future contexts and past contexts. Our results suggest that they did — participants' PRTs were related to their experienced transition probabilities between planet types and to the preceding planet type in certain contexts.

Deploying a model-based strategy requires both inferring an appropriate model of the environment and navigating over that internal model to construct a plan. The canonical measure of whether behavior reflects model-based planning, the two-step task (Daw et al., 2011), primarily reflects the latter (Konovalov & Krajbich, 2020). We've presented a novel foraging task that provides a measure of individuals' structure learning ability.

A potential direction for future work is to explore how incorrect inference of the environment's structure may explain overharvesting. Proposed mechanisms for overharvesting include subjective costs to leaving the patch (Wikenheiser et al., 2013), decreasing marginal utility (Constantino & Daw, 2015), and discounted future rewards (Kane et al., 2019). However, another way such biases could emerge is though learning (Niv et al., 2002; Garrett & Daw, 2020). Other proposed mechanisms may even be subsumed under these learning biases.

Acknowledgements

This work was supported by NIMH P50MH096889 to AMB. NCH was supported by a National Defense Science and Engineering Graduate fellowship.

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.*

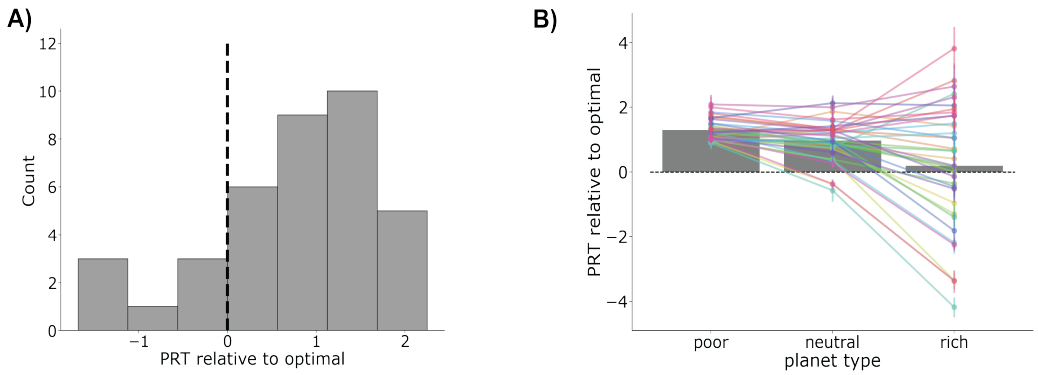

Figure 4: Behavioral results. Error bars are S.E.M. **A. Overall effect** Participants' average PRT relative to the MVT-optimal policy (dashed line) averaged across all planet types. A majority of participants overharvested across the experiment. **B. Effect of current context**. Individual participant's average PRT relative to MVT-optimal (dashed line) for each planet type (grey bars are the average across all participants). On poor and neutral planets, most participants overharvested. However, on rich planets, participants displayed considerable variability in their over/underharvesting.

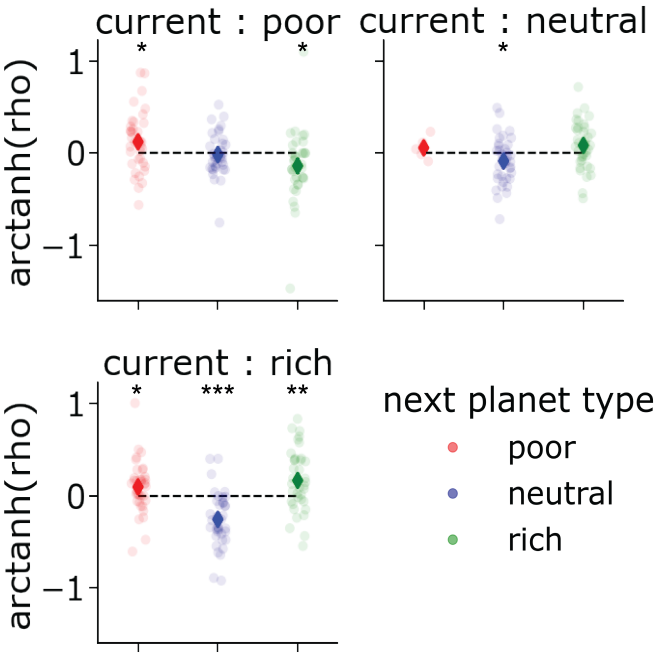

Figure 5: Behavioral results.Effect of future potential context Each subject's inverse hyperbolic tangent transformed correlation coefficient (between transition probability and PRT). Participants stay/leave decisions were modulated by the transition probabilities between planet types.x

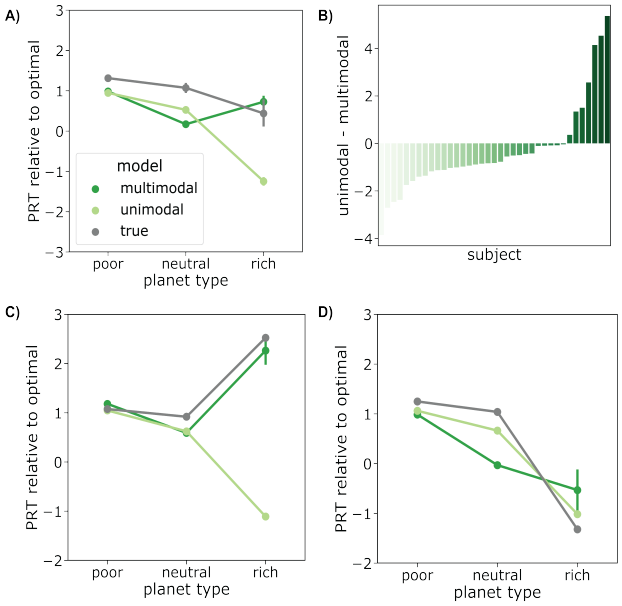

Figure 6: Model fitting results. Error bars are S.E.M. **A.** Comparison of participant data with the simulated data from the multimodal and unimodal models. **B.** Comparison of cross validation scores for each subject. Positive values indicate the participant's data was better fit with the multimodal model. **C.** One example participant best fit with the multimodal model. **D.** One example participant best fit with the unimodal model.

Blanchard, T. C., & Hayden, B. Y. (2015). Monkeys are more patient in a foraging task than in a standard intertemporal choice task. *PLoS One*, *10*(2), e0117057.

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

Constantino, S. M., & Daw, N. D. (2015). 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.

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

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

Hall-McMaster, S., & Luyckx, F. (2019, April). Revisiting foraging approaches in neuroscience. *Cogn. Affect. Behav. Neurosci.*, *19*(2), 225–230.

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*.

Kareiva, P. M., & David Tilman, G. (1997). Spatial ecology : the role of space in population dynamics and interspecific interactions. In *Published in **1997** in princeton NJ) by princeton university press.* Princeton (N.J.) : Princeton university press, 1997.

Kolling, N., & Akam, T. (2017). (reinforcement?) learning to forage optimally. *Curr. Opin. Neurobiol.*, *46*, 162–169.

Konovalov, A., & Krajbich, I. (2020, April). Mouse tracking reveals structure knowledge in the absence of model-based choice. *Nat. Commun.*, *11*(1), 1893.

McNamara, J. M., & Houston, A. I. (1985). Optimal foraging and learning. *J. Theor. Biol.*, *117*(2), 231–249.

Mobbs, D., Trimmer, P. C., Blumstein, D. T., & Dayan, P. (2018). Foraging for foundations in decision neuroscience: insights from ethology. *Nat. Rev. Neurosci.*, *19*(7), 419–427.

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adapt. Behav.*, *10*(1), 5–24.

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

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.*, *117*(4), 1144–1167.

Sparrow, A. D. (1999). A heterogeneity of heterogeneities. *Trends Ecol. Evol.*, *14*(11), 422–423.