Cogn Affect Behav Neurosci
DOI 10.3758/s13415-015-0350-y

Learning the opportunity cost of time in a patch-foraging task

Sara M. Constantino[1] · **Nathaniel D. Daw**[1,2]

Abstract Although most decision research concerns choice between simultaneously presented options, in many situations options are encountered serially, and the decision is whether to exploit an option or search for a better one. Such problems have a rich history in animal foraging, but we know little about the psychological processes involved. In particular, it is unknown whether learning in these problems is supported by the well-studied neurocomputational mechanisms involved in more conventional tasks. We investigated how humans learn in a foraging task, which requires deciding whether to harvest a depleting resource or switch to a replenished one. The optimal choice (given by the marginal value theorem; MVT) requires comparing the immediate return from harvesting to the opportunity cost of time, which is given by the long-run average reward. In two experiments, we varied opportunity cost across blocks, and subjects adjusted their behavior to blockwise changes in environmental characteristics. We examined how subjects learned their choice strategies by comparing choice adjustments to a learning rule suggested by the MVT (in which the opportunity cost threshold is estimated as an average over previous rewards) and to the predominant incremental-learning theory in neuroscience, *temporal-difference learning* (TD). Trial-by-trial decisions were explained better by the MVT threshold-learning rule. These findings expand on the foraging literature, which has focused on steady-state behavior, by elucidating a computational mechanism for learning in switching tasks that is distinct from those used in traditional tasks, and suggest connections to research on average reward rates in other domains of neuroscience.

Keywords Computational model · Decision making · Dopamine · Reward · Patch foraging · Reinforcement learning

Extensive research in neuroscience, psychology, and economics has concerned choice between a number of simultaneously presented alternatives, as in economic lotteries, "bandit" problems, and choice between concurrent schedules in operant conditioning (Barraclough, Conroy, & Lee, 2004; Baum, 1974; Behrens, Woolrich, Walton, & Rushworth, 2007; Frank, Seeberger, & O'Reilly, 2004; Hampton, Bossaerts, & O'Doherty, 2006; Hare, Schultz, Camerer, O'Doherty, & Rangel, 2011; Herrnstein, 1961, 1991; Krajbich, Armel, & Rangel, 2010; Sugrue, Corrado, & Newsome, 2004; Tom, Fox, Trepel, & Poldrack, 2007). In such problems, attention has centered on a hypothesized neural mechanism for learning an estimate of the values of different options (Schultz, Dayan, & Montague, 1997). More recently, there has been increased interest in neuroscience in a different class of decision problems, in which alternatives are not compared simultaneously but are instead considered serially (Cain, Vul, Clark, & Mitroff, 2012; Hayden, Pearson, & Platt, 2011; Hutchinson, Wilke, & Todd, 2008; Jacobs & Hackenberg, 1996; Kolling, Behrens, Mars, & Rushworth, 2012; Wikenheiser, Stephens, & Redish, 2013). The relevant decision in this class of problems is whether to engage with a current option or search for a better one. Such switching-or-stopping problems arise in many real-world settings, such as employment (whether to accept a job

✉ Sara M. Constantino
sara.constantino@gmail.com

[1] Department of Psychology, New York University, 8th floor, 6 Washington Place, New York, NY 10003, USA

[2] Center for Neural Science, New York University, New York, NY, USA

Published online: 28 April 2015

offer or candidate), Internet search, mate selection, and foraging, and have a rich theoretical and experimental history in ethology, ecology, and economics (Charnov, 1976; Freidin & Kacelnik, 2011; Kacelnik, 1984; McCall, 1970; McNickle & Cahill, 2009; Smith & Winterhalder, 1992; Stephens & Krebs, 1986).

Decisions of this sort pose a dilemma for the widely studied neurocomputational mechanisms of choice, which have largely centered on comparing estimated values across the available alternatives (Rangel, Camerer, & Montague, 2008; Rustichini, 2009), as well as for the related psychological mechanisms of matching and melioration in operant choice (Baum, 1974; Herrnstein, 1961, 1991), which also require balancing time among multiple alternatives. If the alternatives are not directly known at choice time, to what alternative value should the current option be compared when deciding whether to accept it or when to leave it? And how is this more nebulous expected value or aspiration level learned, adjusted, or optimized from previous experience? Predominant theories of choice in the ethological foraging literature have suggested quite different answers to this learning question (Bernstein, Kacelnik, & Krebs, 1988; Charnov, 1976; McNamara & Houston, 1985; Stephens & Krebs, 1986) than would be provided by standard neurocomputational theories of learning (Sutton, 1988; Sutton & Barto, 1998).

The ethology literature has considered a class of stylized switching tasks modeling foraging problems, in which an animal encounters a series of depleting "patches" of resources and must decide whether to spend time exploiting the current patch or instead allocate that time toward seeking a new, replenished one. In such tasks, it has been proved (the marginal value theorem [MVT]; Charnov, 1976) that the reward-rate-maximizing choice of whether to stay or search at each step simply requires comparing the current option's *immediate* reward to a threshold given by the opportunity cost of the time spent engaging with it. The opportunity cost of time is given by the long-run average reward per timestep—a measure of the overall environmental richness that is foregone by harvesting. Whenever you expect to earn less than this quantity, you would be better off doing something else. The MVT thus poses an answer to the question of how to value the nebulous alternative of searching: equate it with the overall average reward rate.

Although the MVT concerns a steady-state choice policy, at the trial-by-trial level it suggests an extremely simple learning rule for deriving that policy by trial and error: Estimate the long-run reward rate by a recency-weighted average of received rewards over time, and use this quantity as a dynamic aspiration level against which to accept or reject the current option (Charnov, 1976; Krebs & Inman, 1992; McNamara & Houston, 1985; Ollason, 1980; Stephens & Krebs, 1986). Variants of this simple model have been suggested in the ethological foraging literature, which has shown that many animals, such as bees and starlings, use dynamically adjusting threshold estimates to inform their search decisions (Cuthill, Kacelnik, Krebs, Haccou, & Iwasa, 1990; Hodges, 1985; Krebs & Inman, 1992; McNamara & Houston, 1985; Ollason, 1980). Although, as we discuss below, this model is quite distinct from common theories of reinforcement learning in the brain, it suggests close connections with a number of other disjoint phenomena in neuroscience that also turn on the average reward as the opportunity cost of time, including response vigor (Beierholm et al., 2013; Guitart-Masip, Beierholm, Dolan, Duzel, & Dayan, 2011; Niv, Daw, & Dayan, 2006; Niv, Daw, Joel, & Dayan, 2007) and temporal discounting (Cools, Nakamura, & Daw, 2011; Daw & Touretzky, 2002; Kacelnik, 1997). These connections relate the simple average reward rate learning in foraging to the suggestion that this rate may be tracked and signaled by tonic levels of the neuromodulator dopamine (Niv et al., 2006; Niv et al., 2007).

An alternative hypothetical approach to learning in the foraging problem is to ignore its special structural features and instead treat the problem of learning the value of searching for a new option as being equivalent to any other case of action-value learning. This requires extending the notion of an action's value to encompass the (nonimmediate) value of seeking new, sequentially encountered options. Accordingly, much research in neuroscience has focused on the temporal-difference (TD) learning algorithm, an incremental update rule that learns, via a series of recursive backups "chaining" rewards to earlier predictors, to estimate the cumulative *future* reward associated with different actions in different circumstances (Sutton, 1988; Sutton & Barto, 1998).

There is considerable neural and behavioral support for TD learning in humans and other animals, notably in recordings from midbrain dopaminergic neurons, whose phasic responses quantitatively match the prediction error signal used in TD learning (Houk, Adams, & Barto, 1995; Montague, Dayan, & Sejnowski, 1996; Schultz et al., 1997). A key feature of the TD rule is that it assigns values to the different options; however, unlike in the classic operant-conditioning theories, this value is defined as the cumulative future return following a choice. In this way, these models extend choice among options to *sequential* decision tasks, in which current choices affect future choices, and the optimal solution requires considering the interdependent consequences of a series of decisions on the cumulative reward ultimately achieved. Importantly for the foraging task—since the value of switching is the deferred consequences of harvesting at subsequent patches—this feature allows for treating the nebulous expected value of switching to an unknown option in the same way that one learns the value of known options. TD learning, applied to these problems, incrementally updates estimates of the cumulative expected future reward associated with each

local stay or switch option and compares these estimates to choose an action.

In this article, we investigate these hypothesized learning rules for serial decision problems by examining human behavior in two sequential patch-foraging experiments. By varying the characteristics of the reward environments (and thus the opportunity cost of time) across blocks, we were able to compare these different accounts of how organisms learn in such tasks. We started by considering the asymptotic strategies. These two learning approaches arrive at the same optimal policy (assuming equivalent choices about parameters such as time discounting or risk sensitivity), and thus make equivalent predictions about asymptotic switching behavior in the different environments. These predictions have been tested extensively in animals (Charnov, 1976; Freidin & Kacelnik, 2011; Kacelnik, 1984; McNickle & Cahill, 2009; Stephens & Krebs, 1986), and more rarely in humans (Hutchinson et al., 2008; Jacobs & Hackenberg, 1996; Kolling et al., 2012). Next we examined the choice adjustments visible in the trial-by-trial dynamics. Because these approaches learn different decision variables (average one-step rewards vs. expected cumulative rewards) by different learning rules, they predict different path dynamics to reach the same asymptotic strategy.

For instance, according to the MVT strategy, an unusually large reward (e.g., a lucky harvest) increases the estimated average reward rate, and thus should directly and immediately raise the leaving threshold, favoring exit. In contrast, TD learns about the values of actions when they are chosen. Here, the long-run value of exiting is estimated indirectly, via chaining. When exit is chosen, the value of exiting is updated according to the expected value of the new tree that one encounters; this value was, in turn, learned from the rewards received from previous stay decisions. The effect of an individual lucky harvest, according to this theory, thus first affects the value of "stay"—increasing the chance of staying the next time that the same tree state is encountered—and only later, through a series of further updates, propagates to increase the value of exiting. Differences of this sort capture the two distinct strategies for how the models search for the optimal policy—by estimating the local cost of time versus the long-run value of switching—and allowed us to compare the two learning models (and some plausible variants) in terms of how well they fit the trial-by-trial fluctuations in choice behavior.

The results suggest that learning in patch-foraging problems implicates a distinct computational mechanism from those that have been successful in more traditional choice problems. This additional mechanism may be broadly applicable to many decisions that can be framed as switching-or-stopping problems and suggests links between choice and other behavioral phenomena, such as response vigor.

Method

Subjects

A total of 52 healthy subjects (age 19–35; 33 female, 19 male) participated in the study: 11 in Experiment 1A, 11 in Experiment 1B, and 30 in Experiment 2. Subjects were paid on the basis of their performance in the task ($15–$22). The study was approved by New York University's Committee on Activities Involving Human Subjects.

A small number of the subjects showed a qualitatively different response strategy (nearly always choosing the "harvest" option, even down to zero apples) and were excluded from all analyses. Specifically, we excluded subjects who had a mean number of harvests per tree that fell more than 2.3 standard deviations (99% quantile) above the group mean; this included one subject from Experiment 1B and three subjects from Experiment 2. Although these subjects were excluded out of caution, their inclusion or exclusion did not appreciably affect the learning analyses depicted in Fig. 3 below. One subject was additionally excluded from Experiment 2, due to a problem with the instructions. Thus, all results reported here concern 11, 10, and 26 subjects from Experiments 1A, 1B, and 2, respectively.

Experimental design and task

Subjects performed a virtual patch-foraging task: a discrete-trial adaptation of a class of tasks from the ecology literature (Charnov, 1976; Cuthill et al., 1990; Hayden et al., 2011; Stephens & Krebs, 1986). On each trial, subjects were presented with a tree and had to decide whether to harvest it for apples or go to a new tree (Fig. 1). Subjects indicated their choice by one of two keypresses when prompted by a response cue. If they decided to harvest the tree, they incurred a short harvest time delay, during which the tree shook and the harvested apples were displayed (as an integer number of apple icons plus a fractional apple icon for the remainder), followed by a response cue. As the subject continued to harvest apples at the same tree, the apples returned were exponentially depleted.

If the subjects chose to go to a new, replenished tree, they incurred a travel time delay, during which the old tree faded and moved off the screen while a new tree moved on to the screen, followed by a response cue. Trees were never revisited; each new tree had never been harvested, and its starting quality was correlated with subsequent outcomes (and thus signaled the quality of the overall tree) in Experiment 2, and uncorrelated in Experiment 1. The total time in the game was fixed, and each choice's reaction time was counted toward the ensuing harvest or travel delay. (Subjects who responded too slowly were penalized by a timeout lasting the length of a single harvest trial.) Thus, subjects visited a different number of trees depending on their

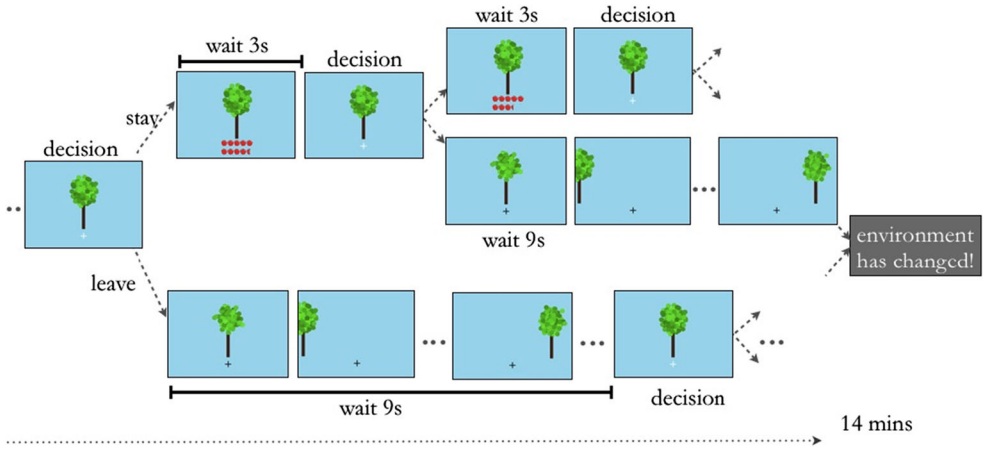

Fig. 1 Task display. Subjects foraged for apples in four 14-min virtual patch-foraging environments. They were presented with a tree and had to decide whether to harvest it for apples and incur a short harvest delay, or move to a new tree and incur a longer travel delay. Harvests at a tree earned apples, albeit at an exponentially decelerating rate. New trees were drawn from a Gaussian distribution, while the environmental richness or opportunity cost of time was varied across blocks by changing the travel time and/or the apple depletion rate. The quality of the tree, depletion rate, and richness of the environment were a priori unknown to the subject (see the Method section for a detailed explanation)

harvest decisions, but apart from timeouts (which occurred on a negligible 1.7% of trials), they were able to influence the reward rate only through their harvest or leave choices, not their reaction times. This design ensured that the optimal choice policy was invariant to the speed of responding.

Subjects experienced four foraging environments in a counterbalanced block design. The decision-relevant parameters that defined an environment were the harvest time, the travel time, the rate at which apples were depleted, and the tree quality distribution. By varying travel time and depletion rate across blocks, we produced environments that differed in terms of richness, with some having a higher achievable average reward rate than others.

The environment changed every 14 min, and this was signaled by a change in background color and a short message. Subjects were not instructed about the type of environment they were entering or what aspects of the environment had changed. They were also not told the form or rate of the depletion or the exact duration of a foraging environment, but they were informed that they would have fixed and equal times in all four environments and that the experiment would last approximately 1 h, that trees could never be revisited, that new trees had never been harvested and were a priori identical, and that harvesting a tree would tend to return fewer apples over time. They were told that they would be paid a half cent for every apple collected and that they should try to collect as many apples as possible.

Each foraging environment was defined by the average initial tree richness S_0, the average depletion rate per harvest κ, the travel time d, and the harvest time h. We denote the state (current expected harvest) of a tree at trial i as s_i.

In Experiments 1A and 1B, each travel decision led to a new tree that was initialized to the same value, $s_i = S_0$. Each harvest decision depleted the tree's state by a fixed multiplicative decay κ, such that $s_{i+1} = \kappa s_i$. The reward r_i returned for harvesting a tree with state s_i was distributed as $\mathcal{N}(s_i, s_i{\cdot}\sigma_r)$. Across environments, the proportional variance of the rewards was chosen such that the probability of the next observed reward falling more than one depletion rate from the current reward was 20% $[P(r_i < \kappa s_i) = .2]$. We varied travel time d or depletion rate κ across blocks in Experiments 1A and 1B, respectively, to create high- and low-average-reward-rate foraging environments. Subjects encountered both environments twice in alternation, with counterbalanced orders ABAB or BABA.

The noise process in Experiment 2 was changed in order to decorrelate a counting-based policy rule from one explicitly based on observed rewards. In this setup, new trees were initialized with a state of variable quality, $s_i{\sim}\mathcal{N}(S_0, \sigma_s)$, and the decay factor applied after each harvest was stochastically drawn, $\kappa_i{\sim}\mathcal{N}(\kappa, \sigma_\kappa)$. This created an effective distribution of different-quality trees with different possible reward paths through the trees. The reward for each harvest was a noiseless reflection of the state of the tree, $r_i = s_i$. We crossed two levels of depletion rate with two levels of travel time, resulting in four environment types. The subjects encountered each environment type once, and the orders were counterbalanced in order to achieve approximately equal numbers of subjects encountering short or long travel delays and steep and shallow depletion rates in the first block.

The parameters for the experiments are shown in Table 1.

The marginal value theorem and optimal behavior

Charnov (1976) proved that the long-run reward-rate optimizing policy for this class of tasks is given by a simple threshold rule. In the context of our discrete-trial version of the task, the

Table 1 Parameter values defining the different environment types in Experiments 1A, 1B, and 2

Experimental Parameters	Experiment 1A: Travel Manipulated		Experiment 1B: Depletion Manipulated	
	Long	Short	Steep	Shallow
h (s)	4.5	4.5	4.5	4.5
d (s)	13.5	4.5	9	9
κ, σ_κ	.85, 0	.85, 0	.68, 0	.89, 0
S_0, σ_S	10, 0	10, 0	10, 0	10, 0
σ_r	.18	.18	.37	.13
	Experiment 2: Travel × Depletion			
	Long–Steep	Long–Shallow	Short–Steep	Short–Shallow
h (s)	3	3	3	3
d (s)	9	9	6	6
κ, σ_κ	.88, .07	.94, .07	.88, .07	.94, .07
S_0, σ_S	10, 1	10, 1	10, 1	10, 1
σ_r	0	0	0	0

optimal policy would be to exit a tree when the expected reward from one more harvest, κs_i, dropped below the opportunity cost of the time that would be spent harvesting it. The opportunity cost of harvesting is the time it takes to harvest, h, times the long-run average reward rate, ρ. Note that in both experiments, the state s_i of a tree was observable: In Experiments 1A and 1B it depended only on the number of times the current tree had been harvested (for a tree harvested n times, $s_i = \kappa^{n-1} S_0$), and in Experiment 2 it was equal to the received reward ($s_i = r_i$).

A sketch of the MVT proof for our setting would be to consider the differential (average reward) Bellman equation for the task (Puterman, 2009). The future value of exiting, $Q^*(s_i, exit) = Q^*(exit)$, is independent of the state; the value of harvesting is

$$Q^*(s_i, \text{harvest}) = -\rho h + \mathbb{E}_{s_{i+1}|s_i}\left[r_{i+1} + \max_{a_{i+1}} Q^*(s_{i+1}, a_{i+1})\right],$$

where s_{i+1} is the state produced by harvesting in state s_i. One should exit when $Q^*(s_i, \text{harvest}) < Q^*(\text{exit})$. Substituting the value of $Q^*(s_i, \text{harvest})$ into this inequality and noting that if *exit* is optimal at i (i.e., if the inequality is satisfied), then the action a_{i+1} that maximizes the continuation value Q^* at the next state would also be *exit* (because trees decay monotonically[1]), results in the following exit rule:

$$-\rho h + \mathbb{E}_{s_{i+1}|s_i}[r_{i+1}] + Q^*(\text{exit}) < Q^*(\text{exit}),$$

which can be simplified to $\mathbb{E}_{s_{i+1}|s_i}[r_{i+1}] < \rho h$, where $\mathbb{E}_{s_{i+1}|s_i}[r_{i+1}] = \kappa s_i$.

[1] This was not strictly true in Experiment 2, in which there was a small chance that $\kappa_i > 1$; however, this did not appreciably impact the optimal policy.

Dependent variable and threshold estimates

Gross, tree-level threshold estimates In Experiments 1A and 1B, the trees' state was discretized and the optimal decision rule could be expressed equivalently as a threshold on the integral number of harvests or on the (real-valued but still quantized) expected reward. To examine behavior and its optimality in units of reward, we estimated a subject's leaving threshold per block as the across-tree average number of expected apples at exit—that is, $\kappa^{n-1} S_0$ —for a tree exited after n harvests. The optimal threshold to which we compared these blockwise estimates was given by the expected number of apples received after n^* harvests, where n^* is the number of harvests that would optimize the total average reward rate.

In Experiment 2, the trees' state was not discretized, so we estimated the subject's exit threshold as the average of the last two rewards r_i and r_{i-1} received before an exit decision. These rewards represent an upper and a lower bound on the (continuously valued) threshold, respectively, since exiting at i implies that κr_i is lower than the subject's threshold, and not exiting in the preceding decision implies that κr_{i-1} is greater. The corresponding optimal threshold is $\rho h/\kappa$, since $\mathbb{E}_i[r_{i+1}] = \kappa s_i$.

For Fig. 2c and f in the Results (the scatterplots), to visualize compliance with the MVT threshold condition $\kappa s_i = \rho h$, we plotted for each subject and block the total number of apples obtained in the block divided by the total length of the block in periods h (ρh), against the expected next reward at exit (κ times s_i, where the threshold s_i was taken as the average of the rewards r_{i-1} and r_i preceding each exit, further averaged over all trees in the block, since the MVT was expressed for both experiments in terms of the continuously valued threshold).

Trial-by-trial choices We can also model the trial-by-trial decisions using a logistic regression. In particular, we assumed

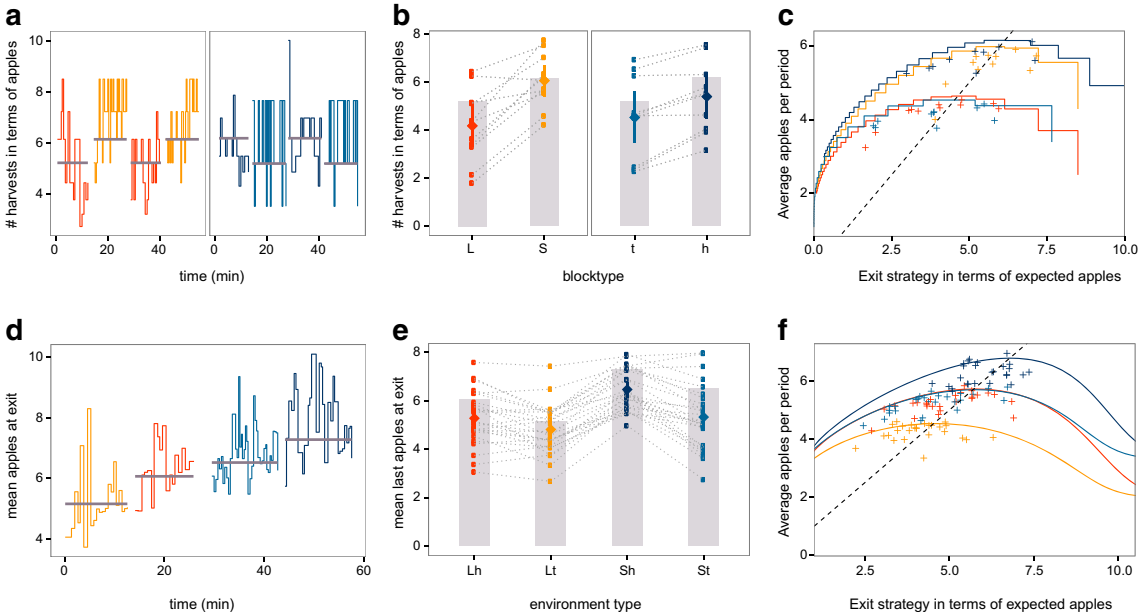

Fig. 2 Foraging behavior: Behavioral results compared to the optimal (ideal observer) performance in the task. (Top) Experiment 1A (travel time varied: L = long, S = short) and 1B (depletion rate varied: t = steep, h = shallow). (Bottom) Experiment 2 (Lh = long–shallow, Lt = long–steep, Sh = short–shallow, St = short–steep). (a, d) Example subject tree-by-tree exit points over time in each experiment. Colors indicate different environments, and gray lines indicate the optimal exit thresholds. (b, e) Group performance by blocks. The heights of the gray bars indicate optimal thresholds; open circles connected by gray lines are individual-subject mean exit thresholds and adjustments across environments; and filled diamonds are mean exit thresholds, with 95% confidence intervals. (c, f) Colored curves show the achievable average rewards per period for any given threshold policy in the different environments. Pluses are individual subjects' mean exit thresholds, and dashed lines indicate the marginal value theorem rule—points at which the average reward rate is equal to the expected reward; these lines intersect the colored curves at the optimal exit thresholds

that subjects made noisy harvest-or-exit choices a_i according to a logistic choice rule of the form

$$P(a_i = \text{stay}) = \frac{1}{1 + \exp\left(-(c_k + \beta[K_k S_i])\right)}$$

for some block-specific threshold (intercept) c_k, trial- and block-specific explanatory variable $\kappa_k s_i$, and logistic weight (akin to a softmax inverse temperature) β. For a given environment with an average reward rate ρ_k, the MVT predicts that $c_k = -\beta\rho_k h$, resulting in a stochastic (logistic decision noise) version of the optimal policy. We estimated the parameters of this model (the thresholds c_k and temperature β) for each subject. The dependent variable was the binary choice a_i on each trial, and the explanatory variables were the reward expected from harvesting on each trial, computed using the block-specific depletion rate, $\kappa_k s_i$, and four block-specific indicators corresponding to c_k.

This model defines a likelihood function over the sequence of choices, allowing us to use Bayesian model comparison to contrast different explanatory variables corresponding to different rules for estimating the state s_i of the tree. In particular, to compare counting- and reward-based strategies in Experiment 2, we tested whether the choices were better fit by assuming $s_i = \kappa_k^{n-1} S_0$ (the number of harvests, expressed in units of reward) or $s_i = r_i$ (the most recently received reward, which was the actual state of the tree in Exp. 2). We used the Bayesian information criterion (BIC) to estimate the marginal likelihood of each model given the data and submitted these values to the spm_BMS function from SPM version 8 to compute exceedance probabilities (Stephan, Penny, Daunizeau, Moran, & Friston, 2009).

Learning

MVT learning model The optimal policy from the MVT is to harvest whenever the immediate expected reward is greater than the average reward: $\kappa s_i \geq \rho h$. Note that the appearance of the fixed harvest time h and the expected depletion rate κ in this equation are artifacts of the discrete-trial/discrete-time structure of our task; they do not appear in the MVT for a more classic continuous-time foraging task, like fishing, in which the agent can exit a patch at any instant. Thus, to execute this policy exactly, the subject must have estimates of both quantities. We assumed that the fixed harvest time h (which was easily observed) was known, or equivalently that subjects learned ρ in units of reward per harvest period, and that subjects estimated κ with a simple within-block running average of experienced depletion rates over adjacent harvests.

The MVT motivates a simple threshold-learning model that requires specifying a learning rule for ρ and deciding whether to harvest by comparing the two sides of the MVT equation, κs_i and ρh (Table 2). We implemented this comparison

Table 2 Marginal value theorem learning update rule

Parameters: α, c, β

$P(a_i = \text{harvest}) = 1/\{1 + \exp[-c - \beta(\kappa_k s_i - \rho_i h)]\}$

$\delta_i \leftarrow r_i/\tau_i - \rho_i$

$\rho_{i+1} \leftarrow \rho_i + [1-(1-\alpha)^{\tau_i}] \cdot \delta_i$

stochastically using a logistic (softmax) rule on the difference $\kappa s_i - \rho h$, with inverse temperature β and intercept c. The decision maker's trial-by-trial estimate of ρ was constructed by a standard delta rule with learning rate α, taking into account the time $\tau_i = h$ or d of each choice step:

$$\rho_i = (1-a)^{\tau_i}\frac{r_i}{\tau_i} + (1-(1-a)^{\tau_i})\,\rho_{i-1}.$$

At the beginning of the experiment, the depletion rate was initialized to 1 and the average reward rate ρ_{init} to the average across the experiments of the reward rates attained by an ideal observer. Fitting the initial condition as an additional free parameter, in both the MVT and TD models, did not appreciably affect the results; these results are not shown. In subsequent environments, the initial average reward rate estimate was taken as the last ρ estimate in the previous environment.

TD learning model The TD algorithm learns the expected future discounted reward of continuing at each state $Q(s, \text{harvest})$ and a value of exiting $Q(s, \text{exit}) = Q(\text{exit})$ that is not state-specific (Table 3). (Note that maintaining separate, state-specific values for exiting would only slow the learning of the task and accentuate the underperformance of TD. Furthermore, this assumption seems natural since subjects were told they could exit a tree on any trial and that this would lead to a new, randomly drawn tree.) The choice is taken to be logistic (softmax) in the difference between these values, with inverse temperature β and intercept c. The action values are incrementally updated by temporal-difference learning.

This model has four parameters—the same three as MVT, and an additional discount factor γ, which ensures that the infinite horizon cumulative future rewards are finite. For states s_i, we used the actual state of the tree as defined by the experiment specifications. Thus, the state s_i was discrete in Experiment 1 and was given by the number n of previous harvests at the current tree, whereas the state in Experiment 2 was the most recently observed reward r_i, a continuous variable. In order to implement the update rule in this case, we approximated the value function by a linear combination of functions that were constant over predefined bins. To match the discretization of the state to the true, exponentially depleting dynamics of the trees, the bins were logarithmically spaced and the width was chosen so that on average, each subsequent harvest would fall in a subsequent bin. More precisely, if b_j and b_{j+1} are the lower and upper bounds of the j th bin, respectively, and $\overline{\kappa}$ is the average depletion across the environments, the bins were spaced according to $\log(b_{j+1}) - \log(b_j) = -\log\overline{\kappa}$. At the beginning of the experiment, the starting values for $Q(s, \text{harvest})$ and $Q(\text{exit})$ were initialized with a constant for all s, equal to the discounted sum of the rewards associated with earning ρ_{init} on average: $\frac{\rho_{\text{init}}}{1-\gamma}$. This starts the algorithm in the approximate range of the correct Q values, given a discount factor of γ and the same initial guess for the long-run reward rate as the MVT learning algorithm. In subsequent environments, the Q values are initialized to the Q values at the end of the previous block.

Note that both models were given explicit knowledge of the delays (h, t).

We fit the free parameters to each subject's choices separately using maximum likelihood, and used these likelihoods to compute per-subject BIC scores as a measure of model fit. We used spm_BMS to compare the models.

Variants of TD In addition to the standard TD learning algorithm, we also compared the performance of two TD variants: TD(λ), which improves the efficiency of the algorithm by allowing faster back-propagation, and R-learning, an undiscounted, average-reward reinforcement learning model (Sutton, 1998; Schwartz, 1993). TD(λ) allows learning to immediately back-propagate not just one step, but through many preceding states by the introduction of an exponentially decaying eligibility trace (Table 4). The decay rate of the eligibility trace is governed by an additional free parameter λ, where $\lambda = 0$ is the one-step backup of the standard TD model presented above.

R-learning aims to optimize the average reward per time step rather than the cumulative discounted reward, and so asymptotically implements the same strategy as MVT. It produces this behavior in a different manner, however, because it

Table 3 Temporal-difference (TD) learning update rule

Parameters: α, γ, c, β

$P(a_i = \text{harvest}) = 1/(1 + \exp\{-c - \beta[Q_i(s_i, \text{harvest}) - Q_i(\text{exit})]\})$

$D_i \sim \text{Bernoulli}[P(a_i)]$

$\delta_i \leftarrow r_i + \gamma^{\tau_i}\,[D_i \bullet Q_i(s_i) + (1 - D_i) \bullet Q_i(\text{exit})] - Q_i(s_{i-1}, a_{i-1})$

$Q_i(s_{i-1}, a_{i-1}) \leftarrow Q_{i+1}(s_{i-1}, a_{i-1}) + \alpha \bullet \delta_i$

Table 4 TD(λ) update rule

Parameters: α_1, λ, γ, c, β

$P(a_i = \text{harvest}) = 1/(1 + \exp\{-c - \beta[Q_i(s_i, \text{harvest}) - Q_i(\text{exit})]\})$

$D_i \sim \text{Bernoulli}\,[P(a_i)]$

$\delta_i \leftarrow r_i + \gamma^{\tau_i}\,[D_i \bullet Q_i(s_i) + (1 - D_i) \bullet Q_i(\text{exit})] - Q_i(s_{i-1}, a_{i-1})$

$E_i(s_i, a_i) \leftarrow 1$

$\forall s, a\; E_{i+1}(s,a) \leftarrow E_i(s,a) + \lambda \bullet \gamma^{\tau_i}\, E_i(s,a)$

$\forall s, a\; Q_{i+1}(s,a) \leftarrow Q_i(s,a) + \alpha_1 \bullet E_i(s,a) \bullet \delta_i$

uses a TD algorithm to learn a different notion of the state–action value appropriate to this goal: the expected *undiscounted* cumulative *average-adjusted* reward. In this model, rewards are measured relative to the long-run average reward per timestep, a stateless quantity that is separately learned according to an additional learning rate parameter α_2. The full R-learning algorithm is presented in Table 5.

In both of these variants, the state-space and initialization were the same as in standard TD. In R-learning, the average reward term was initialized as in MVT learning.

Overharvesting

The intercept in the learning models above can capture a fixed policy bias: early or late exiting relative to the optimal exit threshold. However, plausible factors that might underlie the observed tendency toward overharvesting could include temporal discounting and decreasing marginal utility for money (one way of parameterizing risk-sensitive preferences). Decreasing marginal utility (sublinear increases in the value of each additional unit of money that produce risk-averse choices in gambling tasks; Bernoulli, 1954) results in increased harvesting relative to a linear utility, since the larger rewards at a replenished tree are worth proportionally less. To test whether decreasing marginal utility captured this tendency, we fitted the MVT model with a power-function utility on rewards $[U(r_i) = r_i^{\eta}]$, which affected both the expected next reward and the average reward rate, and looked at the effect this had on the intercept term estimate using an across-subjects *t* test. To examine discounting, we were constrained to the TD algorithm, since there is no discounted equivalent of the MVT model. We again ran an across-subjects *t* test on the intercept estimate to test whether the deviations from optimal were fully captured by the discount parameter.

Results

Across two experiments, we presented human subjects (n = 47) with foraging environments that varied in richness, as measured by the maximally achievable average reward rate, across a series of blocks (see Fig. 1 and the Methods section). At each step of the task, subjects were presented with apples earned from the last harvest at the current tree and had to decide whether to continue harvesting or switch to a new, randomly drawn tree. Harvesting caused trees to deplete—each successive harvest earned fewer apples, on average—but traveling to a new, replenished tree cost additional time. Apples were converted to money at the end of the experiment.

Environmental richness was manipulated by changing the travel delay between trees (Exp. 1A), the tree depletion rate (Exp. 1B), or both factors simultaneously (Exp. 2). All else held constant, a longer travel delay or a steeper depletion rate reduces the rate at which apples can be earned. This reduces the opportunity cost of time spent harvesting and leads an ideal forager to harvest a tree down to a lower number of apples. Accordingly, the MVT states that the optimal policy for this class of tasks is to abandon a tree when the expected marginal intake from one more harvest falls below the overall average reward rate of the environment (see the Methods section; Charnov, 1976; Stephens & Krebs, 1986). The characteristics and richness of the environment were a priori unknown by the subject and had to be learned through experience.

Table 5 R-learning update rule

Parameters: α_1, α_2, γ, c, β

$P(a_i = \text{harvest}) = 1/(1 + \exp\{-c - \beta[Q_i(s_i, \text{harvest}) - Q_i(\text{exit})]\})$

$D_i \sim \text{Bernoulli}[P(a_i)]$

$\delta_i \leftarrow r_i - \rho_i \bullet \tau_i + D_i \bullet Q_i(s_i, \text{havest}) + (1 - D_i) \bullet Q_i(\text{exit}) - Q_i(s_{i-1}, a_{i-1})$

$Q_i(s_{i-1}, a_{i-1}) \leftarrow Q_{i+1}(s_{i-1}, a_{i-1}) + \alpha_1 \bullet \delta_i$

$\rho_{i+1} \leftarrow \rho_i + \alpha_2 \bullet \delta_i$

Experiment 1

Each subject completed four 14-min task blocks that alternated between high and low travel delay (Exp. 1A, n = 11) or depletion rate (Exp. 1B, n = 10), in counterbalanced order. The number of apples earned for each harvest was drawn randomly around an underlying average, which decayed exponentially with harvests in a manner that was deterministic and identical across trees. Thus, the expected reward for a harvest was a function of the number of times the tree had already been harvested. Subjects made an average of 611 harvest-or-exit decisions and visited an average of 103 trees.

First, we asked whether subjects' overall strategies were modulated by environmental richness in the manner predicted by the optimal analysis. The optimal rule compares the expected reward for a harvest to a threshold; in Experiments 1A and 1B, the expected reward from harvesting was a function of the number of times the current tree had been harvested. Therefore, as an indicator of subjects' exit thresholds, we considered the number of times that each tree had been harvested and expressed this quantity in units of the equivalent expected apple reward at the final harvest (Fig. 2b). Comparing the empirical thresholds to the values predicted under optimal switching, we found a tendency across all conditions to harvest longer (i.e., to exhibit a lower exit threshold) than was optimal, but this trend was only significant in the long-travel-time condition ($t_9 = -2.3, p = .045$; all other environments, $p > .1$). Next, comparing thresholds within subjects and across blocks to examine whether, notwithstanding any over- or underharvesting, subjects adjusted to changing reward rates in the optimally predicted direction, we found that

subjects indeed harvested to lower thresholds in lower-quality environments (paired t tests across subjects: $t_{10} = 5.7, p < .001$, for travel delay; $t_9 = 5.1, p < .001$, for depletion), with almost every subject adjusting in the predicted direction. These results suggest that subjects behaved in a way qualitatively consistent with the MVT, though potentially with a slight bias to overstay.

Another way to visualize compliance with the optimal policy is to consider overall earnings. The points in Fig. 2c. show the obtained average reward for each subject and environment pair plotted against the empirical average exit thresholds. The dotted line shows the MVT prediction (where the threshold equals the average earnings), and the step functions show the achievable average rewards per period as a function of different possible fixed exit thresholds in each environment. Subjects clustered around the MVT threshold, but with a tendency to lie to the left (reflecting overharvesting) and below the reward expected if the mean exit strategy had been executed consistently (implying variation in the strategy over the block, which is apparent in the individual-subject data shown in Fig. 2a). Accordingly, subjects earned on average \$17.9 ± \$1.4 (mean ± *SD*), which represents a 10% loss on optimal earnings. In addition to comparing earnings to an upper bound given by the optimal strategy, it can be useful to situate them relative to a lower benchmark, given by random stay-or-exit responding according to a coin flip at each decision. To be conservative, we defined this "random" policy generously by optimizing the weight on the coin to maximize earnings in each environment. The best constant-hazard-rate policy deviated from optimal by –33%.

Learning and trial-by-trial choices

Next, rather than considering overall blockwise measures of thresholds and earnings, we examined trial-by-trial choices using models to explain the series of stay-or-exit decisions in terms of time series of candidate decision variables. Since subjects were not explicitly informed about the parameters of the environments, their ability to systematically adjust their exit thresholds between environments must reflect learning. Choice adjustments due to learning might also help explain the substantial within-block dynamics of the exit strategies (as in the individual raw data shown in Fig. 2a). We compared the fits of two qualitatively different candidate models.

The first was a learning rule that has been suggested in the foraging literature and is based directly on the MVT policy: a stateless model in which subjects simply estimate the long-run average reward per timestep by taking a recency-weighted average over observed rewards. An exit decision occurs when the expected reward for harvesting drops below this opportunity cost estimate (Krebs & Inman, 1992; McNamara & Houston, 1985). An alternative model, the *Q-learning algorithm*, instead learns a *state–action value function*, which represents the cumulative future expected value of harvesting or leaving a tree at each state (where the state is given by the number of previous harvests on a tree). These action values are compared in each state to reach a decision (Watkins, 1989).

Although TD methods like Q learning are more general and learn the full value function, which is required for optimal performance in many sequential decision tasks, MVT learns a summary variable that is sufficient for optimal choice in this class of tasks (see the Method section). Learning these different decision variables results in different trial-by-trial choice adjustments. Generally, the two algorithms differ in how individual experiences affect future choices over the entire state space of the task. Whereas the MVT updates a single threshold (and the behavioral policy) at each observation, TD maintains a set of state–action values specifying the appropriate behavior for each step in the tree. These values are more difficult to learn, both because there are many of them and because they represent a prediction about long-run cumulative future reward, which is updated only locally via a process of chaining or "bootstrapping" received rewards and value estimates across a series of successively encountered states (Sutton, 1988; Sutton & Barto, 1998).

Because both models specify different, complex trial-by-trial relationships between received rewards and subsequent choices, we used model comparison to determine which provided a better fit to the data. We computed the two learning models' fits to each subject's choice data using the BIC, having optimized the free parameters using maximum likelihood, and then compared the fits at the population level using Stephan et al.'s (2009) group model comparison technique. This model comparison method allows for the possibility that the true model varies across subjects and estimates the proportion of subjects in the population expressing either model. This analysis showed overwhelming evidence in favor of the simpler MVT learning rule (Fig. 3, left; with an expected frequency of .83 and an exceedance probability, or posterior probability, that it was the more common model of .999). Note that the MVT model optimizes undiscounted reward rate, whereas TD optimizes cumulative exponentially discounted reward with a free discount rate. However, the model fit difference was not simply due to the penalty for the inclusion of the additional discount rate parameter in TD, since the BIC penalty for a single parameter is small with respect to the differences in model fit. Neglecting the penalty for the extra parameter still produced an exceedance probability of .999. The model differences remained significant even when the likelihood was computed without the initial 20% of observations, suggesting that the differences were not due to starting values (exceedence probability of .999), and when removing the first half of each block (exceedence probability of .993), suggesting that MVT does better not only at the environment transitions, but also at explaining the within-environment dynamics.

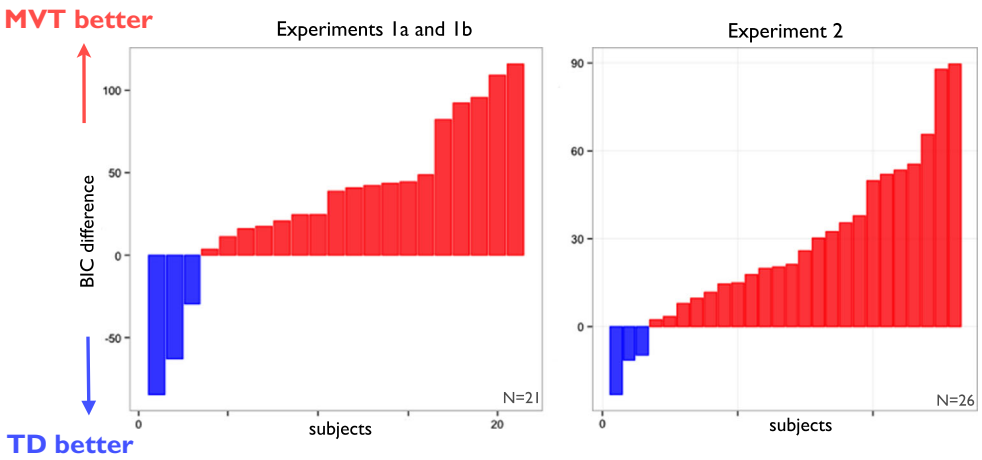

Fig. 3 Model comparison: Approximate log Bayes factors (difference in Bayesian information criterion scores) favoring marginal value theorem (MVT) versus temporal-difference (TD) learning models, shown for each subject separately for Experiments 1A and 1B (left) and 2 (right)

In order to assess the robustness of our results, we also compared the MVT threshold learning model to two other variants of TD learning. First, motivated by the possibility that TD's poor performance was due to slow learning over the state space, we considered TD(λ), a generalization of the standard TD model that allows for faster, nonlocal back-propagation of information across multiple states. Second, we considered R-learning (Schwartz, 1993), an average-reward reinforcement-learning algorithm that updates state–action values relative to a stateless average-reward term that is updated at every step. We reasoned that R-learning might perform more similarly to the MVT rule, because it optimizes the same objective as MVT and does so, in part, by learning an additional average-reward rate term similar to the MVT's decision variable. These analyses again showed overwhelming evidence in favor of the simpler MVT learning rule as compared to TD(λ) (expected frequency of .78 and exceedence probability of .997) and R-learning (expected frequency of .87 and exceedence probability of .999), suggesting that the simple, stateless, threshold-learning rule outperforms a broad class of TD models in this task.

Overharvesting

The above learning models each contain an intercept parameter that encodes any constant bias toward or away from harvesting, in addition to the effect of the learned decision variables. A tendency toward overharvesting, which was noticeable but for the most part nonsignificant in the cruder blockwise analyses above, should be visible here as a negative estimate for the intercept term in the MVT learning model. Indeed, the estimated intercept was significantly negative in the MVT model ($t_{20} = -4.87, p < .001$, across subjects for MVT), demonstrating a bias toward overharvesting.

Several factors might jointly contribute to this overharvesting tendency. First, any persistent behavioral variability or deviations from consistently executing the optimal policy would reduce the obtained average reward and imply a lower steady-state opportunity cost of time and exit threshold. In other words, the extent to which a subject's actual long-run earnings fall below the rewards expected for consistently executing a strategy (the stepped functions in Fig. 2c) implies variability around her average strategy. Her best response to the resulting effective reward environment would be to aim to harvest longer than would otherwise be optimal (dotted line). Two such sources of variability are already accounted for in the learning model estimated above: trial-to-trial threshold variation (due to learning) and additional decision stochasticity (captured by the logistic choice rule). The finding that the intercept is still significantly negative demonstrates residual overharvesting, even after taking these sources of variability into account.

The MVT-predicted policy maximizes reward rate, but subjects may differ from this objective in their preferences over delay (time discount factor, as assumed in TD) or amount (a nonlinear marginal utility of money, a standard way to parameterize risk-sensitive preferences; Bernoulli, 1954), which would also contribute to overharvesting. For a subject with decreasing marginal utility (i.e., one for whom \$10 was worth less than twice \$5, which in expected utility is functionally equivalent to risk aversion), exiting later would be predicted because the marginal value of a replenished tree would be reduced. The reverse pattern would be predicted for subjects with increasing marginal utility (i.e., risk-seeking preferences). We reestimated the MVT model with an additional risk sensitivity parameter (i.e., curvature in the function mapping money to utility). With this parameter, the intercept was no longer significant ($t_{20} = -1.33, p = .2$, across subjects), suggesting that overharvesting was substantially accounted for by risk aversion.

Additionally, a decision-maker who discounts future rewards would exit a patch later, because the value of leaving

would be discounted by the travel delay, predicting discount-dependent overharvesting. Indeed, the TD model optimized cumulative exponentially discounted reward and included a free parameter controlling the sharpness of discounting. The intercept in the TD model was not significantly different from zero ($t_{20} = -1.03, p = .317$, across subjects), indicating that exponential discounting can, on average, also account for overharvesting. (However, note that some individual subjects *underharvest*. Underharvesting cannot be explained by conventional time discounting, though it could in principle be taken as a preference for later rewards—for example, "savoring" delays. This is empirically unusual and also computationally problematic, since the infinite horizon value function then diverges. Such behavior can be explained more straightforwardly by convex utility curvature, which is less exotic.)

Experiment 2

A disadvantage of the design in Experiment 1 was that, because the expected reward decayed deterministically with harvests, the reward-thresholding strategy predicted by the MVT was equivalent to a counting strategy: harvesting a fixed number of times. Therefore, we conducted an additional study, in which the initial qualities of new trees and the rates of depletion following each harvest were drawn randomly. These features would require the ideal agent to monitor the obtained rewards rather than simply to count harvests. In this version, the obtained rewards at each step noiselessly reflected the current state of the tree, and the expected reward from harvesting could thus be constructed by depleting the last observed reward. The optimal policy thus involved directly comparing the obtained reward to a threshold at which a tree should be exited, regardless of its starting quality or the number of preceding harvests.

In this experiment, we simultaneously varied both depletion rate and travel time within subjects, crossing two levels of each to create four environment types with three levels of achievable average reward rates: one high, two medium, and one low. Subjects made an average of 913 harvest-or-exit decisions and visited an average of 95 trees over the experiment.

The results of Experiment 2 echoed those of Experiment 1 and verified that the behaviors studied there extended to this more complex setting. Exit thresholds were estimated as the averages of rewards obtained before and at each exit in each condition, since these rewards should bracket the true threshold (Fig. 2e). These were significantly lower than optimal for three of the four conditions ($< 3.6, ps < .002$; with a trend in the long–steep block, $t_{25} = -1.7, p = .09$), again demonstrating an overharvesting tendency. Despite this tendency, most individual subjects adjusted their thresholds in the expected direction in response to blockwise changes in the average reward rate, a consistency that was reflected in a repeated measures analysis of variance as main effects of depletion, $F(1, 25) = 82, p < .001$, and travel time, $F(1, 25) = 42.1, p < .001$.

A quantitative analysis of the deviations from optimal showed that subjects earned on average \$18.7 ± \$1.4 (mean ± *SD*), which represents a 9% loss relative to optimal earnings. For comparison, the best constant-hazard-rate policy, optimized to each environment, deviated from optimal earnings by –25%. When earnings were plotted against thresholds, strategies again clustered around the MVT's predictions (Fig. 2f), albeit with a tendency to fall to the left of the line, indicating overharvesting.

Trial-by-trial choices

Unlike Experiment 1, this experiment allowed us to distinguish between an exit policy based on thresholding observed rewards, as the MVT predicts, and a simple count of the number of times that a tree had been harvested. We investigated this question by fitting two logistic regression models to subjects' trial-by-trial stay-or-exit decisions, differing only in whether the main explanatory variable was the preceding reward or the preceding number of harvests. The analysis showed strong evidence in favor of the optimal reward-based strategy (Stephan et al.'s, 2009, Bayesian model selection: expected frequency .93, exceedance probability .999).

Learning

Finally, individual behavioral traces again demonstrate rapid threshold adjustments between and within environments (Fig. 2d). We compared the TD and MVT incremental-learning model fits to the trial-by-trial decisions and found evidence in favor of the simpler MVT learning rule (Fig. 3, right; expected frequency .85, exceedance probability .999). Robustness checks showed that the MVT model was still favored, even without including the initial 20% of observations, suggesting that the differences were not due to starting values (exceedence probability of .823), and when removing the first half of each block, suggesting that MVT does better not only at the environment transitions, but also at explaining the within-environment dynamics (exceedence probability of .999).

A comparison between MVT and the two TD variants mentioned above suggests that this simple threshold-learning model better describes decisions in this type of task than a broad class of reinforcement-learning models. A Bayesian model comparison showed an expected frequency of .93 and an exceedence probability of .999 in favor of MVT over R-learning, and an expected frequency of .73 and exceedence probability of .9941 when compared to TD(λ).

Overharvesting

The results of Experiment 2 again revealed a tendency to overharvest relative to the reward-maximizing policy. This was reflected in the MVT learning model fits, which included a negative intercept estimate (bias toward harvesting) that was significant across subjects ($t_{20} = -7.48, p < .001$). This result again suggests that the tendency toward overharvesting was not fully captured by decision noise or threshold variation (due to learning). We again investigated whether this tendency could be explained by delay discounting or decreasing marginal utility (e.g., risk sensitivity) by looking at the effects of these additional preference parameters in TD and MVT, respectively. As in Experiment 1, the TD model incorporated a free time-discounting parameter, but in this case still had a significantly negative intercept, suggesting that time discounting cannot by itself explain the result ($t_{20} = -4.47, p < .001$, across subjects). However, when we added an additional parameter controlling risk sensitivity (value curvature) to the MVT model, the mean intercept was close to zero and no longer significant across subjects ($t_{20} = -1.71, p = .10$), suggesting that overharvesting may be substantially accounted for by risk sensitivity.

Discussion

Following on a long tradition of work in neuroscience and psychology about decision making in tasks in which multiple options are presented simultaneously, there has been a recent interest in a different class of tasks, largely inspired by the ethological foraging literature: switching-or-stopping tasks, in which options are presented serially and the choice is whether to accept the current option or search for a better one (Cain et al., 2012; Hayden et al., 2011; Hutchinson et al., 2008; Jacobs & Hackenberg, 1996; Kolling et al., 2012; Wikenheiser et al., 2013). Although most previous work in this class of foraging task has considered steady-state choices, these tasks pose a particularly interesting problem for learning, which we consider here: how to learn the (deferred and implicit) value of the alternative option.

Consistent with a body of ethological work in animals (Charnov, 1976; Freidin & Kacelnik, 2011; Hayden et al., 2011; Kacelnik, 1984; McNickle & Cahill, 2009; Stephens & Krebs, 1986) and a smaller literature in humans (Hutchinson et al., 2008; Jacobs & Hackenberg, 1996; Kolling et al., 2012; McCall, 1970; Smith & Winterhalder, 1992), we showed across two experiments that humans consistently adjust their foraging choices to manipulations of the richness of the environment in the directions predicted by the optimal analysis. However, both experiments revealed a tendency to harvest a patch longer than would be optimal for maximizing reward rate, which is consistent with previous findings in patch-foraging experiments (Hutchinson et al., 2008). In our data, this apparent suboptimality disappeared if we assumed that subjects optimized a currency with a nonlinear marginal utility for money (e.g., risk sensitivity). Thus, the choices are consistent with the MVT, when stated in terms of utility, rather than the raw monetary amount. A less satisfactory explanation for overharvesting (e.g., because it failed to fully account for the overharvesting in Exp. 2) is that subjects discounted delayed rewards. Had we needed to appeal to time discounting, this would have been harder to reconcile with the MVT, which concerns the undiscounted reward rate. A more direct test of either explanation would require assessing risk sensitivity (or time discounting) independently, in order to investigate whether those estimates predict a subject's overharvesting. It is also possible that many factors may contribute jointly to this tendency.

With respect to learning, subjects' trial-by-trial choice adjustments are consistent with a simple learning rule in which they track the average reward per timestep and use it as a threshold or aspiration level against which to compare their immediate gains (Krebs & Inman, 1992; McNamara & Houston, 1985). This learning rule is notably different from the reinforcement-learning rules more often studied in neuroscience, which solve sequential decision tasks by estimating cumulative long-term rewards (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997).

In analyzing learning, we relied on formal model comparison to examine which of the two learning accounts better explained the observed relationship between outcomes and choices. This was necessitated because the correlated structure of rewards over time in the task—especially in Experiment 2—made it difficult to find a simple, local feature of the reward–choice relationship (such as the relationship between individual rewards and subsequent exits) that would unambiguously capture the qualitative difference between the models. Thus, we compared how well each model explained the full joint probability of all the choices, conditional on all the rewards. However, the disadvantages of this approach, interpretationally, are that it is relatively opaque to what features of the data drive the difference in model fit and that the conclusion is also specific to the particular models tested. The possibility therefore exists that some variant model not explored, and in particular some version of TD, might outperform the models considered here. We addressed this problem by considering what we take to be a canonical representative of the set of TD models, further improved with task-specific knowledge (e.g., a stateless value of exiting), and by additionally considering other TD variants, which together represent a reasonably large class of TD models. The magnitudes of the differences in fit and their robustness to these algorithmic variations suggest that MVT learning is a better model of subjects' trial-by-trial decisions than are a broad range of plausible TD models.

One important dimension along which there are many possible variants of TD learning is how the algorithm construes the space of possible states of the task. This was less of an issue in Experiment 1, in which the objective state space of the task was small, discrete, and well-defined and did not require any generalizing function approximator. Two other TD variants that we considered were TD(λ), which increases the efficiency of the algorithm by introducing eligibility traces that allow for faster back-propagation of information across states (making the particular representation of states less important), and R-learning, a TD algorithm that learns state–action values relative to an average-reward term equivalent to that used in the MVT. Since this model maximizes the same currency as MVT learning, it rules out the possibility that the difference in fits between MVT and the other TD variants was driven by the slightly different objective (cumulative discounted reward) optimized by the latter.

Another difference between the models is that the MVT rule, as we have simulated it, relies on learning the depletion rate of the trees, and (unlike TD) is given the knowledge that this process takes an exponential form.[2] Although it is difficult to see how to imbue TD with equivalent knowledge (since its action values are derived in a complex way from the one-step rewards and continuation values), we believe that this disadvantage does not explain away the difference in fits between the models. In particular, although we included this factor in the MVT for consistency with the ideal observer (where this term occurs as an algebraic infelicity, owing to the discrete trial setup of our task), leaving it out altogether (i.e., omitting κ) from the modeled MVT choice rule, and thus removing any built-in knowledge about the form of the depletion, results only in a slight decision bias relative to optimality and does not change the dynamics of the algorithm. Accordingly, repeating the model comparison analyses omitting the depletion rate term from the decision rule would not appreciably change the model comparison (these results are not presented).

Our results suggest an important role for the average reward rate in guiding choice in serial switching-or-stopping tasks. These results tie decision-making in this class of tasks to a diverse set of behavioral phenomena and neural mechanisms that have been previously associated with average reward rates. Notably, the average reward rate has arisen in the analysis of a number of other seemingly disparate behaviors, including behavioral vigor, risk sensitivity, labor–leisure trade-offs, self-control, and time discounting (Cools et al., 2011; Daw & Touretzky, 2002; Gallistel & Gibbon, 2000; Guitart-Masip et al., 2011; Kacelnik, 1997; Keramati, Dezfouli, & Piray, 2011; Kurzban, Duckworth, Kable, & Myers, 2012; Niv et al., 2006; Niv et al., 2007; Niyogi et al., 2014). The notion of the average reward rate as the opportunity cost of time links these seemingly disparate domains: It indicates, for instance, the potential reward foregone by behaving less vigorously, by waiting for a delayed reward in a discounting task, or, in the case of foraging, by spending more time with the current option rather than searching for a better one (Niv et al., 2006; Niv et al., 2007). A very similar comparison—of current rewards to the long-run average—has also been suggested to govern the explore–exploit trade-off (Aston-Jones & Cohen, 2005; Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010). Indeed, in the case of vigor, it has been shown that trial-by-trial fluctuations in behavioral vigor, as assessed by reaction times, are consistent with modulation by an ongoing estimation of the average reward rate (Beierholm et al., 2013; Guitart-Masip et al., 2011). The present results extend this rate-sensitive adjustment from modulations of arousal to discrete foraging decisions.

That such diverse phenomena implicate the same decision variable invites the possibility that they might share neural mechanisms. In particular, whereas TD learning about cumulative future action values is widely linked to *phasic* dopamine signaling (Houk et al., 1995; Montague et al., 1996; Schultz et al., 1997), it has been argued that longer-timescale ("tonic") extracellular dopamine levels might be a neural substrate for tracking average rewards and for controlling the behaviors that depend on them (Beierholm et al., 2013; Cools, 2008; Niv et al., 2006; Niv et al., 2007). Accordingly, dopamine is well known to affect response vigor and time discounting, and we hypothesize that it may play a similar role in the present task (Beierholm et al., 2013; Robbins & Everitt, 2007; Salamone, 1988). However, whereas dopaminergic involvement in response vigor appears to focus on the nucleus accumbens (Lex & Hauber, 2008; Salamone, 1988), foraging decisions, switching to nondefault courses of action, average-reward tracking, cognitive control, and decisions involving costs all appear to involve another important dopaminergic target, the anterior cingulate cortex (ACC; Boorman, Rushworth, & Behrens, 2013; Curtis & Lee, 2010; Gan, Walton, & Phillips, 2009; Hayden et al., 2011; Kolling et al., 2012; Kurzban et al., 2012; Seo, Barraclough, & Lee, 2007; Shenhav, Botvinick, & Cohen, 2013; Walton et al., 2009). Neurons in the ACC track reward history with a range of time constants (Bernacchia, Seo, Lee, & Wang, 2011; Seo et al., 2007), and the slower of these may also serve as a signal for the average reward rate. Finally, the cingulate's projections to another neuromodulatory nucleus, the locus coeruleus–norepinephrine system, have been implicated in average-reward effects that govern explore–exploit trade-offs (Aston-Jones & Cohen, 2005).

A different choice mechanism that has been argued to be at least partly distinct from the dopaminergic/TD system is "model-based" learning. Such a system learns a cognitive map or model of the structure of the task and then evaluates

[2] Note that this knowledge does not enter the estimation of the average-reward threshold, but only the decision rule or subjective value difference, where it multiplies the observed reward.

options by explicit simulation or dynamic programming (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Daw, Niv, & Dayan, 2005; Doya, 1999). Although model-based learning in the present task—that is, incrementally estimating the state transitions and rewards and solving the Bellman equation at each step for the cumulative expected values of harvesting and exiting—would likely result in behavior quite similar to the MVT learning rule, this seems an implausible account of choice behavior in the present study. Humans have been shown to use model-based evaluation in problems with very small state spaces (most often for sequences of only two actions; Daw et al., 2011), whereas quantitatively computing the optimal exit threshold by dynamic programming would require a very deep and precise search. TD and the MVT rule represent two different and more plausible "model-free" ways to short-circuit this computation, by learning decision variables directly while foregoing the global model.

Indeed, it has recently been proposed that, perhaps in addition to the model-free/model-based distinction, the brain contains psychologically and anatomically distinct mechanisms for solving (bandit-type) simultaneous choice tasks versus (patch-foraging type) serial stopping-or-switching tasks (Kolling et al., 2012; Rushworth, Kolling, Sallet, & Mars, 2012), as with our TD versus MVT models. On the basis of the imaging and physiological evidence discussed above, the ACC seems a likely substrate for foraging, whereas a long line of research has implicated nearby ventromedial prefrontal cortex in more-symmetric economic choices, such as bandit tasks (e.g., Behrens et al., 2007; Hare et al., 2011; O'Doherthy, 2011). The two sorts of choices have been most directly contrasted in a recent fMRI study (Kolling et al., 2012), supporting this anatomical dissociation, although another study using the task has argued for a different intepretation of the ACC response (Shenhav, Straccia, Cohen, & Botvinick, 2014). The present study, in the context of previous reinforcement-learning work, suggests a computational counterpart to such a distinction. In particular, learning in bandit tasks has been well described by TD and related delta-rule learning models (Barraclough et al., 2004; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Ito & Doya, 2009; Li & Daw, 2011; Sugrue et al., 2004). Our finding that learning in a serial switching problem is instead more consistent with a structurally different threshold-learning rule is consistent with the idea that these two sorts of choices employ distinct computational mechanisms and helps to flesh out the function of this proposed additional mode of choice.

This computational distinction sets the stage for future work directly comparing both approaches against a range of tasks, so as to understand under what circumstances the brain favors each approach. It has been suggested that organisms may approach many problems like foraging tasks, including even classic *symmetric* or *simultaneous* ones like economic lotteries and bandit tasks (Brandstätter, Gigerenzer, & Hertwig, 2006; Hills & Hertwig, 2010; Hills, Jones, & Todd, 2012; Kacelnik, Vasconcelos, Monteiro, & Aw, 2010). One might hypothesize a rational trade-off: that the brain favors the simpler MVT rule in tasks for which it is optimal, or nearly so, but favors more complex decision rules in tasks in which these would earn more effectively. The MVT threshold-learning rule differs from TD in two respects, by essentially neglecting two sorts of information that are inessential to patch-foraging and analogously structured tasks. The first difference is that foraging models neglect the value of any other alternatives in a simultaneous choice set, since they assume that options are evaluated sequentially. Subjects must accept or reject each option by comparing it to some global aspiration level (the average reward), which can account for alternatives only in the aggregate via their long-run statistics. This approach could be applied to arbitrary problems, such as bandit tasks, by assuming that even when options are proffered simultaneously, subjects evaluate them serially, taking some option as a "default." Regardless of how the default is set, this should produce a sort of satisficing relative to a maximizing model (like TD) that evaluates each option and chooses among them, since subjects would select a "good enough" default even when a better option was available. This inefficiency should be especially pronounced—and, perhaps, TD methods most favored—when the choice set is changed from trial to trial (e.g., Behrens et al., 2007).

The second difference has to do with sequential credit assignment. An MVT rule, unlike TD, neglects the longer-run consequences of engaging with an option by myopically considering only its instantaneous or next-step reward rate. For patch-leaving, such myopic evaluation results in the optimal long-run choice rule as long as patches degrade monotonically. (The same rule is optimal in models of serial prey selection—e.g., Krebs, Erichsen, Webber, & Charnov, 1977—because encounters are independent and the gain for processing each prey is a single event.) However, in a modified foraging-like task in which options can improve—for instance, deciding whether to sell a stock whose dividend might decrease and increase cyclically, or to fire an employee whose productivity could be improved with training—then the MVT would undervalue staying. The myopia of foraging rules has also been noted in time-discounting problems (e.g., Stephens, Kerr, & Fernandez-Juricic, 2004). Finally, long-run credit assignment is important to many differently structured multistep choice tasks, such as spatial navigation or multiplayer games. For instance, in a two-choice Markov decision task, which we have elsewhere argued is learned by a combination of TD and model-based reinforcement learning (Daw et al., 2011), the MVT rule would be at chance in selecting a first-stage move, since it earns no immediate reward and is only distinguished by the reward earned following an additional state transition and choice. Different tasks will vary in the degrees

to which myopic credit assignment results in inefficient returns.

Overall, although the class of switching tasks for which the MVT solution is optimal is small, the idea of staying or switching by comparing short-run returns to the expected long-run average is a plausible, albeit suboptimal, heuristic across many domains. MVT may be an overall favorable algorithm across a broader range of tasks because the resources saved in simplified computation may outweigh the reward losses. Thus, the choice and learning mechanisms isolated in simple patch-foraging problems may shed light on broadly applicable systems, different from but complementary to those that have seen more attention so far.

Author note This research was funded by Human Frontiers Science Program Grant No. RGP0036/2009-C and by Grant No. R01MH087882 from the National Institute of Mental Health. N.D.D. is supported by a Scholar Award from the McDonnell Foundation. We thank Paul W. Glimcher for helpful discussions, and Dylan S. Simon for technical assistance.

References

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus–norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, 28,* 403–450. doi:10.1146/annurev.neuro.28.061604.135709

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience, 7,* 404–410.

Baum, W. M. (1974). Choice in free-ranging wild pigeons. *Science, 185,* 78–79.

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

Beierholm, U., Guitart-Masip, M., Economides, M., Chowdhury, R., Düzel, E., Dolan, R., & Dayan, P. (2013). Dopamine modulates reward-related vigor. *Neuropsychopharmacology, 38,* 1495–1503.

Bernacchia, A., Seo, H., Lee, D., & Wang, X.-J. (2011). A reservoir of time constants for memory traces in cortical neurons. *Nature Neuroscience, 14,* 366–372.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica, 22,* 23–36.

Bernstein, C., Kacelnik, A., & Krebs, J. (1988). Individual decisions and the distribution of predators in a patchy environment. *Journal of Animal Ecology, 57,* 1007–1026.

Boorman, E. D., Rushworth, M. F., & Behrens, T. E. (2013). Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. *Journal of Neuroscience, 33,* 2242–2253.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review, 113,* 409–432. doi:10.1037/0033-295X.113.2.409

Cain, M. S., Vul, E., Clark, K., & Mitroff, S. R. (2012). A Bayesian optimal foraging model of human visual search. *Psychological Science, 23,* 1047–1054.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology, 9,* 129–136.

Cools, R. (2008). Role of dopamine in the motivational and cognitive control of behavior. *The Neuroscientist, 14,* 381–395.

Cools, R., Nakamura, K., & Daw, N. D. (2011). Serotonin and dopamine: Unifying affective, activational, and decision functions. *Neuropsychopharmacology, 36,* 98–113.

Curtis, C. E., & Lee, D. (2010). Beyond working memory: The role of persistent activity in decision making. *Trends in Cognitive Sciences, 14,* 216–222.

Cuthill, I. C., Kacelnik, A., Krebs, J. R., Haccou, P., & Iwasa, Y. (1990). Starlings exploiting patches: The effect of recent experience on foraging decisions. *Animal Behaviour, 40,* 625–640.

Daw, N. D., & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation, 14,* 2567–2583.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8,* 1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876–879.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69,* 1204–1215. doi:10.1016/j.neuron.2011.02.027

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks, 12,* 961–974.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science, 306,* 1940–1943.

Freidin, E., & Kacelnik, A. (2011). Rational choice, context dependence, and the value of information in european starlings (*Sturnus vulgaris*). *Science, 334,* 1000–1002.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review, 107,* 289–344. doi:10.1037/0033-295X.107.2.289

Gan, J. O., Walton, M. E., & Phillips, P. E. M. (2009). Dissociable cost and benefit encoding of future rewards by mesolimbic dopamine. *Nature Neuroscience, 13,* 25–27.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience, 10,* 252–269. doi:10.3758/CABN.10.2.252

Guitart-Masip, M., Beierholm, U. R., Dolan, R., Duzel, E., & Dayan, P. (2011). Vigor in the face of fluctuating rates of reward: An experimental examination. *Journal of Cognitive Neuroscience, 23,* 3933–3938.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *Journal of Neuroscience, 26,* 8360–8367.

Hare, T. A., Schultz, W., Camerer, C. F., O'Doherty, J. P., & Rangel, A. (2011). Transformation of stimulus value signals into motor commands during simple choice. *Proceedings of the National Academy of Sciences, 108,* 18120–18125.

Hayden, B. Y., Pearson, J. M., & Platt, M. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience, 14,* 933–939.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior, 4,* 267–272. doi:10.1901/jeab.1961.4-267

Herrnstein, R. J. (1991). Experiments on stable suboptimality in individual behavior. *American Economic Review, 81,* 360–364.

Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological Science, 21,* 1787–1792.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review, 119,* 431–440. doi:10.1037/a0027373

Hodges, C. M. (1985). Bumble bee foraging: Energetic consequences of using a threshold departure rule. *Ecology, 66,* 188–197.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). Cambridge, MA: MIT Press.

Hutchinson, J. M. C., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: Can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour, 75,* 1131–1349.

Ito, M., & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience, 29,* 9861–9874. doi:10.1523/JNEUROSCI. 6157-08.2009

Jacobs, E. A., & Hackenberg, T. D. (1996). Humans' choices in situations of time-based diminishing returns: Effects of fixed-interval duration and progressive-interval step size. *Journal of the Experimental Analysis of Behavior, 65,* 5–19.

Kacelnik, A. (1984). Central place foraging in starlings (*Sturnus vulgaris*): I. Patch residence time. *Journal of Animal Ecology, 53,* 283–299.

Kacelnik, A. (1997). Normative and descriptive models of decision making: Time discounting and risk sensitivity. *Ciba Foundation Symposium, 208,* 51–70.

Kacelnik, A., Vasconcelos, M., Monteiro, T., & Aw, J. (2010). Darwin's "tug-of-war" vs. starlings' "horse-racing": How adaptations for sequential encounters drive simultaneous choice. *Behavioral Ecology & Sociobiology, 65,* 547–558.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology, 7,* e1002055. doi:10.1371/journal.pcbi.1002055

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. (2012). Neural mechanisms of foraging. *Science, 336,* 95–98.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature Neuroscience, 13,* 1292–1298.

Krebs, J. R., & Inman, A. J. (1992). The University of Chicago learning and foraging: Individuals, groups, and populations. *American Naturalist, 140,* S63–S84.

Krebs, J. R., Erichsen, J. T., Webber, M. I., & Charnov, E. L. (1977). Optimal prey selection in the great tit (*Parus major*). *Animal Behaviour, 25,* 30–38.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2012). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences, 36,* 697–698.

Lex, A., & Hauber, W. (2008). Dopamine D1 and D2 receptors in the nucleus accumbens core and shell mediate Pavlovian-instrumental transfer. *Learning and Memory, 15,* 483–491.

Li, J., & Daw, N. D. (2011). Signals in human striatum are appropriate for policy update rather than value prediction. *Journal of Neuroscience, 31,* 5504–5511.

McCall, J. J. (1970). Economics of information and job search. *Quarterly Journal of Economics, 84,* 113–126.

McNamara, J. M., & Houston, A. I. (1985). Optimal foraging and learning. *Journal of Theoretical Biology, 117,* 231–249.

McNickle, G. G., & Cahill, J. F. (2009). Plant root growth and the marginal value theorem. *Proceedings of the National Academy of Sciences, 106,* 4747–4751. doi:10.1073/pnas.0807971106

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience, 16,* 1936–1947.

Niv, Y., Daw, N., & Dayan, P. (2006). How fast to work: Response vigor, motivation and tonic dopamine. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in neural information processing systems* (Vol. 18, pp. 1019–1026). Cambridge, MA: MIT Press.

Niv, Y., Daw, N. D., Joel, D., & Dayan, P. (2007). Tonic dopamine: Opportunity costs and the control of response vigor. *Psychopharmacology, 191,* 507–520. doi:10.1007/s00213-006-0502-4

Niyogi, R. K., Breton, Y.-A., Solomon, R. B., Conover, K., Shizgal, P., & Dayan, P. (2014). Optimal indolence: A normative microscopic approach to work and leisure. *Interface, 11,* 91.

O'Doherthy, J. P. (2011). Contributions of the ventromedial prefrontal cortex to goal-directed action selection. *Annals of the New York Academy of Sciences, 1239,* 118–129.

Ollason, J. G. (1980). Learning to forage-optimally? *Theoretical Population Biology, 56,* 44–56.

Puterman, M. L. (2009). *Markov decision processes: Discrete stochastic dynamic programming*. New York, NY: Wiley.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience, 9,* 545–556.

Robbins, T. W., & Everitt, B. J. (2007). A role for mesencephalic dopamine in activation: Commentary on Berridge (2006). *Psychopharmacology, 191,* 433–437.

Rushworth, M. F. S., Kolling, N., Sallet, J., & Mars, R. B. (2012). Valuation and decision-making in frontal cortex: One or many serial or parallel systems? *Current Opinion in Neurobiology, 22,* 946–955. doi:10.1016/j.conb.2012.04.011

Rustichini, A. (2009). Neuroeconomics: Formal models of decision making and cognitive neuroscience. In P. W. Glimcher, C. F. Camerer, E. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision making and the brain* (pp. 33–46). London, UK: Elsevier Academic Press.

Salamone, J. D. (1988). Dopaminergic involvement in activational aspects of motivation: Effects of haloperidol on schedule-induced activity, feeding, and foraging in rats. *Psychobiology, 16,* 196–206.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599.

Schwartz, A. (1993). A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the Tenth International Conference on Machine Learning (ICML '93)* (pp. 298–305). Piscataway, NJ: IEEE Press.

Seo, H., Barraclough, D. J., & Lee, D. (2007). Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. *Cerebral Cortex, 17,* 110–117.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron, 79,* 217–240.

Shenhav, A., Straccia, M. A., Cohen, J. D., & Botvinick, M. M. (2014). Anterior cingulate engagement in a foraging context reflects choice difficulty, not foraging value. *Nature Neuroscience, 17,* 1249–1254. doi:10.1038/nn.3771

Smith, E. A., & Winterhalder, B. (1992). *Evolutionary ecology and human behavior*. New York, NY: Aldine De Gruyter.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage, 46,* 1004–1017.

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

Stephens, D. W., Kerr, B., & Fernandez-Juricic, E. (2004). Impulsiveness without discounting: The ecological rationality hypothesis. *Proceedings of the Royal Society B, 271,* 2459–2465.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science, 304,* 1782–1787.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning, 3,* 9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science, 315,* 515–518.

Walton, M. E., Groves, J., Jennings, K. A., Croxson, P. L., Sharp, T., Rushworth, M. F. S., & Bannerman, D. M. (2009). Comparing the role of the anterior cingulate cortex and 6-hydroxydopamine nucleus accumbens lesions on operant effort-based decision making. *European Journal of Neuroscience, 29,* 1678–1691. doi:10.1111/j.1460-9568.2009.06726.x

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. PhD thesis, Cambridge University, Cambridge, UK.

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