Humans adapt their foraging strategies and computations to environment complexity

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

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

Abstract

Foraging has been suggested to provide a naturalistic context for studying decision-making. In the wild and in the laboratory, foragers come close to approximating the optimal decision strategy given by Marginal Value Theorem (MVT; Charnov, 1976). Recent work has used reinforcement learning to understand how the variables for decision-making under an MVT-policy are learned (Garrett & Daw, 2020; Simon & Daw, 2011). This work often implicitly assumes the forager begins with a specific, fixed representation of the environment. However, it is likely that this representation is also something that must be learned. Here we ask — can foragers learn a representation of the environment and, importantly, do they adapt their decision strategies and value computations to this representation as it evolves? We propose a model of how foragers could use principled statistical inference to organize their past experiences into a representation that guides decision-making. The model was tested in a variant of a serial stay/switch foraging task with multimodal reward distributions and non-uniform transition structure between patch types. In this task, participants adapted their foraging to both the richness of the local context and their internal uncertainty. These results are consistent with participants having learned and used a model of the environment to guide their decisions. Overall, these findings demonstrate the utility of combining representation learning and reinforcement learning to understand foraging behavior.

Keywords: foraging, structure learning, model-based reinforcement learning

Acknowledgements

This work was supported by NIMH P50MH096889 and a NARSAD Young Investigator Award by the Brain and Behavior Research Foundation to AMB. NCH was supported by a National Defense Science and Engineering Graduate fellowship.

1 Introduction

Decision-makers commonly choose between staying with a current option or foregoing it in hope of a better future alternative. Such decisions arise in ethology where they are known as patch leaving problems in which a patch refers to a concentration of resources in the environment. In solving these problems, foragers must weigh the the costs and benefits of harvesting an often depleting resource against those associated with searching for a new, unharvested one. An optimal solution is given by Marginal Value Theorem under a certain set of assumptions (MVT; Charnov, 1976) — a forager should leave the current depleting patch once its reward rate falls below the overall reward rate of the environment. Relative to MVT, it's been widely observed that foragers from rodents to humans stay longer than prescribed (Blanchard & Hayden, 2015; Constantino & Daw, 2015; Kane et al., 2019). This is known as overharvesting.

MVT's predictions assume the forager has complete knowledge of the environment and its dynamics. Consequently, MVT provides how the decision should be made, but it does not explicitly describe how its key decision variables, the local and global reward rates, should be learned. However, this assumption of complete knowledge is not often met in the real world. This suggests more naturalistic patching leaving is both a decision-making and a learning problem. A potential simple learning rule involves keeping a running average of rewards across all past patch experiences in the environment (Constantino & Daw, 2015).

These simple learning rules work well in simple, homogeneous environments in which patches are similar to one another. However, real world environments are often complex with regions varying in richness (McNamara & Houston, 1985; Sparrow, 1999). In more naturalistic environments, it may be beneficial to group patches of similar richness together to form a multi-state representation that affords contextually-appropriate and dynamic estimates of the local and global reward rates. In standard reinforcement settings, humans use a similar strategy — they track environmental statistics and leverage them to adaptively adjust reward-related computations (Behrens, Woolrich, Walton, & Rushworth, 2007; Simon & Daw, 2011). Thus, we asked — in a foraging context, do decision-makers learn an internal representation of the environment and adapt their strategies and computations with respect to it? We propose a model of how foragers may incrementally build such a representation from past experiences and use it during decision-making. We then test its predictions with a novel serial stay-switch task.

2 Latent Cause Model

2.1 Learning a state representation of the environment

Latent-cause inference provides a framework for building state space representations out of past experiences (Courville, Daw, & Touretzky, 2006; Gershman, Norman, & Niv, 2015). Under this framework, representation learning is treated as a clustering problem in which an experience is assigned to a pre-existing cluster based on its similarity to experiences previously assigned to the cluster. In a new environment, the learner begins with a single cluster, or state, to which experiences can belong. New experiences that differ significantly from past ones can initiate the creation of a new cluster. Thus, the complexity of the representation is allowed to grow incrementally as experience warrants it. Through principled statistical learning, the learner develops a representation with a useful number of clusters — enough to allow for contextually-specific predictions but few enough to enable generalization across experiences.

Within a foraging context, past experiences could correspond to reward decays experienced while harvesting patches and states to patch types that differ in their richness. To infer the current patch type or state, the observer must combine their past experiences with current experience. The prior probability of a patch belonging to a patch type, p, at time t is given by:

$$P(p) = \begin{cases} \frac{N_p}{t-1+} & \text{if p is an old patch type} \\ \frac{}{t-1+} & \text{if p is a new patch type} \end{cases} \tag{1}$$

Where N_p is the number of patches already assigned to that patch type and is the prior over environment complexity. This formally instantiates the assumptions that 1) the current patch type is more likely to be a frequently visited one and 2) there always remains some probability that a new patch belongs to a previously unobserved patch type. In our model, we allow the structure learning parameter, , to be a free parameter fit to individual participants' choice data.

A set of reward decays at time t, D_t, can be combined with the prior probability specified in Equation 1 to generate a posterior distribution over patch types.

$$P(p_t j D_t) = \frac{P(D_t j p_t) P(p_t)}{p(D_t)} \tag{2}$$

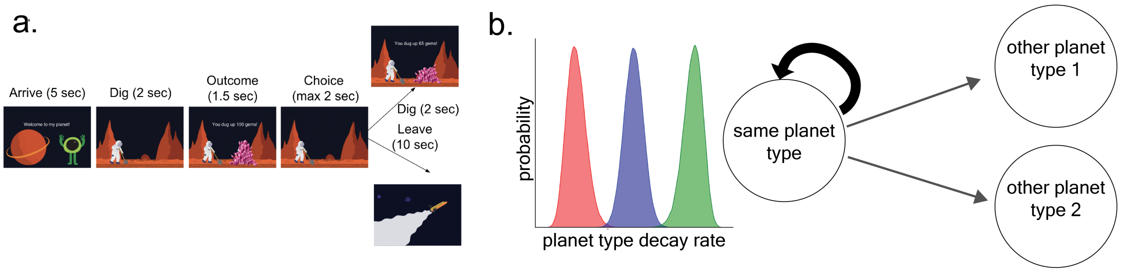

Figure 1: Task structure. **A.** Participants sequentially decided whether to stay dig from a depleting gem mine or incurring a time cost to leave for a new planet with a replenished mine. **B.** The decay rate distributions associated with rich, neutral, and poor planet types and and transition probabilities between planet types when leaving for a new planet.

where p_t is a potential patch type. Each patch type has a unique distribution over decay rates associated with it that determines $P(D_t j p_t)$.

Exact computation of this posterior is computationally demanding, so we use particle filtering as an approximate inference algorithm (Gershman et al., 2015; Sanborn, Griffiths, & Navarro, 2006). Harhen, Hartley, and Bornstein (2021) contains further implementation details.

2.2 Using structure to inform stay/leave decisions

The forager compares the value of staying and leaving and selects the higher-valued option. The value of staying is taken as the reward received on the last harvest, r_t, multiplied by the predicted decay rate, $\hat{d}$, if the forager were to stay again.

$$V_{stay} = r_t \quad \hat{d} \tag{3}$$

To generate $\hat{d}$ the agent samples from patch type-specific decay rate distributions. The probability of sampling from a planet type is proportional to its posterior probability of the current planet belonging to that cluster, $P(p_t j D_t)$.

The value of leaving is estimated by averaging over the reward rates from all previously encountered patches discounted by some factor, .

$$V_{leave} = \frac{r_{total}}{t_{total}} t_{harvest} \tag{4}$$

Theoretical work has suggested that discounting factors that adapt to an agent's internal uncertainty can be beneficial in complex environments (Jiang, Kulesza, Singh, & Lewis, 2015). Following this, we allow to be dynamic, flexibly adjusting to the individual's uncertainty over the accuracy of their internal representation. Here, we compute uncertainty as being proportional to the entropy of the samples drawn to generate, $\hat{d}$.

We compared this structure learning model to two models previously used to explain human foraging behavior in Constantino & Daw (2015) – a temporal difference learning model (TD) and a MVT learning model that learns the mean decay rate and global reward rate of the environment (MVT learn). Each model's fit to the data was evaluated using a 10-fold cross validation procedure. For each participant, we shuffled their PRTs on all visited planets and split them into 10 separate training/test datasets. The best fitting parameters were those that minimized the sum of squared error (SSE) between the participant's PRT and the model's predicted PRT on each planet in the training set. Then, with the held out test dataset, the model was simulated with the best fitting parameters and the SSE was calculated between the participant's true PRT and the model's PRT. To compute the model's final cross validation score, we summed over the test SSE from each fold.

3 Methods

We tested whether participants could learn an internal representation of a three patch type environment and use it to guide their foraging decisions. Past human foraging work has focused on either single patch type environments or multi-patch type environments in which patch types of differing richness are blocked off from one another. To more closely mimic real world conditions, we interleaved the three patch types and did not indicate that patches could differ from one another.

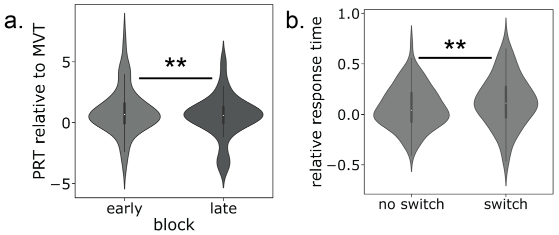

Figure 2: Behavioral results. **A.** With experience, participant's planet residence times (PRT) became closer to those prescribed by an MVT-optimal policy (at 0). **B.** Participants were slower in making their first decision on a new planet following a switch to a different planet type relative to if there was no switch.

Specifically, we investigated how humans learn in a serial stay/switch foraging task (Constantino & Daw, 2015). Participants visited planets where they would dig for space treasure (Figure 1A). On a planet, participants had to choose between staying and digging from a depleting mine or incurring a time cost to travel to a new planet with a replenished mine. Their goal was to collect as many gems as possible across the span of the game. Participants completed 5 blocks lasting 6 minutes each. The environment consisted of three planet types differing in richness – rich, neutral, and poor (Figure 1B). Rich planets had decay rates sampled from a distribution with a higher expected value and hence would deplete more slowly relative to the poor and neutral planets. Participants were not told that planets differed in quality requiring them to infer this from experienced rewards alone. Planets of a similar type temporally clustered together resembling natural environments' spatiotemporal correlations in richness. When the participant left a planet, there was a 80% probability they would travel to a planet of the same type. If traveling to a planet of differing quality, it was equally likely to be one of the two remaining planet types.

We recruited 198 participants from Amazon Mechanical Turk (ages 23-64, Mean=39.79, SD=10.56). 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 82 participants for having average planet residence times 2 standard deviations above or below the group mean, failing a quiz on the task instructions more than 2 times, and/or missing catch questions. The catch questions asked participant's to press the letter "Z" on their keyboard at random intervals throughout the task. This was meant to "catch" participants who were repeatedly making repetitive choices in a manner not guided by value.

4 Results

As predicted by MVT, participants stayed longer the richer the planet was (rich vs. neutral - t(115)= 19.77, $p < 0:0001$; neutral vs. poor - t(115) = 12.57, $p < 0:0001$). When directly comparing to MVT, participants on average overharvested across the entire experiment (t(115) = 3.88, $p = 0:00018$). MVT assumes perfect knowledge of the environment's structure. If participants learned the structure of the task environment, then the extent of over harvesting should diminish as they accumulate more experience. Participants did just that, overharvesting more in the initial two blocks relative to the final two (Figure 2A, t(115) = 3.27, $p = 0:0014$). Further suggesting a multi-state representation, participants demonstrated context sensitivity, overharvesting only on poor and neutral planets but not on rich (Figure 2d; poor - t(115) = 6.92, $p < 0:0001$; neutral - t(115) = 9.00, $p < 0:0001$; rich - t(115) = 1.38, $p = 0:17$). Participants' reaction times provided further evidence of structure learning. We reasoned that learners sensitive to structure should demonstrate switch costs when transitioning between planets of a different type. Consistent with this, participants were slower in making their initial choice on planets who differed in type from the most recent prior planet (Figure 2B, t(115) = 2.65,$p = 0:0093$).

Computational modeling results further supported participants' use of an internal model of the environment. Based on cross validation scores, the adaptive discounting model provided a better account of participants' choices relative to the two other models that assumed no structure learning (Figure 2AB). In the adaptive discounting model, the structure learning parameter must be greater than 0 to allow for multi-state inference. In simulation, the lowest setting of that lead to multiple states being inferred in at least 90% of simulation runs was $0:8$. Thus, this value was used as our baseline for assessing structure learning. We found that 76% of participants had a fit greater than than this threshold (Figure 3C). Validating as a measure of individual structure learning ability, participant's with higher fit demonstrated greater switch costs (Figure 3D, Kendall's = 0.24, $p = 0:00076$).

Prior theoretical work has demonstrated that decision-makers should monitor the uncertainty over the accuracy of their model of the environment and adapt their planning horizon to it (Jiang et al., 2015). Based on this normative work and empirical findings that humans similarly adapt their discounting of future value to internal uncertainty (Gershman & Bhui, 2020), we reasoned that participants' choices in this more complex, naturalistic environment environment would