Overharvesting in human patch foraging reflects rational structure learning

Nora C. Harhen[1] and Aaron M. Bornstein[1,2]

[1]Department of Cognitive Sciences, University of California, Irvine.
[2]Center for the Neurobiology of Learning and Memory, University of California, Irvine.

Abstract

Patch foraging presents a sequential decision-making problem widely studied across organisms — stay with a current option or leave it in search of a better alternative? Behavioral ecology has identified an optimal strategy for these decisions, but, across species, foragers systematically deviate from it, staying too long with an option or "overharvesting", relative to this optimum. Despite the ubiquity of this behavior, the mechanism underlying it remains unclear. Here, we address this gap, by approaching foraging as both a decision-making and learning problem. Specifically, we propose a model in which foragers 1) rationally infer the structure their environment and 2) use their uncertainty over the inferred structure representation to adaptively discount future rewards. We find that overharvesting can emerge from this rational statistical inference and uncertainty adaptation process. In a patch leaving task, we show that human participants adapt their foraging to the richness and dynamics of the environment in ways consistent with our model. These findings suggest that definitions of optimal foraging could be extended by considering how foragers reduce and adapt to uncertainty over representations of their environment.

Introduction

Many real world decisions are sequential in nature. Rather than selecting from a set of known options, a decision-maker must choose between accepting a current option or rejecting it for a potentially better future alternative. Such decisions

arise in a variety of contexts including choosing an apartment to rent, a job to accept, or a website to browse. In ethology, these decisions are known as patch leaving problems. Optimal foraging theory suggests that the current option should be compared to the quality of the overall environment. An agent using the optimal choice rule given by Marginal Value Theorem (MVT[1]) will leave once the local reward rate of the current patch, or concentration of resources, drops below the global reward rate of the environment.

Foragers largely abide by the qualitative predictions of MVT, but deviate quantitatively in systematic ways - staying longer in a patch relative to MVT's prescription. Known as overharvesting, this bias to overstay is widely observed across organisms. Despite this, how and why it occurs remains unclear. Proposed mechanisms include a sensitivity to sunk costs[2,3], diminishing marginal utility[4], discounting of future rewards[3–5], and underestimation of post-reward delays[6]. Critically, these all share MVT's assumption that the forager has accurate and complete knowledge of their environment, implying that deviations from MVT optimality emerge in spite of this knowledge. However, an assumption of accurate and complete knowledge often fails to be met in dynamic real world environments[7]. Relaxing this assumption, how might foragers learn the quality of the local and global environment?

Previously proposed learning rules include recency-weighted averaging over all previous experiences[4,8] and Bayesian updating[9]. In this prior work, learning of environment *quality* is foregrounded while knowledge of environment *structure* is assumed. In a homogeneous environment, as is nearly universally employed in these experiments, this is a reasonable assumption as a single experience in a patch can be broadly generalized from across other patches. However, it may be less reasonable in more naturalistic heterogeneous environments with regional variation in richness. To make accurate predictions within a local patch, the forager must learn the heterogeneous structure of the broader environment. How might they rationally do so?

In standard economic choice tasks, humans have been shown to act in accordance with rational statistical inference of environment structure. Furthermore, by assuming humans must learn the structure of their environment from experience, seemingly suboptimal behaviors can be rationalizing including prolonged exploration[10], melioration[11], and overgeneralization[12]. By building on this proposal and extending it to a foraging context, we suggest that overharvesting, another seemingly suboptimal behavior, could be a byproduct of a rational agent inferring the latent structure of their environment and, critically, adapting their decision computations to the qualities of this inferred structure.

We formalize this suggestion using an infinite capacity mixture model[13,14] and test its predictions with a novel variant of a serial stay-switch task (Fig. 1A; Constantino and Daw[4], Decker et al.[15]). Participants visited different planets to mine for "space treasure" and were tasked to collect as much space treasure as possible over the course of a fixed length game. On each trial, they had to decide between staying on the current planet to dig from a depleting treasure mine or traveling to a new planet with a replenished mine at the cost of a time delay. To

mimic naturalistic environments, we varied planet richness across the broader environment while locally correlating richness in time. More concretely, planet richness was drawn from a multimodal distribution (Fig. 1B) and transitions between planets of a similar richness were more likely (Fig. 1C). Our model predicted distinct behavioral patterns from structure learning individuals versus their non-structure learning counterparts in our task. Specifically, within the multimodal environment, non-structure learners are predicted to underharvest on average, while structure learners overharvest. Furthermore, structure learners' extent of overharvesting are predicted to vary across the task — decreasing with experience and increasing following rare transitions between planets. In contrast, non-structure learners should consistently underharvest.

We found that principled inference of environment structure and adaptation to this structure can produce key deviations from MVT that have been widely observed in participant data across species. Taken together, these results reinterpret overharvesting: Rather than reflecting irrational choice under a fixed representation of the environment, it can be seen as rational choice under a dynamic representation.

Methods

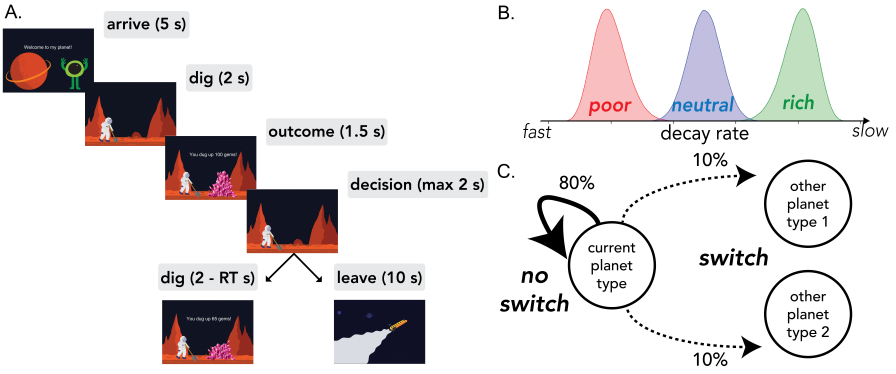

Fig. 1 A. Serial stay-switch task. Participants traveled to different planets and mined for space gems across 5 6-minute blocks. On each trial, they had to decide between staying to dig from a depleting gem mine or incurring a time cost to travel to a new planet. **B. Environment structure.** Planets varied in their richness or, more specifically, the rate at which they exponentially decayed with each dig. There were three planet types — poor, neutral, and rich — each with their own characteristic distribution over decay rates. **C. Environment dynamics.** Planets of a similar type clustered together. A new planet had an 80% probability of being the same type as the prior planet ("no switch"). However, there was a 20% probability of transitioning or "switching" to a planet of a different type.

Participants

We recruited 176 participants from Amazon Mechanical Turk (111 male, 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 earned $6 as a base payment and could earn a bonus contingent on performance ($0-$4). We excluded 60 participants according to one or more of three criteria: 1. having average planet residence times 2 standard deviations above or below the group mean (36 participants) 2. failing a quiz on the task instructions more than 2 times (33 participants) or 3. failing to respond appropriately to one or more of the two catch trials (17 participants). On catch trials, participants were asked to press the letter "Z" on their keyboard. These questions were meant to "catch" any participants repeatedly choosing the same option (using key presses "A" or "L") independent of value.

Task Design

Participants completed a serial stay-switch task adapted from previous human foraging studies[4,16]. With the goal of collecting as much space treasure as possible, participants traveled to different planets to mine for gems. Upon arrival to a new planet, they performed an initial dig and received an amount of gems sampled from a Gaussian distribution with a mean of 100 and standard deviation (SD) of 5. Following this initial dig, participants had to decide between staying on the current planet to dig again or leaving to travel to a new planet (Fig 1A). Staying would further deplete the gem mine while leaving yielded a replenished gem mine at the cost of a longer time delay. They made these decisions in a series of five blocks, each with a fixed length of 6 minutes. Blocks were separated by a break of participant-controlled length, up to a maximum of 1 minute.

On each trial, participants had 2 seconds to decide via key press whether to stay ("A") or leave ("L"). If they decided to stay, they experienced a short delay before the gem amount was displayed (1.5 s). The length of the delay was determined by the time the participant spent making their previous choice (2 - RT s). This ensured participants could not affect the environment reward rate via their response time. If they decided to leave, they encountered a longer time delay (10 s) after which they arrived on a new planet and were greeted by a new alien (5 s). On trials where a decision was not made within the allotted time (2 s), participants were shown a timeout message for two seconds.

Unlike previous variants of this task, planets varied in their richness within and across blocks, introducing greater structure to the task environment. Richness was determined by the rate at which the gem amount exponentially decayed with each successive dig (Fig. 1B). If a planet was “poor”, there was steep depletion in the amount of gems received. Specifically, its decay rates were sampled from a beta distribution with a low mean (mean = 0.2; sd = 0.05; $\alpha = 13$ and $\beta = 51$). In contrast, rich planets depleted more slowly (mean = 0.8; sd = 0.05; $\alpha = 50$ and $\beta = 12$). Finally, the quality of the third planet type

— neutral — fell in between rich and poor (mean = 0.5; sd = 0.05; $\alpha = 50$ and $\beta = 50$). The environment dynamics were designed such that planet richness was correlated in time. When traveling to a new planet, there was an 80% probability of it being the same type as the prior planet ("no switch"). If not of the same type, it was equally likely to be of one of the remaining two types ("switch", Fig. 1C). This information was not communicated to participants, requiring them to infer the environment's structure and dynamics from rewards received alone.

Comparison to Marginal Value Theorem

Participants' planet residence times, or PRTs, were compared to those prescribed by MVT. Under MVT, agents are generally assumed to act as though they have accurate and complete knowledge of the environment. For this task, that would include knowing each planet type's unique decay rate distribution and the total reward received and time elapsed across the environment.

Knowledge of the decay rate distributions is critical for estimating V_{stay}, the anticipated reward if the agent were to stay and dig again.

$$V_{stay} = r_t * d \tag{1}$$

where r_t is the reward received on the last dig and d is the upcoming decay.

$$d = \begin{cases} 0.2 & \text{if planet is poor} \\ 0.5 & \text{if planet is neutral} \\ 0.8 & \text{if planet is rich} \end{cases}$$

V_{leave} is estimated using the total reward accumulated, r_{total}, total time passed in the environment, t_{total}, and the time delay to reward associated with staying and digging, t_{dig}.

$$V_{leave} = \frac{r_{total}}{t_{total}} * t_{dig} \tag{2}$$

$\frac{r_{total}}{t_{total}}$ estimates the average reward rate of the environment. Multiplying it by t_{dig} gives the opportunity cost of the time spent exploiting the current planet.

Finally, to make a decision, the MVT agent compares the two values and acts greedily, always taking the higher valued option.

$$\text{choice} = \text{argmax}(V_{stay}, V_{leave}) \tag{3}$$

Model

Making the stay-leave decisions

We assume that the forager compares the value for staying, V_{stay}, to the value of leaving V_{leave}, to make their decision. Similar to MVT, we assume foragers act greedily with respect to these values.

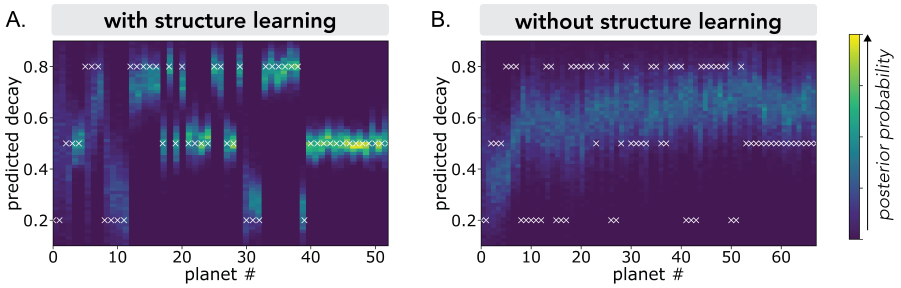

Fig. 2 Structure learning improves prediction accuracy. A. With structure learning A simulated agent's posterior probability over the upcoming decay rate on each planet is plotted. If the forager's prior allows for the possibility of multiple clusters ($\alpha > 0$), they learn with experience the cluster-unique decay rates. Initially, the forager is highly uncertain of their predictions. However, with more visitations to different planets, the agent makes increasingly accurate and precise predictions. **B. Without structure learning** If the forager's prior assumes a single cluster ($\alpha = 0$), the forager makes inaccurate and imprecise predictions - either over or underestimating the upcoming decay, depending on the planet type. This inaccuracy persists even with experience because of the strong initial assumption.

Learning the structure of the environment affords more accurate and precise predictions which support better decision-making. Here, the forager predicts how many gems they'll receive if they stay and dig again and this determines the value of staying, V_{stay}. To generate this prediction, a forager could aggregate over all past experiences in the environment[4]. This may be reasonable in homogeneous environments but less so in heterogeneous ones where it could introduce substantial noise and uncertainty. Instead, in these varied environments, it may be more reasonable to cluster patches based on similarity and only generalize from patches belonging to the same cluster as the current one. This selectivity enables more precise predictions of future outcomes.

Clusters are latent constructs. Thus, it is not clear how many clusters a forager *should* divide past encounters into. Non-parametric Bayesian methods provide a potential solution to this problem. They allow for the complexity of the representation — as measured by the number of clusters — to grow freely as experience accumulates. These methods have been previously used to explain phenomena in category learning[13,17], task set learning[12], fear conditioning[14], and event segmentation[18].

To initiate this clustering process, the forager must assume a model of how their observations, decay rates, are generated by the environment. The generative model we ascribe to the forager is as follows. Each planet belongs to some cluster, and each cluster is defined by a unique decay rate distribution:

$$d_k \sim Normal(\mu_k, \sigma_k) \tag{4}$$

where k denotes cluster number. The generative model takes the form of a *mixture model* in which normal distributions are mixed together according to some distribution $P(k)$ and observations are generated from sampling from the distribution $P(d|k)$.

Before experiencing any decay on a planet, the forager has prior expectations regarding the likelihood of a planet belonging to a certain cluster. We assume that the prior on clustering corresponds to a "Chinese restaurant process"[19]. If previous planets are clustered according to $p_{1:N}$, then for the current planet:

$$P(k) = \begin{cases} \frac{n_k}{N+\alpha} & \text{if k is old} \\ \frac{\alpha}{N+\alpha} & \text{if k is new} \end{cases}$$

Where n_k is the number of planets assigned to cluster k, α is a clustering parameter, and N is the total number of planets encountered. The probability of a planet belonging to an old cluster is proportional to the number of planets already assigned to it. The probability of it belonging to a new cluster is proportional to α. Thus, α controls how dispersed the clusters are — the higher α is the more new cluster creation is encouraged. The ability to incrementally add clusters as experience warrants it makes the generative model an *infinite capacity mixture model*.

After observing successive depletions on a planet, the forager computes the posterior probability of a planet belonging to a cluster:

$$P(k|D) = \frac{P(D|k)P(k)}{\sum_{j=1}^{J} P(D|j)P(j)} \tag{5}$$

Where J is the number of clusters created up until the current planet, D is a vector of all the depletions observed on the current planet, and all probabilities are conditioned on prior cluster assignments of planets, $p_{1:N}$.

Exact computation of this posterior is computationally demanding as it requires tracking all possible clusterings of planets and the likelihood of the observations given those clusterings. Thus, we approximate the posterior distribution using a particle filter[20]. Each particle maintains a hypothetical clustering of planets which are weighted by the likelihood of the data under the particle's chosen clustering. All simulations and fitting were done with 1 particle which is equivalent to Anderson's local MAP algorithm[21].

With 1 particle, we assign a planet definitively to a cluster. This posterior then determines (a) which cluster's parameters are updated and (b) the inferred cluster on subsequent planet encounters.

If the planet is assigned to an old cluster, k, the existing μ_k and σ_k are updated analytically using the standard equations for computing the posterior

for a normal distribution with unknown mean and variance:

$$\begin{aligned}
\bar{d} &= \frac{1}{n}\sum_{i=1}^{n} d_i \\
\mu_0' &= \frac{n_0\mu_0 + n\bar{d}}{n_0 + n} \\
n_0' &= n_0 + n \\
\nu_0' &= \nu_0 + n \\
\nu_0'\sigma_0^{2\prime} &= \nu_0\sigma_0^2 + \sum_{i=1}^{n}(d_i - \bar{d})^2 + \frac{n_0 n}{n_0 + n}(\mu_0 - \bar{d})^2
\end{aligned} \tag{6}$$

where d is a decay observed on the current planet, n is the total number of decays observed on the current planet, n_0 is the total number of decays observed across the environment before the current planet, μ_0 is the prior mean of the cluster-specific decay rate distribution and ν_0 is its precision. $\mu_0^{'}$ and $\nu_0^{'}$ are the posterior mean and variance respectively.

If the planet is a assigned to a new cluster, then a new cluster is initialized with the following distribution:

$$d_{new} \sim Normal(0.5, 0.25) \tag{7}$$

This initial distribution is updated with the depletions encountered on the current planet upon leaving.

The goal of this learning and inference process is to support accurate prediction. To generate a prediction of the next decay, the forager samples a cluster according to $P(k)$ or $P(k|D)$ depending on whether any depletions have been observed on the current planet. Then, a decay rate is sampled from the cluster specific distribution, d_k. The forager averages over these samples to produce the final prediction.

To demonstrate structure learning's utility for prediction, we show in simulation the predicted decay rates on each planet with structure learning (Fig. 2A) and without (Fig. 2B). With structure learning, the forager's predictions approach the mean decay rates of the true generative distributions. Without structure learning, however, the forager is persistently inaccurate, underestimating the decay rate on rich planets and overestimating it on poor planets.

Adapting the model of the environment

Because the inference process is an approximation and foragers' experience is limited, their inferred environment structure may be inaccurate. Theoretical work has suggested that a rational way to compensate for this inaccuracy is to discount future values in proportion to the agent's uncertainty over their representation of the environment [22]. We quantified an agent's uncertainty by taking the entropy of the approximated posterior distribution over clusters (Fig

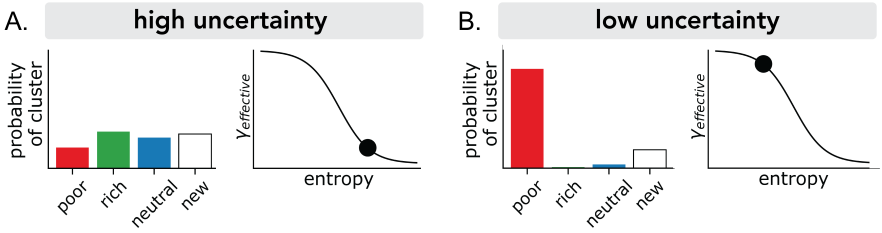

Fig. 3 Uncertainty adaptive discounting. A. High uncertainty When clusters are similarly probable, the posterior entropy is high. This entropy is taken as the forager's internal uncertainty and is used to adjust their discounting rate, $\gamma_{effective}$. When uncertainty is high, they discount future value more heavily. **B. Low uncertainty** When one cluster is much more likely than the others, entropy or uncertainty is low and consequently, future value is discounted less heavily.

3). We sample clusters 100 times proportional to the posterior. These samples are multinomially distributed. We represent them with the distribution, X:

$$X \sim Multinomial(100, K) \tag{8}$$

Where K is a vector containing the counts of clusters from sampling 100 times from the distribution, $P(k)$ or $P(k|d)$ depending on whether depletions on the planet have been observed. Uncertainty is quantified as the Shannon entropy of distribution X.

We implemented this proposal in our model by discounting the value of leaving as follows:

$$V_{leave} = \frac{r_{total}}{t_{total}} * t_{dig} * \gamma_{effective} \tag{9}$$

$$\gamma_{effective} = \frac{1}{1 + e^{(-\gamma_{base} + \gamma_{coef} * H(X))}} \tag{10}$$

where γ_{base} and γ_{coef} are free parameters and $H(X)$ is the entropy of the distribution X.

Model fitting

We compared participant PRTs on each planet to those predicted by the model. A model's best fitting parameters were those that minimized the difference between the true participant's and simulated agent's PRTs. We considered 1000 possible sets of parameters generated by quasi-random search using low-discrepancy Sobol sequences[23]. Prior work has demonstrated random and quasi-random search to be more efficient than grid search[24] for parameter optimization. Quasi-random search is particularly efficient with low-discrepancy sequence, more evenly covering the parameter space relative to true random search.

Because cluster assignment is a stochastic process, the predicted PRTs vary slightly with each simulation. Thus, for each candidate parameter setting, we simulated the model 50 times and averaged over the mean squared error (MSE)

between participant PRTs and model-predicted PRTs for each planet. The parameter configuration that produced the lowest MSE on average was chosen as the best fitting for the individual.

Model Comparison

We compared three models: the structure learning and adaptive discounting model described above, a temporal difference model previously applied in a foraging context, and a MVT model that learns the mean decay rate and global reward rate of the environment.

MVT-Learning In this model, the agent learns a threshold for leaving which is determined by the global reward rate, ρ[4]. ρ is learned with a simple delta rule with α as a learning rate and taking into account the temporal delay accompanying an action τ. The value of staying is $d * r_i$ where d is the predicted decay and r_i is the reward received on the last time step. The value of leaving,V_{leave}, is the opportunity cost of the time spent digging, $\rho * t_{dig}$. The agent chooses an action using a softmax policy with temperature parameter, β which determines how precisely the agent represents the value difference between the two options.

$$\begin{aligned} P(a_i = dig) &= \frac{1}{(1 + e^{(-c-\beta(d*r_i-\rho*t_{dig}))})} \\ \delta_i &= \tfrac{r_i}{\tau_i} - \rho_i \\ \rho_{i+1} &= \rho_i + (1 - (1-\alpha)^{\tau_i}) * \delta_i \end{aligned} \tag{11}$$

TD-Learning The temporal difference (TD) agent learns a state-specific value of staying and digging, $Q(s, dig)$ and a non-state specific value of leaving, $Q(leave)$. The state, s is defined by the gem amounts offered on each dig. The state space is defined by binning the possible gems that could be earned from each dig. The bins are spaced are according to $log(b_{j+1})$ - $log(b_j)$ = $log(\bar{k})$ where b_{j+1} and b_j are the upper and lower bounds of the bins and $\bar{d}$ is the mean decay rate. This state space specification is taken from[4]. We set b_{j+1} to 135 and b_j to 0 as these were the true bounds on gems received per dig. We set $\bar{k}$ to 0.5 because this would be the mean decay rate if one were to average the depletions experienced over all planets. The agent compares the two values and makes their choice using a softmax policy.

$$\begin{aligned} P(a_i = dig) &= \frac{1}{(1 + e^{(-c-\beta(Q_i(s_i,dig)-Q_i(leave)))})} \\ \mathrm{D}_i &\sim Bernoulli(P(a_i)) \\ \delta_i = r_i + \gamma\tau_i(D_i * Q_i(s_i) + (1 - D_i) * Q_i(leave)) - Q_i(s_{i-1}, a_{i-1}) \\ \mathrm{Q}_i(s_{i-1}, a_{i-1}) &= Q_{i+1}(s_{i-1}, a_{i-1}) + \alpha * \delta_i \end{aligned} \tag{12}$$

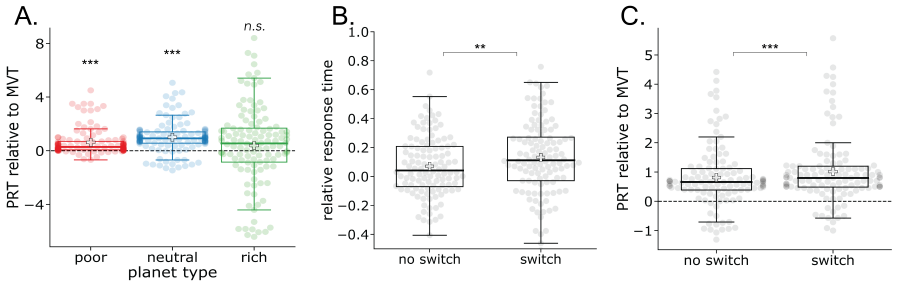

Fig. 4 Model-free results A. Planet richness influences over and underharvesting behavior. Planet residence times (PRT) relative to Marginal Value Theorem's (MVT) prediction are plotted as the median (± one quartile) across participants. The grey line indicates the median while the white cross indicates the mean. Individuals' PRTs relative to MVT are plotted as shaded circles. In aggregate, participants overharvested on poor and neutral planets and acted MVT optimally on rich planets. **B. Decision times are longer following rare switch transitions.** If a participant has knowledge of the environment's planet types and the transition structure between them, then they should be surprised following a rare transition to a different type. Consequently, they should take longer to decide following these transitions. As predicted, participants spent longer making a decision following transitions to different types ("switch") relative to when there was transition to a planet of the same type ("no switch"). This is consistent with having knowledge of the environment's structure and dynamics. **C. Overharvesting increases following rare switch transitions.** On poor and neutral planets, participants overhavested to a greater extent following a rare "switch" transition relative to when there was a "no switch" transition. This is consistent with uncertainty adaptive discounting. Switches to different planet types should be points of greater uncertainty. This greater uncertainty produces heavier discounting and in turn staying longer with the current option.*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

where c, α, β, γ are free parameters. c is a perseveration term, α is the learning rate, β is the softmax temperature, and γ is the temporal discounting factor.

Cross Validation 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.

Results

Model-free analyses

Participants adapt to local richness

We first examined a prediction of MVT — foragers should adjust their patch leaving to the richness of the local patch. In the task environment, planets varied in their richness or how quickly they depleted. Slower depletion causes

the local reward rate to more slowly approach the global reward rate of the environment. Thus, MVT predicts that stay times should increase as depletion rates slow. As predicted, participants stayed longer on rich planets relative to neutral ($t(115) = 19.77$, $p < .0001$) and longer on neutral relative to poor ($t(115) = 12.57$, $p < .0001$).

Experience decreases overharvesting

Despite modulating stay times in the direction prescribed by MVT, participants stayed longer or overharvested relative to MVT when averaging across all planets ($t(115) = 3.88$, $p = .00018$). However, the degree of overharvesting diminished with experience. Participants overharvested more in the first two blocks relative to the final two ($t(115) = 3.27$, $p = .0014$). Our definition of MVT assumes perfect knowledge of the environment. Thus, participants approaching the MVT optimum with experience is consistent with learning the environment's structure and dynamics.

Local richness modulates overharvesting

We next considered how participants' overharvesting varied with planet type. As a group, participants overharvested only on poor and neutral planets while behaving MVT optimally on rich planets (Fig. 4A; poor - $t(115) = 6.92$, $p < .0001$; neutral - $t(115) = 9.00$, $p < .0001$; rich - $t(115) = 1.38$, $p = .17$).

Environment dynamics modulates decision time and overharvesting

We also asked how participants adapted their foraging strategy to the environment's dynamics or transition structure. Upon leaving a planet, it was more common to transition to a planet of the same type (80%, "no switch") than transition to a planet of a different type ("switch"). Thus, we reasoned that switch transitions should be points of maximal surprise and uncertainty given their rareness. However, this would only be the case if the participant could discriminate between planet types and learned the transition structure between them.

If surprised, a participant should take longer to make a choice following a rare "switch" transition. So, we next examined participants' reaction times (z-scored and log-transformed) for the decision following the first depletion on a planet. We compared when there was a switch in planet type versus where there was none. As predicted, participants showed longer decision times following a "switch" transition suggesting they were sensitive to the environment's structure and dynamics (Fig. 4B; $t(115) = 2.65$, $p = 0.0093$).

If uncertain, our adaptive discounting model predicts that participants should discount remote rewards more heavily and, consequently, overharvest to a greater extent. To test this, we compared participants overharvesting following rare "switch" transitions to their overharvesting following the more common "no switch" transitions. Following the model's prediction, participants

marginally overharvested more following a change in planet type (t(115) = 1.86, $p = 0.065$). When considering only planets that participants overharvested on on average (poor and neutral), overharvesting was significantly greater following a change (Fig. 4C; t(115) = 4.67, $p < .0001$).

Computational modeling

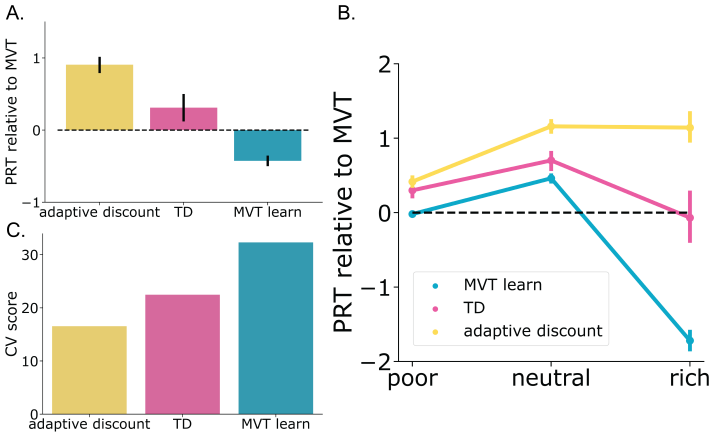

Fig. 5 Modeling results A. The adaptive discounting model predicts overharvesting. Averaging across all planets, only the adaptive discounting model predicts overharvesting while the temporal-difference learning model predicts MVT optimal behavior and the MVT learning model predicts underharvesting. This demonstrates that overharvesting, a seemingly suboptimal behavior, can emerge from principled statistical inference and adaptation. **B. Model predictions diverge most on rich planets.** Similar to participants, the greatest differences in behavior between the models occurred on rich planets. **C. The adaptive discounting model provides the best account for participant choices.** The adaptive discounting model had the lowest mean cross validation score indicating it provided the best account of participant choice at the group level.

Structure learning with adaptive discounting provide the best account of participant choice

To check the models' goodness of fit, we asked whether the compared models could capture key behavioral results found in participants' data. For each model and participant, we simulated an agent with the best fitting parameters estimated for them under the given model. Only the adaptive discounting model was able to account for overharvesting when averaging across all planets (Fig. 5A, t(115) = 9.03, $p < .0001$). The temporal-difference learning model predicted MVT optimal choices on average (t(115) = 1.09, $p = .28$) while the MVT learning model predicted underharvesting (t(115) = -7.17, $p < .0001$). These differences were primarily driven by predicted behavior on the rich planets (Fig. 5B).

Model fit was also assessed at a more granular level (stay times on individual planets) using 10-fold cross validation. Comparing cross validation scores as a group, participants' choices were best captured by the adaptive discounting model (Fig. 5C; mean cross validation scores — adaptive discounting: 16.55, TD: 22.47, MVT learn: 32.31). At the individual level, 64% of participants were best fit by the adaptive discounting model, 14% by TD, and 22% by MVT learn.

Adaptive discounting model parameter distribution

Because the adaptive discounting model provided the best account of choice for most participants, we examined the distribution of individuals' best fitting parameters for the model. Specifically, we compared participants' estimated parameters to two thresholds. These thresholds were used to identify whether a participant 1) inferred and assigned planets to multiple clusters and 2) adjusted their overharvesting in response to internal uncertainty.

The threshold for multi-cluster inference, 0.8, was computed by simulating the adaptive discounting model 100 times and finding the lowest value that produced multi-cluster inference in 90% of simulations. 76% of participants were above this threshold. Thus, most participants were determined to be "structure learners" using our criteria.

The threshold for uncertainty-adaptive discounting was assumed to be 0. A majority of participants, 93%, were above this threshold. These participants were determined to be "adaptive discounters", those who dynamically modulated their discounting factor in accordance with their internal uncertainty.

We next looked for relationships between parameters. Uncertainty should be greatest for individuals who have prior expectations that do not match the environment's true structure, whether too complex or too simple. Consistent with this, there was a non-monotonic relationship between the structure learning and discounting parameters. γ_{base} and γ_{coef} were greatest when α was near its lower bound, 0, and upper bound, 10 (γ_{base}: $\beta = 0.080$, $p < .0001$; γ_{coef}: $\beta = 0.021$, $p < .0001$). An individual's base level discounting constrains the range over which uncertainty can adapt the effective discounting. Reflecting this, the two discounting parameters were positively related to one another (τ = -0.33, $p < .0001$).

Parameter validation

Correlations with model-free measures of task behavior confirmed the validity of the model's parameters. We interpret α as reflecting an individual's prior expectation of environment complexity. α must reach a certain threshold to produce inference of multiple clusters and consequently, sensitivity to the transitions between clusters. Validating this interpretation, participants with higher fit α demonstrated greater switch costs between planet types (Kendall's $\tau = 0.24$, $p = 0.00076$). Moreover, this relationship was specific to α. γ_{base} and γ_{coef} were not significantly correlated with switch cost behavior (γ_{base}: τ = -0.036, $p = .57$; γ_{coef}: τ = -0.10, $p = 0.11$). This is a particularly strong

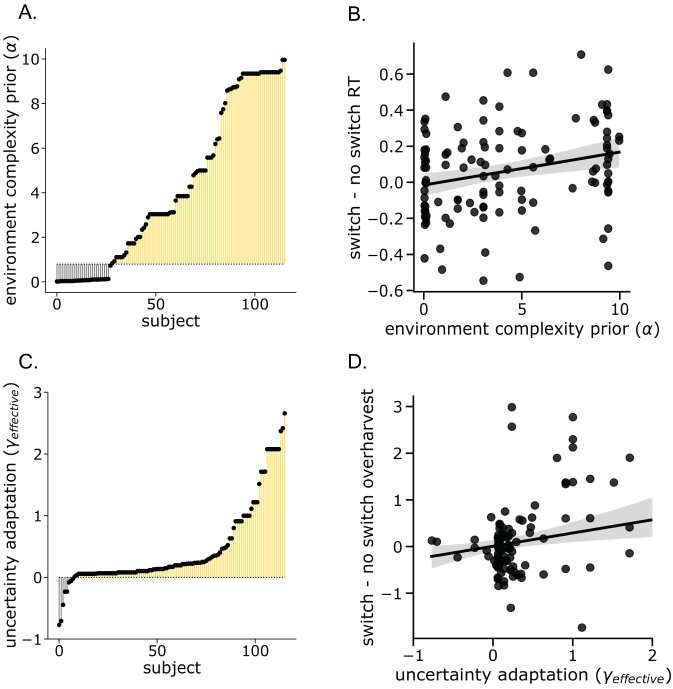

Fig. 6 Parameter distributions A. Participants learned the structure of the environment. Distribution of participants' priors over environment complexity, α. Each individual's parameter is shown relative to a baseline threshold, 0.8. This threshold is the lowest value that produced multi-cluster inference in simulation. Most participants (76%) fall above this threshold indicating a majority learned the environment's multi-cluster structure. **B. Environment complexity parameters were positively related to reaction time sensitivity to transition frequency.** An individual must infer multiple planet types to be sensitive to the transition structure between them. In terms of the model, this would correspond to having a sufficiently high environment complexity parameter. Validating this parameter, it was positively correlated with individual's modulation of reaction time following a rare transition to a different planet type. **C. Participants adapted their discounting computations to their uncertainty over environment structure.** Distribution of participant's uncertainty adaptation parameter, γ_{coef}. Each individual's parameter is shown relative to a baseline of 0. A majority were above this threshold (93%) indicating most participants dynamically adjusted their discounting, increasing it when they experienced greater internal uncertainty. **D. Uncertainty adaptation parameters were positively related to overharvesting sensitivity to transition frequency.** If an individual increases their discounting to their internal uncertainty over environment structure, then they should discount more heavily following rare transitions and stay longer with the current option. Consistent with this, we found that the extent an individual increased their overharvesting following a rare transition was related to their uncertainty adaptation parameter.

validation as the model was not fit to reaction time data. Validating γ_{coef} as reflecting uncertainty-adaptive discounting, the parameter was correlated with the extent overharvesting increased following a rare transition or "switch" between different planet types (τ = 0.15, p = 0.016). This was not correlated with α nor the baseline discounting factor γ_{base} (α: τ = -0.011, p = .86; γ_{base}: τ = 0.082, p = .20).

Discussion

While Marginal Value Theorem (MVT) provides an optimal solution to patch leaving problems, organisms systematically deviate from it, staying too long or overharvesting. A critical assumption of MVT is that the forager has accurate and complete knowledge of the environment. Yet, this is often not the case in real world contexts — the ones in which foraging behaviors are likely to have adapted[25]. We propose a model of how foragers could rationally learn the structure of their environment and adapt their foraging decisions to it. In simulation, we demonstrate how seemingly irrational overharvesting can emerge as a byproduct of a rational dynamic learning process. In a heterogeneous, multimodal environment, we compared how well our structure learning model predicted participants' choices relative to two other models — one implementing a MVT choice rule with a fixed representation of the environment and the other a standard temporal-difference learning algorithm. Importantly, only our structure learning model predicted overharvesting in this environment. Participants' choices were most consistent with learning a representation of the environment's structure through individual patch experiences. They leveraged this structured representation to inform their strategy in multiple ways. One way determined the value of staying. The representation was used to predict future rewards from choosing to stay in a local patch. The other modulated the value of leaving. Uncertainty over the accuracy of the representation was used to set the discount factor over future value. These results suggest that to explain foraging as it occurs under naturalistic conditions optimal foraging may need to provide an account of how the forager learns to acquire accurate and complete knowledge of the environment, and how they adjust their strategy as their representation is refined with experience.

If foragers are learning a model of the environment and using it to make decisions for reward, then this suggests that they may be doing something like model-based reinforcement learning (RL). Seemingly contrary to this, Constantino and Daw [4] found human foragers' choices to be better explained by a MVT model augmented with a learning rule than a standard reinforcement learning model. However, it is important to note that the task environment in that study was homogeneous and the RL model tested was model-free (temporal-difference learning). Thus, the difference in our results could be due to different task environments and class of models. A key way our model deviates from a model-based RL approach is that prospective prediction is only applied in computing the value of staying while the value of leaving is similar to MVT's threshold for leaving – albeit discounted proportionally to the agent's internal uncertainty over their representation's accuracy. In the former respect, our model parallels the one proposed by Kolling and Akam [26] to explain humans sensitivity to the gradient of reward rate change during foraging observed by Wittmann et al. [27]. Given that computing the optimal exit threshold under a pure model-based strategy would be highly computationally expensive, Kolling and Akam [26]'s model pairs model-based patch evaluation with a model-free, MVT-like exit threshold. Specifically, under this model, the agent leaves once

the local patch's average predicted reward rate over n time steps in the future falls below the global reward rate. Both our model and Kolling and Akam [26]'s model demonstrate the potential role of representation learning and planning over these representations in explaining foraging behavior.

In standard economic choice, prior work has demonstrated that individuals' behavior can be rationalized by considering their potential uncertainty over environment structure and how they attempt to resolve it. For instance, deviations from the optimal balance of exploration and exploitation can be rationalized by considering that individuals must explore both at the level of actions and at the level of hypotheses over environment structure and dynamics [10]. These hypotheses could include the probability of reward produced by an action changing over time or the probability of reward being correlated across actions. Assuming one of these hypotheses will have implications for how widely new information should be generalized across time and actions. This reveals how a representation of the environment supports efficient learning, allowing large gains in knowledge from sparse experience. However, information can sometimes be generalized too broadly. Learners may inappropriately transfer information from an old environment to a novel one if they overestimate the shared structure between the environments [28]. Importantly, humans are able to improve their ability to transfer information appropriately ("meta-learning"). While occasionally costly, structure representations more often than not facilitate robust learning particularly when coupled with meta-learning abilities, allowing information to be spread between options both within and across environments. The underweighting of a choice's delayed outcomes has also been rationalized by considering a decision-maker's uncertainty over environment structure, or more specifically, the relationship between actions and their long-term consequences [11]. By assuming that this relationship must be learned, a preference for lesser, local gains over greater, later gains becomes reasonable. Here, the desire to learn the consequences of an action over multiple time scales reveals the utility of structure representations in goal-directed planning [29–31]. By considering the seemingly suboptimal patterns of choice introduced by structure learning, we gain insight into the motivations for structure learning itself. Amongst them are efficient learning of generalizable knowledge and the ability to learn and leverage long-term dependencies.

While structure learning is beneficial, it is also challenging and computationally costly. With limited experience and computational noise, an inaccurate model of the environment may be inferred. An inaccurate model, however, can be counteracted by adapting certain computations. For example, lowering the temporal discounting factor acts as a form of regularization or variance reduction [22,32–35]. Empirical work has found humans appear to do something like this in standard intertemporal choice tasks. Gershman and Bhui [36] found evidence that individuals rationally set their temporal discounting as a function of the imprecision or uncertainty of their internal representations. Here, we found that humans while foraging act similarly, overharvesting to a greater extent at points of peak uncertainty. While temporal discounting has been proposed as a

mechanism of overharvesting previously [3–5], the discounting factor is usually treated as a fixed, subject-level parameter, inferred from choice. Thus, it provides no mechanism for how the factor is set let alone dynamically adjusted with experience. In contrast, our model proposes a mechanism through which the discounting factor is rationally set in response to both the external and internal environment.

The costs and benefits of inferring a structure representation and storing it in memory is determined by qualities of the environment. Foraging strategies leveraging mental maps of previous reward locations are particularly useful in heterogeneous but predictable environments [37–39]. This characterizes the environments that humans typically forage in — we have a tendency to rely on high-value but costly to obtain and widely-dispersed resources. A reliance on these resources could have pressured us, whether over a developmental or evolutionary timescale, towards a predisposition to seek out structure. This idea has been explored in non-human animals. Animals whose food resources are patchily and widely distributed tend to rely on memory-based foraging strategies and have greater spatial memory ability than animals whose resources are more homogenously distributed. For instance, dolphins dive to varying depths to collect their prey [40]. Constrained by their to need to surface to breathe, dolphins appear to recall prey distributions at different depths and plan their foraging dives accordingly [41]. Primates memory ability and use also appears to adapt in response to the environment. Rosati [42] found that chimpanzees, who feed on more variably and patchily distributed fruit, demonstrated greater spatial memory ability than bonobos, who rely on more homogeneously distributed terrestrial herbs. Furthermore, spider monkeys and baboons appear to rely on detailed mental maps of prior reward sites that guide their trajectories through forest habitats [43,44]. Potentially, careful analysis of the statistics and structure of naturalistic environments may explain how and why humans adjust their use of structure representations across different decision contexts and may rationalize apparent suboptimalities.

By assuming complete and accurate knowledge of the environment, MVT has limited explanatory reach in most naturalistic environments. Optimal foraging theory has provided behavioral rules for how key decision variables could be learned and optimally used assuming a fixed representation of the environment [45,46]. However, it remains unclear how decision-making should adapt to dynamic representations. Different assumptions lead to different definitions of optimality. Thus, relaxing a key assumption of MVT suggests optimal foraging theory must be expanded. Potentially, its applicability could be broadened by considering decision rules that are flexible enough to support representational change.

References

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

2. Andrew M Wikenheiser, David W Stephens, and A David Redish. 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, May 2013.
3. Evan C Carter and A David Redish. Rats value time differently on equivalent foraging and delay-discounting tasks. *J. Exp. Psychol. Gen.*, 145(9):1093–1101, September 2016.
4. Sara M Constantino and Nathaniel D Daw. Learning the opportunity cost of time in a patch-foraging task. *Cogn. Affect. Behav. Neurosci.*, 15(4): 837–853, December 2015.
5. Tommy C Blanchard and Benjamin Y Hayden. Monkeys are more patient in a foraging task than in a standard intertemporal choice task. *PLoS One*, 10(2):e0117057, February 2015.
6. Gary A Kane, Aaron M Bornstein, Amitai Shenhav, Robert C Wilson, Nathaniel D Daw, and Jonathan D Cohen. Rats exhibit similar biases in foraging and intertemporal choice tasks. *Elife*, 8, September 2019.
7. Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.
8. Neil Garrett and Nathaniel D Daw. Biased belief updating and suboptimal choice in foraging decisions. *Nat. Commun.*, 11(1):3417, July 2020.
9. Zachary P Kilpatrick, Jacob D Davidson, and Ahmed El Hady. Uncertainty drives deviations in normative foraging decision strategies. April 2021.
10. Daniel E Acuña and Paul Schrater. Structure learning in human sequential decision-making. *PLoS Comput. Biol.*, 6(12):e1001003, December 2010.
11. Chris R Sims, Hansjörg Neth, Robert A Jacobs, and Wayne D Gray. Melioration as rational choice: sequential decision making in uncertain environments. *Psychol. Rev.*, 120(1):139–154, January 2013.
12. Anne G E Collins and Michael J Frank. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychol. Rev.*, 120 (1):190–229, January 2013.
13. Adam N Sanborn, T L Griffiths, and D J Navarro. A more rational model of categorization. January 2006.
14. Samuel J Gershman, David M Blei, and Yael Niv. Context, learning, and extinction. *Psychol. Rev.*, 117(1):197–209, January 2010.
15. Johannes H Decker, A Ross Otto, Nathaniel D Daw, and Catherine A Hartley. From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological science*, 27(6):848–858, 2016.
16. Jennifer K Lenow, Sara M Constantino, Nathaniel D Daw, and Elizabeth A Phelps. Chronic and acute stress promote overexploitation in serial decision making. *J. Neurosci.*, 37(23):5681–5689, June 2017.
17. Adam N Sanborn, Thomas L Griffiths, and Daniel J Navarro. Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.*, 117(4):1144–1167, October 2010.

18. Yeon Soon Shin and Sarah DuBrow. Structuring memory through Inference-Based event segmentation. *Top. Cogn. Sci.*, 13(1):106–127, January 2021.
19. Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Ann. Stat.*, 2(6):1152–1174, 1974.
20. Paul Fearnhead. Particle filters for mixture models with an unknown number of components. *Stat. Comput.*, 14(1):11–21, January 2004.
21. John R Anderson. The adaptive nature of human categorization. *Psychol. Rev.*, 98(3):409–429, 1991.
22. Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. https://nanjiang.cs.illinois.edu/files/gamma-AAMAS-final.pdf. Accessed: 2022-2-18.
23. Illya M Sobol. Distribution of points in a cube and approximate evaluation of integrals. *Zh. Vych. Mat. Mat. Fiz.*, 7:784–802, 1967.
24. James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf, 2012. Accessed: 2021-5-6.
25. Benjamin Y Hayden. Time discounting and time preference in animals: a critical review. *Psychonomic bulletin & review*, 23(1):39–53, 2016.
26. Nils Kolling and Thomas Akam. (reinforcement?) learning to forage optimally. *Curr. Opin. Neurobiol.*, 46:162–169, October 2017.
27. Marco K Wittmann, Nils Kolling, Rei Akaishi, Bolton K H Chau, Joshua W Brown, Natalie Nelissen, and Matthew F S Rushworth. Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex. *Nat. Commun.*, 7:12327, August 2016.
28. Anne G E Collins. The cost of structure learning. *J. Cogn. Neurosci.*, 29 (10):1646–1655, October 2017.
29. Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.
30. Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, 8(12):1704–1711, December 2005.
31. Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, March 2011.
32. Marek Petrik and Bruno Scherrer. Biasing approximate dynamic programming with a lower discount factor. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/file/08c5433a60135c32e34f46a71175850c-Paper.pdf.
33. Vincent Francois-Lavet, Guillaume Rabusseau, Joelle Pineau, Damien Ernst, and Raphael Fonteneau. On overfitting and asymptotic bias in batch reinforcement learning with partial observability. *J. Artif. Intell. Res.*, 65: 1–30, May 2019.

34. Harm van Seijen, Mehdi Fatemi, and Arash Tavakoli. Using a logarithmic mapping to enable lower discount factors in reinforcement learning. June 2019.
35. Ron Amit, Ron Meir, and Kamil Ciosek. Discount factor as a regularizer in reinforcement learning. July 2020.
36. Samuel J Gershman and Rahul Bhui. Rationally inattentive intertemporal choice. *Nat. Commun.*, 11(1):3365, July 2020.
37. Denis Boyer and Peter D Walsh. Modelling the mobility of living organisms in heterogeneous landscapes: does memory improve foraging success? *Philos. Trans. A Math. Phys. Eng. Sci.*, 368(1933):5645–5659, December 2010.
38. Chloe Bracis, Eliezer Gurarie, Bram Van Moorter, and R Andrew Goodwin. Memory effects on movement behavior in animal foraging. *PLoS One*, 10 (8):e0136057, August 2015.
39. Louise Riotte-Lambert, Simon Benhamou, and Simon Chamaillé-Jammes. How memory-based movement leads to nonterritorial spatial segregation. *Am. Nat.*, 185(4):E103–16, April 2015.
40. G D Hastie, B Wilson, and P M Thompson. Diving deep in a foraging hotspot: acoustic insights into bottlenose dolphin dive depths and feeding behaviour. *Mar. Biol.*, 148(5):1181–1188, March 2006.
41. Patricia Arranz, Kelly J Benoit-Bird, Brandon L Southall, John Calambokidis, Ari S Friedlaender, and Peter L Tyack. Risso's dolphins plan foraging dives. *J. Exp. Biol.*, 221(Pt 4), February 2018.
42. Alexandra G Rosati. Foraging cognition: Reviving the ecological intelligence hypothesis. *Trends Cogn. Sci.*, 21(9):691–702, September 2017.
43. Rahel Noser and Richard W Byrne. Mental maps in chacma baboons (papio ursinus): using inter-group encounters as a natural experiment. *Anim. Cogn.*, 10(3):331–340, July 2007.
44. Alejandra Valero and Richard W Byrne. Spider monkey ranging patterns in mexican subtropical forest: do travel routes reflect planning? *Anim. Cogn.*, 10(3):305–315, July 2007.
45. John M McNamara and Alasdair I Houston. Optimal foraging and learning. *J. Theor. Biol.*, 117(2):231–249, November 1985.
46. John M McNamara, Richard F Green, and Ola Olsson. Bayes' theorem and its applications in animal behaviour. *Oikos*, 112(2):243–251, February 2006.