Decision

2325-9965/17/$12.00 http://dx.doi.org/10.1037/dec0000074

Exploratory Choice Reflects the Future Value of Information

Alexander S. Rich and Todd M. Gureckis
New York University

The tension between exploration and exploitation is a primary challenge in decision making under uncertainty. Optimal models of choice prescribe that individuals resolve this tension by evaluating how information gained from their choices will improve future choices. However, research in behavioral economics and psychology has yielded conflicting evidence about whether people consider the future during exploratory choice, particularly in complex, uncertain environments. Adding to the empirical evidence on this issue, we examine exploratory decision making in a novel approach-avoid paradigm. In the first set of experiments we find that people parametrically increase their exploration when the expected number of future encounters with a prospect is larger. In the second we demonstrate that when the number of future encounters is unknown, as is often the case in everyday life, people are sensitive to the relative frequency of future encounters with a prospect. Our experiments show that people adaptively utilize information about the future when deciding to explore, a tendency that may shape decisions across several real-world domains.

Keywords: decision making, exploration, information sampling, learning

Supplemental materials: http://dx.doi.org/10.1037/dec0000074.supp

People are often faced with decisions among uncertain alternatives (Mehlhorn et al., 2015; Sutton & Barto, 1998). To behave effectively, they must balance exploration of relatively unknown options with exploitation of those currently believed to be most rewarding.

An optimal decision maker achieves this balance by considering the future when deciding whether to explore. Rather than evaluate only the immediate reward from choosing an option, he or she also considers the degree to which learning about the option is expected to improve future choices, known as the value of information (Bellman, 1957; Howard, 1966). This means that expectations about future encounters with an option—how soon they will occur, how often they will occur, or if they will occur at all—are relevant to the immediate decision to explore something new. For example, the value of learning about a new local café is lower for someone who will move away in two weeks than for someone who will move in two years, and is higher for someone who drinks coffee frequently than for someone who drinks it rarely.

Work in economics and marketing has argued that people make exploratory choices in a manner well-described by forward-looking models that match or approach optimal behavior (Aguirregabiria & Mira, 2010; Ching, Erdem, & Keane, 2013; Erdem & Keane, 1996; Kreps & Porteus, 1978). Although this class of models has been used to explain consumer choice (Erdem & Keane, 1996) and behavior ranging from medical decision making to college enrollment (Chintagunta, Goettler, & Kim, 2012; Stange, 2012), it is difficult to determine from field data whether people are truly forward-looking in their exploration (Ching et al., 2013). Experimental methods from behavioral economics may more clearly identify the signature of future-sensitive exploratory behavior, but such studies have produced mixed evidence,

Alexander S. Rich and Todd M. Gureckis, Department of Psychology, New York University.

This work was supported by grant BCS-1255538 from the National Science Foundation and a John S. McDonnell Foundation Scholar Award to T.M.G. The authors thank Dan Bartels, Paul Glimcher, Gregory Murphy, Yael Niv, and Peter Todd for helpful discussions in the development of this work.

Correspondence concerning this article should be addressed to Alexander S. Rich, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003. E-mail: asr443@nyu.edu

particularly in cases where the number of future choices is not precisely known (Banks, Olson, & Porter, 1997; Lee, Zhang, Munro, & Steyvers, 2011; Meyer & Shi, 1995; Wilson, Geana, White, Ludvig, & Cohen, 2014).

In the current study, we report on a series of large, online experiments using a set of novel approach-avoid decision-making tasks. We find that people are sensitive to multiple forms of expectations about the future and use these expectations to guide exploratory choice. In Experiment Sequence 1 we clarify the way in which the future number of encounters with a prospect affects exploration, a relationship that has been identified in past work (Lee et al., 2011; Meyer & Shi, 1995; Wilson et al., 2014). In Experiment Sequence 2 we extend our paradigm to include uncertainty about the number of future encounters, a situation common to daily life but in which past studies have failed to uncover future-sensitive exploration (Banks et al., 1997). We show that when expectations about the future are expressed as the frequency of future encounters with a prospect, people effectively make use of this relative frequency information to guide their exploration.

Approach–Avoid Decision Making

We focus on a class of decision-making dilemmas that we term "approach-avoid" decision making, in which a person must choose whether to approach and sample an uncertain prospect, or avoid the prospect in favor of a well-known default. Many common decisions reduce to approach-avoid dilemmas. For example, a person may choose to either try a new café or maintain a default routine, which (depending on the individual) may mean going without coffee, making coffee at home, or going to a well-known café. Similarly, a consumer may choose between familiar and unfamiliar brands, and a doctor may select between standard and novel treatments. Critically, we focus on situations where this decision is not "one-off," but where instead the decision-maker may expect to make the same or similar decisions in the future.

This kind of scenario, in which an agent makes a series of choices between one alternative with an uncertain reward distribution and one with a known reward distribution, is known in the statistics literature as a one-armed bandit problem (Berry & Fristedt, 1979). In the formal definition of the problem, an agent has a distribution of beliefs F over the reward distribution of the uncertain alternative (e.g., approaching and trying the new café), and we assume without loss of generality that the known alternative (e.g., avoiding the café and skipping coffee today) yields a mean reward of zero. The agent's goal is to maximize summed expected reward over a sequence of choices weighed by a *discount sequence* $A = (\alpha_1, \alpha_2, \ldots)$, a sequence of non-negative numbers that determine the importance of rewards received from each choice.

The discount sequence represents the agent's expectations about and valuation of future choices. Two types of discount sequences are of particular interest. In the first, $\alpha_m = 1$ for $m \leq n$ and $\alpha_m = 0$ for $m > n$. This is known as a finite horizon (Sutton & Barto, 1998), and corresponds to cases where a person makes a known number of choices, n, and cares equally about the outcomes from each. In the second, $\alpha_m = d^{m-1}$ and d is a *discount rate* that is non-negative and less than 1. This is known as an infinite horizon (Sutton & Barto, 1998) and corresponds to cases where a person is unsure of the number of future choices. The progressively decreasing weight of future choices reflects uncertainty over whether a given future choice and its resultant rewards will occur. (If the decision maker intrinsically prefers earlier rewards to later rewards (Frederick, Loewenstein, & O'Donoghue, 2002), this time preference can also be incorporated into the weight of future choices.)

Within both finite and infinite horizons, we can compare the *length* of two horizons. One finite horizon is longer than another when the number of future choices is higher, which occurs when n is larger. One infinite horizon is longer than another when the *expected* number of future choices is higher, which occurs when d is larger. In both cases, for a decision maker who is sensitive to future choices, the weight of future rewards relative to immediate reward will increase as the horizon lengthens.

Effect of Horizon on Optimal Exploration

A general relationship holds between expectations about the future and optimal choice: as the horizon grows longer, the value of approaching relative to avoiding increases (or remains the same in the limiting case; see Proof of

the nondecreasing relative value of approaching in the supplemental material available online.). More precisely, let $V_{ap}(F, A)$ and $V_{av}(F, A)$ be the expected value of first choosing the uncertain (approach) or certain (avoid) alternative, respectively, and subsequently following an optimal strategy, and let discount sequence A^+ be longer than A. Then for any belief distribution F,

$$V_{ap}(F, A^+) - V_{av}(F, A^+) \geq V_{ap}(F, A) - V_{av}(F, A)$$

The horizon has no effect on the immediate expected reward of either action, so this change in relative values reflects solely the increasing value of approaching as an exploratory, information-seeking action. When there is only a single choice remaining or the discount rate is zero, information about the uncertain alternative cannot be used for consequential future choices and thus has no value. As the horizon lengthens, collecting information to improve future choices becomes increasingly valuable. Because gaining information is *contingent* on approaching, the relative value of approaching increases.

To illustrate this relationship and its consequences, consider the example of a finite-horizon problem where the known alternative has a constant payoff of 0, and the uncertain alternative produces payoffs of 1 and −1. Suppose the uncertain alternative is expected to produce the higher payoff on either 1/3 or 2/3 of trials, with these two possibilities equally likely a priori.

Figure 1 shows the behavior of an optimal agent engaging in problems of this type with horizons ranging from one to 32 choices. When there is a single choice, the agent is indifferent between approaching and avoiding because the expected value of both options is zero. When information can be used to inform at least one future choice, the agent initially approaches, and persists longer in approaching when the horizon is longer. This persistence represents a trade-off that harms the agent in some cases and helps it in others, because it occurs both when the prospect's true expected value is positive and when it is negative. But while approaching a mostly negative prospect yields information that corrects the agent's beliefs and causes only

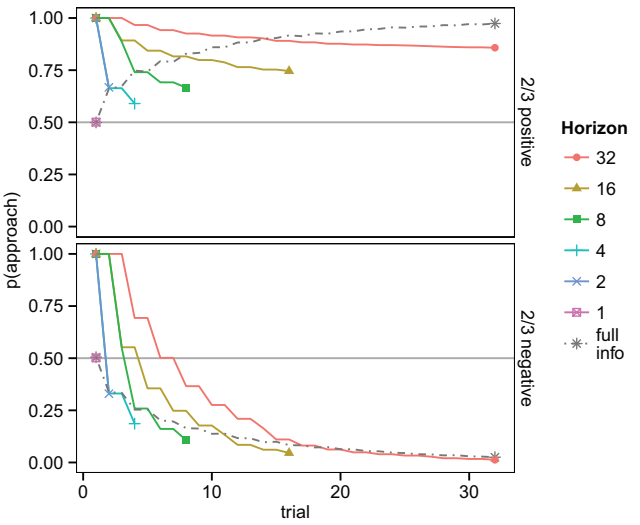

Figure 1. Simulated behavior of an optimal agent when encountering a prospect over a finite horizon. The top panel shows behavior when the prospect is truly 2/3 positive, and the bottom panel shows behavior when it is truly 2/3 negative. With choice-contingent information, the agent approaches on early trials to collect information, and approaches more persistently in longer horizons. With full information, the agent's behavior simply tracks the prospect's immediate expected reward given the observed outcomes, regardless of horizon. The model was simulated over 10,000 task iterations. See the online article for the color version of this figure.

short term costs, avoiding a mostly positive prospect yields no belief-correcting information and inflicts long-term costs (Denrell & March, 2001). This means that as the horizon lengthens, the relative cost of avoiding a good prospect grows relative to that of approaching a bad one.

The role of the value of information in determining optimal behavior is made particularly clear by comparing the standard one-armed bandit problem, with choice-contingent information, to a problem with *full* information where feedback about the foregone payoff is provided upon avoiding. Optimal behavior in this situation is plotted as a dotted gray line in Figure 1. The expected immediate reward of approaching is the same as in the original formulation, but approaching no longer has additional value because information is provided regardless of choice. As a result, horizon loses its influence on choice and the optimal policy becomes a myopic reward-maximizing policy that tracks the expected immediate return from each arm.

Past Experimental Tests of Forward-Looking Exploration

As the rational analysis in the previous section makes clear, the expectation of future encounters should cause a bias toward approaching an uncertain prospect, and this bias should increase as the horizon lengthens. Past experiments have yielded mixed results as to how much human exploratory choice reflects these two patterns.

First, some work has suggested that there is no bias toward choosing uncertain options, and that people are uncertainty-insensitive (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) or even uncertainty-averse (Payzan-LeNestour & Bossaerts, 2011). However, studies using more sophisticated models (Speekenbrink & Konstantinidis, 2014) or more constrained tasks (Knox, Otto, Stone, & Love, 2012) report that people do in fact tend to explore uncertain options. In addition, experiments in which participants can make a series of purely information-seeking actions before making a single consequential choice show that people are willing to sacrifice both time and money to reduce uncertainty (Hertwig, Barron, Weber, & Erev, 2004; Juni, Gureckis, & Maloney, 2016).

These studies establish that uncertainty can drive exploratory choice. A smaller number of studies have more directly examined whether exploration is forward-looking by testing the effects of horizon length, particularly with finite horizons. Although these studies have generally found that exploration is uncertainty-sensitive, they vary in their reported effect of horizon. Meyer and Shi (1995) found that people chose the uncertain option more in a one-armed bandit task of 20 trials than in one of 5 trials, and Lee et al. (2011) found evidence of horizon-sensitivity among some participants in 8-trial and 16-trial four-armed bandit tasks. Most recently, Wilson et al. (2014) found that people explored more when six trials remained than on the final trial of a bandit task, but that exploration did not increase further when 11 trials remained. They proposed that people may have a heuristic to explore more when future encounters are expected than when they are not.

To our knowledge, Banks et al. (1997) conducted the only experiment examining the effect of infinite horizons with differing discount rates on exploration. To manipulate discount rate, participants in two conditions of a one-armed bandit task were informed that the probability of the task continuing after each trial was .8 or .9, respectively. Although participants tended to choose the uncertain arm, there was no difference in this tendency between the two conditions. It might be that the small difference in stopping probabilities along with a low-power design contributed to this reported null effect.

In summary, there is evidence that people show some form of sensitivity to the future value of information when the horizon is finite. However, there is no existing evidence that people are sensitive to the value of information in an infinite-horizon setting, despite finite horizons being relatively uncommon in everyday life.

In the following experiments, we first test whether approach behavior parametrically increases with the length of a finite horizon, as should be the case for a forward-looking decision-maker. These studies are designed to confirm the existing literature and test the generality of those findings. We then provide a novel test of forward-looking exploration in an infinite horizon. This represents an important contribution to the literature, especially given the relevance of infinite-horizon tasks to everyday choice behavior. In both cases we also compare

behavior in the standard, contingent-information task with behavior in a noncontingent, full-information task where foregone payoffs are revealed. These control conditions verify that horizon-dependent behavior is in fact attributable to information seeking rather than being an idiosyncratic aspect of our experimental task. Although several studies have described the effects of receiving foregone payoffs (Grosskopf, Erev, & Yechiam, 2006; Yechiam & Busemeyer, 2006), little work has documented the effect of *anticipating* foregone payoff information on exploratory choice.

Experiment Sequence 1—Finite Horizons

We conducted two experiments to investigate whether the length of a finite horizon affects approach-avoid decision-making, and whether this effect is linked to information-seeking. In Experiment 1a, we tested whether exploration increased parametrically with horizon when information is choice-contingent. In Experiment 1b, we replicated the results of Experiment 1a in a setting where participants were given more precise prior information about the environment, and also added a control condition in which participants were given full information regardless of choice, and in which no effect of horizon was predicted.

Participants completed a sequence of one-armed bandit problems in the form of a mushroom-foraging game. They visited patches that each contained a unique mushroom species and had different numbers of mushrooms available, creating different horizons. They encountered each mushroom in the patch in turn, and chose to either eat it to receive an uncertain payoff, or avoid it to receive a payoff of zero. This simple scenario mimics the formal analysis described above. Compared to past experiments, we used a relatively simple task, wider range of horizons, and greater number of participants.

Method

Participants. Participants were recruited via Amazon Mechanical Turk using the psiTurk framework (Gureckis et al., 2016) and compensated with a monetary payment and performance-based bonus. Past work has shown that data collected using AMT is comparable with data collected in a lab setting (Crump, McDonnell, & Gureckis, 2013). Participants were tested on their comprehension of the experiment instructions, and data from participants who failed the comprehension test more than twice were excluded. Based on model simulations we predicted that a clear effect of horizon would emerge with 100 participants. For both experiments we conducted a preliminary qualitative analysis of the first 50 participants before completing participant recruitment. Final sample sizes were 143 (3 excluded) in Experiment 1a and 254 (22 excluded) across the two conditions of Experiment 1b. No variables or conditions were dropped from our analysis.

Design and procedure. Participants in both experiments played a game based around foraging for edible mushrooms. Mushrooms were represented by color illustrations. Figure 2 shows examples of the task in each experiment.

Experiment 1a. Participants sequentially encountered patches of mushrooms that contained 1, 2, 4, 8, 16, or 32 exemplars of a single species, and were informed that each species was unique to a single patch. The goal of the task was to eat healthy mushrooms while avoiding poisonous ones. Participants were told that species varied in their proportion of healthy mushrooms, from "almost-always healthy" to "almost-always poisonous." They encountered four patches of each length; one each of proportions $p(healthy) = \{.125, .375, .625, .875\}$. The patches were pseudorandomly ordered.

In each patch participants first observed the set of available mushrooms represented as a group of empty circles. On each trial, participants chose whether to eat or avoid the next mushroom in the patch by clicking buttons labeled "eat" and "avoid." Upon eating a mushroom, it turned green (if healthy) or red (if poisonous) and moved to a group of healthy or poisonous mushrooms. Upon avoiding a mushroom, it turned gray and moved to group of avoided mushrooms. The number of remaining mushrooms in the patch was denoted at the top of the screen at all times.

Participants started with a bonus of \$.25. They earned \$.02 for each healthy mushroom eaten and lost \$.02 for each poisonous mushroom eaten. They did not gain or lose money for avoiding a mushroom. The bonus was cumulative over all patches, and its value was visible throughout the game.

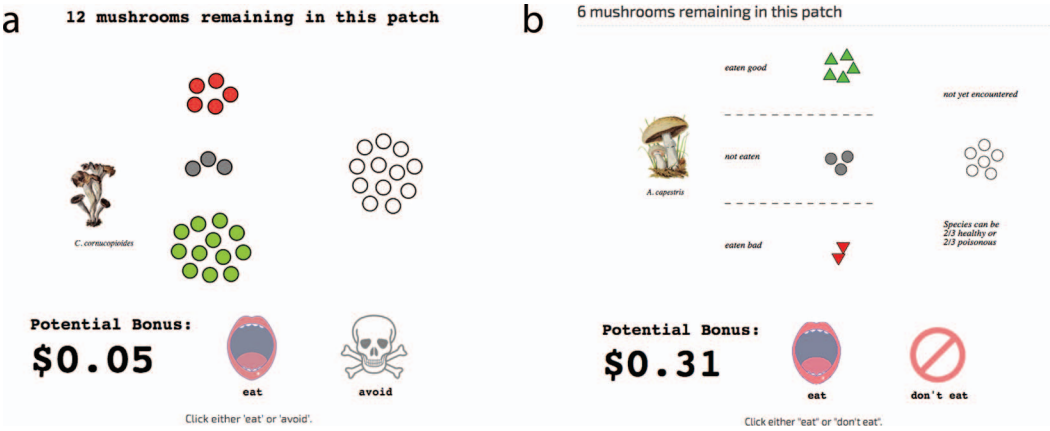

Figure 2. Example of the Experiment 1a task (a), and of the Experiment 1b task (b) in the contingent-information condition. See the online article for the color version of this figure.

Experiment 1b. As in Experiment 1a, participants encountered patches of mushrooms that contained 1, 2, 4, 8, 16, or 32 exemplars. Participants were given more precise information that each species was either of the "mostly-healthy" type, for which 2/3 of individuals were healthy, or the "mostly-poisonous" type, for which 2/3 of individuals were poisonous, and that these two types were equally common. These environmental statistics match those for which the optimal policy is shown in Figure 1. Participants encountered four patches of each length, with two of each type.

Participants were split into two conditions. In the contingent-information condition, the button to avoid (labeled "don't eat") was described simply as not eating the mushroom. In the full-information condition, participants were told that when they chose not to eat the mushroom, they would put it into a "mushroom-testing kit" and learn its value.

Eaten mushrooms were represented as upward-pointing green triangles if healthy and downward-pointing red triangles if poisonous. Not-eaten mushroom were represented as gray circles in the contingent-information condition. In the full-information condition, they were represented as grayish-green upward-pointing or grayish-red downward-pointing triangles, depending on their healthiness.

As in Experiment 1a, participants earned a cumulative bonus that started at $.25 and changed in increments of $.02 as they ate healthy or poisonous mushrooms.

Results

Participants' probability of approaching a mushroom on each trial within patches of each length is shown in Figure 3. In both Experiment 1a and the contingent-information condition of Experiment 1b, participants' behavior resembles that of the forward-looking, information-sensitive model (see Figure 1), with a high rate of exploration in early trials and more persistent exploration in larger patches. In the full-information condition of Experiment 1b, participants appeared to begin approaching at a rate near chance and then to modify their behavior based on observed outcomes, similar to the myopic reward-maximizing policy shown in Figure 1.

To quantify the effects of horizon and other variables on trial-by-trial behavior we used a hierarchical Bayesian logistic regression that allowed for individual differences among participants (Gelman et al., 2013; see Description of data analysis in the supplemental material available online). We included five predictors: a bias term (capturing overall tendencies to approach or avoid), immediate expected reward (i.e., expected payoff from approaching on the next trial), number of remaining trials in the patch (horizon length), trial number within the patch (which may have an independent effect if participants were uncertainty- or novelty-seeking), and the interaction between trial number and horizon.

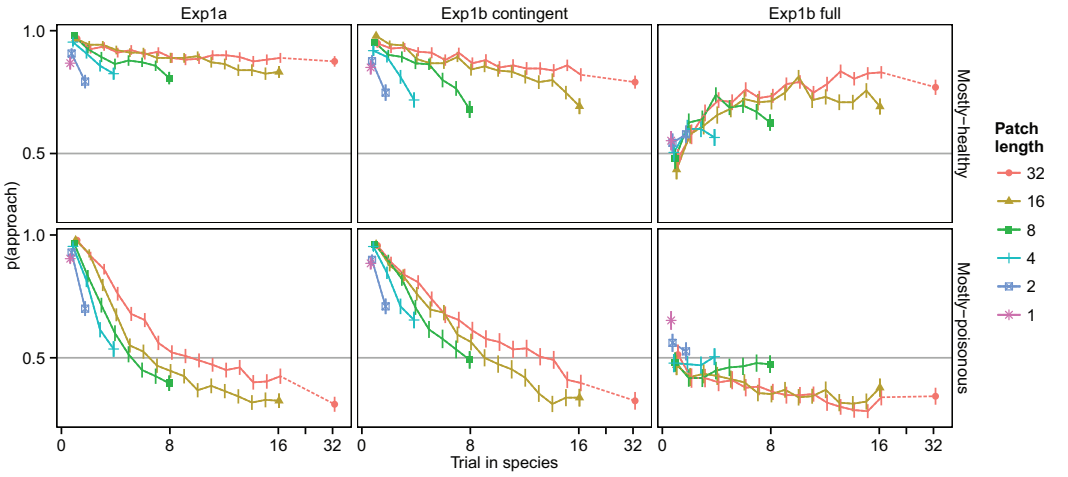

Figure 3. Probability of approaching a mushroom on each encounter within a patch for each finite-horizon experiment and condition. Top panels show participant behavior for patches where *p*(*healthy*) was .625 or .875 (Exp. 1a), or .67 (Exp. 1b), and bottom panels show behavior for patches where *p*(*healthy*) was .125 or .375 (Exp. 1a), or .33 (Exp. 1b). Error bars show standard error of the mean. In order to focus on comparisons among patch lengths, the *x* axis is compressed after Trial 16 and the change in *p*(*approach*) from Trial 16 to Trial 32 is shown as a dotted line. Participants tended to approach early on and to have higher *p*(*approach*) for longer horizons in Experiment 1a and in the contingent-information condition of Experiment 1b, but not in the full-information condition of Experiment 1b. See the online article for the color version of this figure.

We calculated expected reward by applying Bayes' rule using the participant's observed outcomes for a patch, with a uniform prior over *p*(*healthy*) in Experiment 1a and the prior given in the instructions in Experiment 1b. Horizon and trial-number were log transformed because the marginal effects of both factors are expected to be decreasing. We rescaled expected reward to equal −1 when *p*(*healthy*) = 1/3 and 1 when *p*(*healthy*) = 2/3, and rescaled log horizon to range from 0 (at horizon 1) to 1 (at horizon 32). Finally, we shifted log trial-number to have zero mean so that the horizon coefficient could be interpreted as an average effect across all trials.

The model posterior was estimated for each experiment and condition using the Stan modeling language (Stan Development Team, 2015). Posterior estimates of the population-level parameters are presented in Figure 4. Simulated data from the model posteriors confirmed a close match to the key features of the data (see Figure S1 in the supplemental material available online).

Participants were positively sensitive to expected reward across all three scenarios. The model posteriors confirm that behavior was similar in Experiment 1a and the contingent-information condition of Experiment 1b. In both scenarios, participants were highly exploratory early on; the 95% posterior predictive intervals for first-trial approach proportion were [.920, .937] and [.894, .916], respectively. Participants became less likely to approach over the course of a patch, as shown by a negative effect of trial number. This decrease in exploration likely results from the decreased uncertainty about and novelty of later mushrooms in a patch.

Critically, as shown in Figure 4, there was a positive population-level effect of horizon in both contingent-information scenarios, suggesting that people were not simply uncertainty-seeking but used a forward-looking strategy that tracked the value of information. This sensitivity to horizon also held broadly at the individual level. In Experiment 1a, the posterior mean effect of horizon was above zero for 81% of participants, and the 95% posterior interval for the effect of horizon was entirely above zero for 51% of participants. In the contingent-information condition of Experiment 1b, the posterior

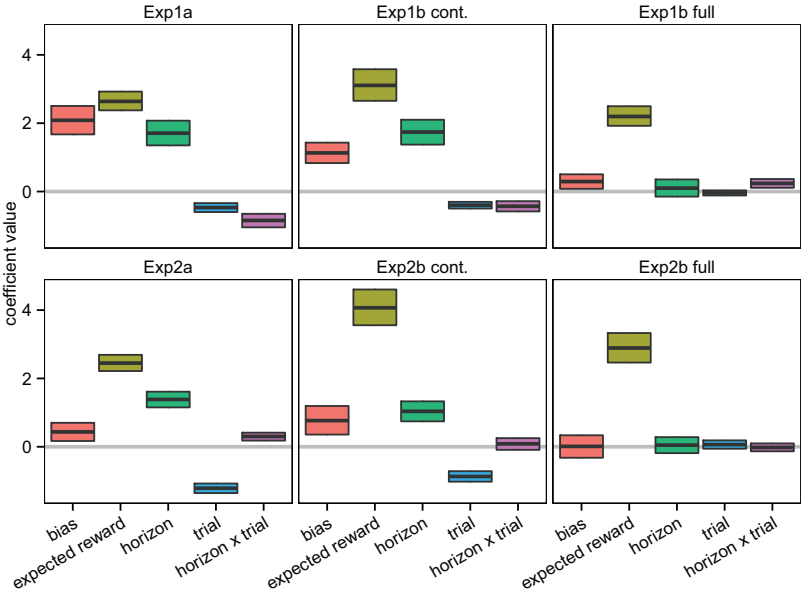

Figure 4. Posterior estimates of the population-level effects of five predictors on behavior, estimated using a hierarchical Bayesian logistic regression of participants' choices in all experiments. The top panels show results for finite-horizon experiments, and the bottom panels show results for infinite-horizon experiments. Thick lines show posterior means, and colored intervals show 95% posterior intervals. As predicted by a forward-looking model, participants exhibited a positive effect of horizon in all four contingent-information scenarios, but not in the two full-information scenarios. See the online article for the color version of this figure.

mean was above zero for 79% of participants, and the posterior interval was entirely above zero for 58% of participants. There was also a negative interaction between horizon and trial, such that the effect of horizon decreased in later trials. This may reflect that further into a patch participants were confident about the healthiness of the species, reducing the potential value of information.

Behavior in the full-information condition of Experiment 1b was markedly different from behavior in the contingent-information scenarios. Participants were roughly indifferent between approaching and avoiding on the first trial, with a 95% posterior predictive interval of [.500, .538]. There was no main effect of trial number or of horizon, and the 95% posterior interval for horizon was completely below those of the two contingent-information scenarios. We also observed that participants started approaching less in the last few trials of mostly healthy patches and more in the last few trials of mostly poisonous patches. This effect is not well-captured by our model and is contrary to our predictions, but is distinct from the information-related horizon effect in that it tends toward increased approach when the horizon is short and expected reward is negative. We hypothesize that this effect may be an instance of the gambler's fallacy, with participants expecting that a string of good outcomes would be balanced out by bad ones and vice versa (Tversky & Kahneman, 1971).

Recent work raises the question of whether the effect of horizon is parametric and smoothly increasing, or categorical with two distinct levels for situations with and without future encounters (Wilson et al., 2014). Although our main analysis assumed that the effect of horizon was logarithmic, we tested this assumption with an expanded model in which each experienced horizon, from 1 to 32, had a unique population-level parameter that was allowed to vary freely (see Description of data analysis in the supplemental material available online). We found that in both contingent-information scenarios

these values increased at a roughly logarithmic rate, supporting our supposition that the effect of horizon increases parametrically but with a decreasing marginal effect (see Figure S2 in the supplemental material available online).

Finally, although participants did modulate their exploration based on the length of the horizon, they also appeared to have a general bias toward information seeking. We tested this by comparing the final trials of the two conditions of Experiment 1b, where the true value of information was zero but participants could still choose to make an information-seeking choice in the contingent-information condition. A regression on these final trials reveals that, controlling for expected reward, participants in the contingent-information condition were more likely to approach, $z = 7.82$, $p < .001$. This bias could reflect a heuristic to collect information even when not clearly useful, for example in case the species was encountered again (although participants were informed it would not be) or the information could improve prior knowledge about future species (although the prior over $p(healthy)$ was given).

Infinite Horizons and Prospect Frequency

In Experiments 1a and 1b, as in most previous studies of task horizon and exploration (Lee et al., 2011; Meyer & Shi, 1995; Wilson et al., 2014), participants knew the exact number of times they would encounter a prospect in the future. This kind of finite horizon, though, is not representative of many of the decisions faced in everyday life. Often, people do not know how many times they will encounter a prospect. For example, there is no precise limit on the number of times someone might have the opportunity to visit a local café. This type of situation is more naturally formalized as an infinite-horizon problem.

Banks et al. (1997) attempted to induce infinite horizons with different discount rates in a one-armed bandit task by informing some participants that the task had a .9 probability of continuing after each trial, and others that the task had a .8 probability of continuing. They found that this manipulation did not affect participants' exploration, though this null result may have been a result of low power. In addition, although a probabilistic experiment length is an effective way to create uncertainty about the horizon (Camerer & Weigelt, 1996), explicit information about ending probabilities may feel artificial to participants and be difficult to integrate into decision-making strategies.

A more natural and common way in which individuals face differing effective discount rates is through the differing frequencies of prospects within a wider environment. Some types of products are purchased more often than others, and likewise some social situations are encountered more than others. Intuitively, if the length of the environment is uncertain (e.g., length of time living in a city, length of life), and encounters with one prospect are less frequent than encounters with another, then a person should expect more future encounters with the frequent prospect and value information about that prospect more highly.

We can formalize this idea by considering an agent facing a *contextual* one-armed bandit problem. The agent has a single infinite-horizon discount sequence A with discount rate d, but rather than always facing the same uncertain alternative, on each trial it encounters one of k independent uncertain alternatives (i.e., the context for that trial). Each alternative has its own starting belief distribution F_i, and each is encountered with a known frequency f_i, such that $\sum_{i=1}^{k} f_i = 1$. For example, the agent may have to decide whether to buy an espresso each morning after observing which of k baristas (with potentially varying skill) are working at the café. We assume the agent knows how frequently each barista works at the café, though this could also be learned from experience.

Since the alternatives are independent, this situation can be reduced to k independent one-armed bandit problems by considering only the sequence of encounters with the k^{th} alternative. However, the timing, and thus the discounted weight, of future encounters with each bandit will depend on its frequency, causing future encounters with rare prospects to tend to receive lower weight. Although the exact discount sequence for a given prospect is unknown in advance, these uncertain sequences can be replaced with their expected values (Berry & Fristedt, 1985), resulting in a set of k independent one-armed bandits with effective discount rates d_k.

Figure 5 shows the behavior of an optimal agent in an environment with an infinite hori-

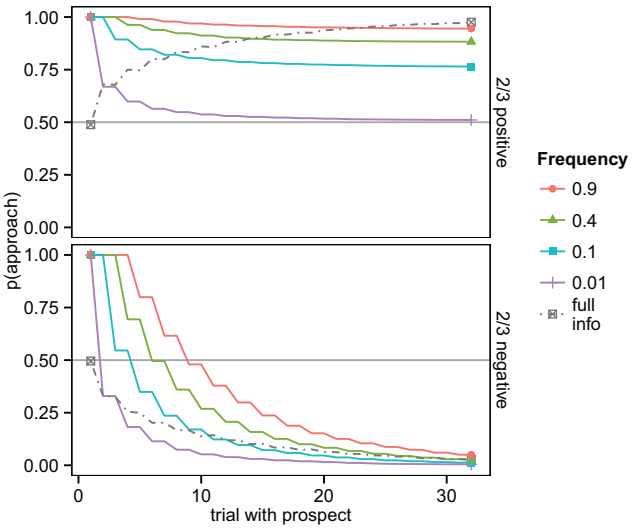

Figure 5. Simulated behavior of an optimal agent with a base discount rate of .99 when encountering prospects with differing frequencies of occurrence over an infinite horizon. Each prospect is known to either yield a payoff of +1 with probability 2/3 or a payoff of −1 with probability 2/3, and the opposite payoff otherwise. The top panel shows behavior when the prospect is truly 2/3 positive, and the bottom panel shows behavior when it is truly 2/3 negative. With choice-contingent information, the agent approaches to collect information, and approaches more persistently when the prospect is more frequent. With full information, the agent's behavior simply tracks the prospect's expected value given the observed outcomes, regardless of frequency. The model was simulated over 10,000 task iterations. See the online article for the color version of this figure.

zon, for prospects that have differing frequencies. To explore optimally, the agent approaches more when the prospect is frequent, just as it does when the finite horizon is longer (see Figure 1).

Experiment Sequence 2—Infinite Horizons

In Experiment Sequence 2, we tested whether participants display a sensitivity to prospect frequency. Although there is a large literature surrounding the effects of rare and common *outcomes* on decision strategies (e.g., Hertwig et al., 2004), to our knowledge there are few studies that examine the effect of rare and common prospects—that is, of rarely and commonly faced decision contexts. In Experiment 2a, participants performed a contingent-information task in which the horizon was unknown. In Experiment 2b, we replicated the results of Experiment 2a in a similar task with more precise information about the possible values of prospects and the distribution of possible horizons, and also added a full-information control condition.

Participants played a mushroom-foraging game that instantiated the kind of contextual one-armed bandit problem described above. Participants played through a series of *habitats* that each had four unique mushroom species, any one of which could appear on a given trial. They were trained on the relative frequency of the species at the beginning of the task, and then made a series of approach-avoid decisions. In Experiment 2a, participants were not told the length of the habitats. In Experiment 2b, we incorporated a probabilistic habitat-ending mechanism so that participants had explicit prior information about the uncertainty over the horizon.

Method

Participants. Participants were recruited via Amazon Mechanical Turk using the psiTurk framework and compensated with a monetary

payment and performance-based bonus. Data from participants who failed the preexperiment comprehension test more than twice were excluded. For both experiments we conducted a qualitative analysis of the first 50 participants before completing data collection. Sample sizes were 152 (3 excluded) in Experiment 2a, and 159 (5 excluded) across the two conditions of Experiment 2b. No variables or conditions were dropped from our analysis.

Design and procedure. Participants in both experiments played a game based around foraging for healthy mushrooms. Mushrooms were represented by color illustrations. The experiments were divided into subtasks called "habitats" (e.g., "New England Forest," "Amazonian Rain-forest"). Each habitat contained four unique mushroom species, two of which occurred with frequency 4/10 and two of which occurred with frequency 1/10. Within each habitat, the game was broken into two phases. In the first phase, participants observed a large, representative sample of the mushrooms in the habitat. Mushrooms encountered in this sample were depicted as circles that appeared without participant input. Once the entire sample had been shown, the species were highlighted one at a time and participants submitted a "field report" by answering questions of the form "If you saw 10 mushrooms on your hike back through the [Habitat Name], how many would you expect to be from the species [Species Name]?" This ensured that participants noticed and encoded the relative frequency of each species. In the second phase, participants played a game similar to those in Experiment Sequence 1, but where the species encountered on any trial was randomly interleaved with other species based on the underlying frequencies. The net result was that the number of trials between successive encounters tended to be greater for infrequent species than for frequent species. Figure 6 shows examples of the decision-making phase of each experiment.

Experiment 2a. Participants completed two habitats. Each habitat had one rare and one common species that were healthy with proportion .7 and one rare and one common species that were healthy with proportion .3. Participants were not informed of the exact possible *p*(*healthy*) values. In each trial of the decision-making phase, one of the four species was highlighted (with frequencies matching those learned in the observation phase) and participants chose whether to eat or avoid a mushroom from that species. Healthy and poisonous eaten mushrooms were represented as green and red dots on a "histogram" of observations, whereas avoided mushrooms were represented as gray dots along with the samples from the observation phase. The decision-making phase of each habitat lasted 120 trials, but participants were not informed of when the habitat would end. Participants started with a bonus of $.50, and gained and lost money in increments of $.05. Potential bonuses were earned separately for the two habitats, and one of the two bonuses was

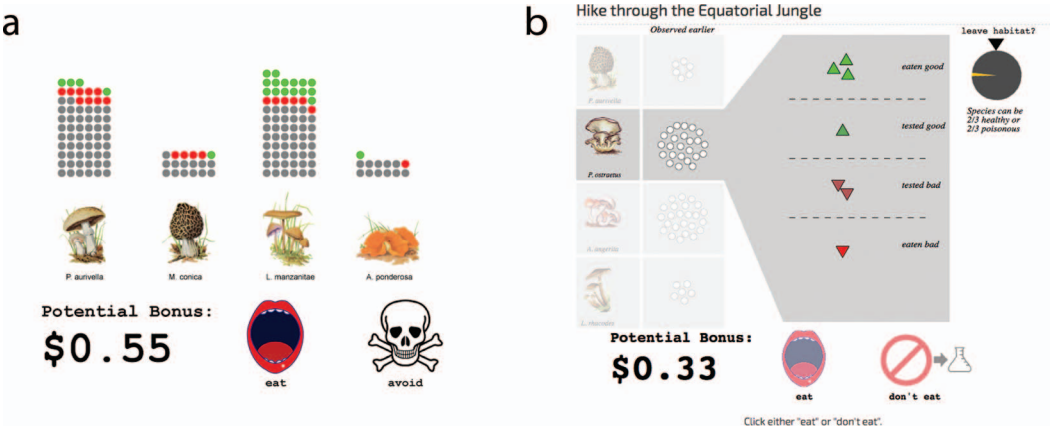

Figure 6. Example of the Experiment 2a task (a), and of the Experiment 2b task (b) in the full-information condition. See the online article for the color version of this figure.

randomly awarded at the conclusion of the experiment.

Experiment 2b. Participants completed four habitats. The task parameters and game interface were designed to be similar to that in Experiment 1b. Participants were informed that mushroom species could be either 2/3 healthy or 2/3 poisonous, and that these types were equally common. Mostly healthy and mostly poisonous species were pseudorandomly distributed across habitats, but each habitat contained at least one healthy species to maintain engagement (this was not revealed to participants in the instructions). The presentation of mushrooms was similar to Experiment 1b, and mushrooms from the observation phase were displayed separately from those from the decision-making phase. Participants were randomly assigned to either a contingent-information or a full-information condition. Upon making a decision a new circle or triangle appeared and was added to the appropriate group for that species as explained for Experiment 1b.

Rather than giving participants no information about habitat length, as in Experiment 2a, participants were informed that after each trial of the decision-making task there was a constant small probability of leaving the habitat and continuing to the next one. This probability was illustrated by a spinner with a small yellow wedge covering 1.5% of its area. The spinner was spun after each trial, and the habitat ended when the black arrow landed on the wedge. The habitat lengths were in fact predetermined to be 40, 60, 80, and 100 trials, randomly ordered. The bonus started at $.25 and changed in increments of $.02, and was cumulative over the four habitats.

Results

When asked in the "field report" how many times out of 10 each species would be encountered in the future, participants in Experiment 2a reported 4.7 for frequent species and 1.6 for rare species, whereas participants in Experiment 2b reported 4.8 for frequent species and 1.6 for rare species, on average. The target values based on the empirical frequencies were 4.0 and 1.0. Furthermore, participants gave the exact correct response 62% of the time, suggesting frequency information was encoded accurately, and critically that the difference between frequent and rare prospects was encoded.

Results of the decision-making phase are shown in Figure 7. Visual inspection shows a high initial exploration rate and greater approach probability for high-frequency species in Experiment 2a and the contingent-information condition of Experiment 2b, similar to the choices of the optimal policy (see Figure 5). In the full-information condition of Experiment 2b, people appeared to behave in a myopic reward-maximizing manner.

We analyzed the choice data using the hierarchical Bayesian logistic regression model introduced in Experiment Sequence 1, and predictors were calculated and rescaled as described above. Horizon was coded as 1 for high-frequency species and 0 for low-frequency species, and trial number represented the encounter number within the species encountered on that trial (rather than the trial number within the habitat). The posterior estimates of the population-level coefficients are presented in Figure 4, and posterior simulations from the model again confirmed a good qualitative match to the data (see Figure S3 in the supplemental material available online).

Participants were positively sensitive to expected reward across all three scenarios. In the two contingent-information scenarios, participants were highly exploratory early on, with 95% posterior predictive intervals for first-trial approach proportion of [.935, .960] and [.905, .936], and became less exploratory in later encounters.

Participants were also more exploratory in encounters with high-frequency species, supporting the hypothesis that they engaged in a forward-looking evaluation of the value of information. This effect of frequency was exhibited by a large portion of the participant population. In Experiment 2a, the posterior mean of the horizon parameter was greater than zero for 96% of participants, and the 95% posterior interval was entirely above zero for 64% of participants. In the contingent-information condition of Experiment 2b, the posterior mean was above zero for 86% of participants and the posterior interval was entirely above zero for 47% of participants. Participants did not exhibit the negative interaction between horizon and trial observed in Experiment Sequence 1, and in fact there was a small positive interaction in

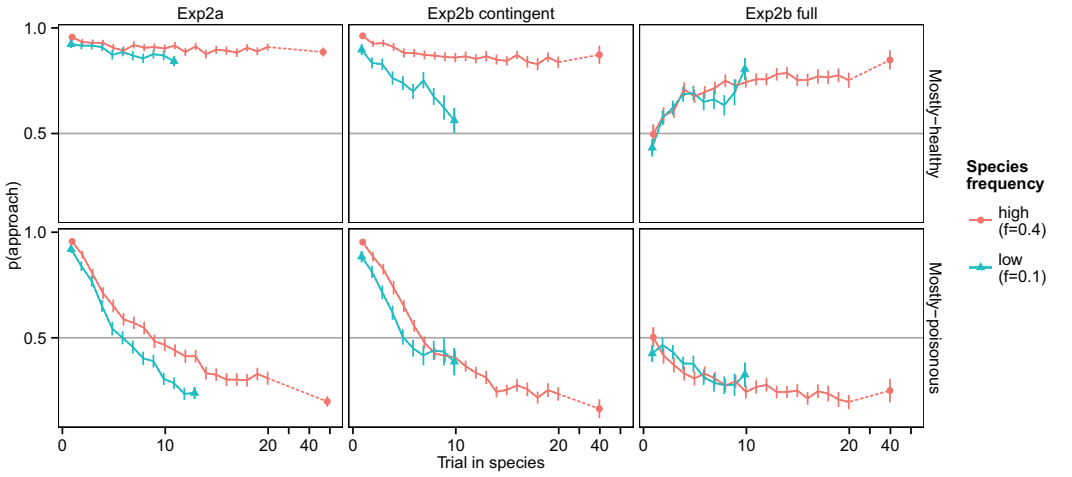

Figure 7. Probability of approaching a mushroom on each encounter within a patch for each infinite-horizon experiment and condition. Top panels show participant behavior for species where *p*(*healthy*) was .7 (Exp. 2a) or .67 (Exp. 2b), and bottom panels show behavior for patches where *p*(*healthy*) was .3 (Exp. 2a) or .33 (Exp. 2b). Error bars show standard error of the mean; error bars are wider on later trials in Experiment 2b because some habitats ended earlier than others. In order to focus on comparisons between patch frequencies, the *x* axis is compressed after Trial 20 and the change in p(approach) from Trial 20 to the final trial with a species is shown as a dotted line. Participants tended to approach early on and to have higher *p*(*approach*) for longer horizons in Experiment 2a and in the contingent-information condition of Experiment 2b, but not in the full-information condition of Experiment 2b. See the online article for the color version of this figure.

Experiment 2a. Further work is needed to investigate how task horizon interacts with past encounters and uncertainty.

In the full-information scenario, expected reward appeared to be the primary driver of behavior. The 95% posterior predictive interval for first-trial approach was [.450, .509], and there was no effect of horizon or encounter number. Furthermore, the 95% posterior interval for the effect of horizon was fully below those for the two contingent-information scenarios.

These findings appear to show a clear effect of frequency on approach behavior when information is choice-contingent. However, one alternative possibility is that participants simply were exploratory in the early trials of each habitat (rather than the early trials of each species), and became less exploratory later on, regardless of what species was encountered on a given trial. Because the n^{th} encounter with a frequent species will tend to come earlier in the habitat than the n^{th} encounter with an infrequent species, this could cause participants to falsely appear frequency-sensitive. We tested this explanation by fitting a logistic regression model to the first encounter with each mushroom species, using species frequency and the trial in the habitat on which the species was encountered to predict choice. There was a large effect of frequency on *p*(*approach*) in both contingent-information scenarios, $z = 2.58$, $p = .01$ and $z = 3.96$, $p < .001$, but no effect of trial in habitat, $z = .71$, $p > .250$ and $z = 0.06$, $p > .250$. Therefore, trial in habitat appears insufficient to explain our full pattern of results.

Discussion

Exploratory decisions are a constant feature of life in an uncertain and changing world, and people's choices about what to explore often determine their later behavior, preferences, and beliefs. Although work in applied fields has posited that people's decisions reflect the future value of exploring various prospects (Ching et al., 2013), experimental work has left unclear whether this is actually the case, particularly in naturalistic environments with an uncertain or "infinite" horizon. Across four experiments, we

found that people are indeed sensitive to the future value of information when making exploratory decisions. Participants increased their exploration as finite task horizon increased, and were sensitive not just to definite, discrete horizons but also to differences in prospect frequency when the horizon was uncertain.

These results show that people can use exploratory strategies that consider the future, but the mechanisms underlying their behavior remain unclear. Optimal information seeking requires complex dynamic programming calculations over possible future outcomes that seem implausible for humans to compute. People may perform simpler computations, such as simulating a few possible chains of future outcomes (Vul, Goodman, Griffiths, & Tenenbaum, 2014), or considering possible outcomes only one or two choices ahead (Zhang & Yu, 2013). Alternatively, they may use simple heuristics (e.g., Seale & Rapoport, 2000) or exploration bonuses (e.g., Daw et al., 2006) that are sensitive to the horizon. Finally, people may mix more and less complex exploration strategies, and the degree of forward-looking thinking may vary across individuals and situations as has been found in other decisions that involve future rewards (Frederick et al., 2002). Although distinguishing these potential mechanisms of forward-looking exploration will be difficult, many interesting avenues of research toward this goal are open. These include studying exploratory choice under cognitive load, investigating the developmental trajectory of forward-looking exploration, and testing the degree to which the exploration-enhancing effect of a long horizon is altered by the context of recently experienced horizons.

Regardless of the mechanisms driving their choices, our finding that people use frequency as a cue to the value of information may offer a better understanding of exploratory behavior in many domains. In our experiments, people were more likely to eat uncertain mushrooms from a species they expected to encounter frequently. If this pattern generalizes widely, then consumers may be more likely to purchase a new brand when shopping for commonly-purchased goods (Ching et al., 2013), and doctors may be more likely to prescribe a new drug for commonly-treated diseases (Chintagunta et al., 2012). Generally, frequency-dependent exploration should allow individuals to make better exploratory decisions compared to non-forward-looking agents (Denrell & March, 2001).

Interestingly, however, this individually rational behavior might amplify societal biases. Research on fads and social influence has described how self-reinforcing cycles of popularity can develop in consumer and cultural settings (Bikhchandani, Hirshleifer, & Welch, 1992; Denrell & Le Mens, 2007). Frequency-dependent exploration may strengthen these cycles, as items that are popular and thus common become more valuable to learn about than those that are unpopular and rare.

Similarly, work on social attitude formation has suggested that one cause of negative attitudes toward outgroups is that outgroup members are more easily avoided than ingroup members, allowing false beliefs about them to persist (Allport, 1979; Denrell, 2005). To the extent that outgroup members are rare in daily life, we predict this tendency to be exacerbated by frequency-dependent exploration. Fortunately, interventions that increase contact with an outgroup may erode prejudice (Shook & Fazio, 2008), possibly in part by increasing the future rewards of interacting and learning. To speculate, it is possible that even an intervention that simply increased a person's belief that outgroup members are a frequent part of their social environment might increase later exploratory interactions. Thus, although forward-looking exploration may cause biases, it might also be leveraged as a tool to reverse them.

References

Aguirregabiria, V., & Mira, P. (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics, 156,* 38–67. http://dx.doi.org/10.1016/j.jeconom.2009.09.007

Allport, G. W. (1979). *The nature of prejudice*. New York, NY: Basic Books.

Banks, J., Olson, M., & Porter, D. (1997). An experimental analysis of the bandit problem. *Economic Theory, 10,* 55–77. http://dx.doi.org/10.1007/s001990050146

Bellman, R. (1957). *Dynamic programming* (1st ed.). Princeton, NJ: Princeton University Press.

Berry, D. A., & Fristedt, B. (1979). Bernoulli one-armed bandits–Arbitrary discount sequences. *Annals of Statistics, 7,* 1086–1105. http://dx.doi.org/10.1214/aos/1176344792

Berry, D. A., & Fristedt, B. (1985). *Bandit problems*. London, UK: Chapman & Hall/CRC. http://dx.doi.org/10.1007/978-94-015-3711-7

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy, 100,* 992–1026. http://dx.doi.org/10.1086/261849

Camerer, C. F., & Weigelt, K. (1996). An asset market test of a mechanism for inducing stochastic horizons in experiments. *Research in Experimental Economics, 57,* 109–142.

Ching, A. T., Erdem, T., & Keane, M. P. (2013). Invited Paper—Learning models: An assessment of progress, challenges, and new developments. *Marketing Science, 32,* 913–938. http://dx.doi.org/10.1287/mksc.2013.0805

Chintagunta, P. K., Goettler, R. L., & Kim, M. (2012). New drug diffusion when forward-looking physicians learn from patient feedback and detailing. *Journal of Marketing Research, XLIX,* 1–43.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8,* e57410. http://dx.doi.org/10.1371/journal.pone.0057410

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. http://dx.doi.org/10.1038/nature04766

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review, 112,* 951–978. http://dx.doi.org/10.1037/0033-295X.112.4.951

Denrell, J., & Le Mens, G. (2007). Interdependent sampling and social influence. *Psychological Review, 114,* 398–422. http://dx.doi.org/10.1037/0033-295X.114.2.398

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science, 12,* 523–538. http://dx.doi.org/10.1287/orsc.12.5.523.10092

Erdem, T., & Keane, M. P. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science, 15,* 1–20. http://dx.doi.org/10.1287/mksc.15.1.1

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature, 40,* 351–401. http://dx.doi.org/10.1257/jel.40.2.351

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. London, UK: Taylor & Francis.

Grosskopf, B., Erev, I., & Yechiam, E. (2006). Foregone with the wind: Indirect payoff information and its implications for choice. *International Journal of Game Theory, 34,* 285–302. http://dx.doi.org/10.1007/s00182-006-0015-8

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., . . . Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods, 48,* 829–842.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15,* 534–539. http://dx.doi.org/10.1111/j.0956-7976.2004.00715.x

Howard, R. A. (1966). Information value theory. *Systems Science and Cybernetics, 2,* 22–26.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2016). Information sampling behavior with explicit sampling costs. *Decision, 3,* 147–168. http://dx.doi.org/10.1037/dec0000045

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology, 2,* 398.

Kreps, D. M., & Porteus, E. L. (1978). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica, 46,* 185–200. http://dx.doi.org/10.2307/1913656

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research, 12,* 164–174. http://dx.doi.org/10.1016/j.cogsys.2010.07.007

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., . . . Fiedler, K. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2,* 191–215. http://dx.doi.org/10.1037/dec0000033

Meyer, R. J., & Shi, Y. (1995). Sequential Choice Under Ambiguity: Intuitive Solutions to the Armed-Bandit Problem. *Management Science, 41,* 817–834. http://dx.doi.org/10.1287/mnsc.41.5.817

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology, 7,* e1001048. http://dx.doi.org/10.1371/journal.pcbi.1001048

Seale, D. A., & Rapoport, A. (2000). Optimal stopping behavior with relative ranks: The secretary problem with unknown population size. *Journal of Behavioral Decision Making, 13,* 391–411. http://dx.doi.org/10.1002/1099-0771(200010/12)13:4<391::AID-BDM359>3.0.CO;2-I

Shook, N. J., & Fazio, R. H. (2008). Interracial roommate relationships: An experimental field test of the contact hypothesis. *Psychological Science, 19,* 717–723. http://dx.doi.org/10.1111/j.1467-9280.2008.02147.x

Speekenbrink, M., & Konstantinidis, E. (2014). Uncertainty and exploration in a restless bandit task. *Proceedings of the 36rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Stan Development Team. (2015). *Stan: A C++ library for probability and sampling* (Version 2.7).

Stange, K. M. (2012). An empirical investigation of the option value of college enrollment. *American Economic Journal Applied Economics, 4,* 49–84. http://dx.doi.org/10.1257/app.4.1.49

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

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110. http://dx.doi.org/10.1037/h0031322

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science, 38,* 599–637. http://dx.doi.org/10.1111/cogs.12101

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General, 143,* 2074–2081. http://dx.doi.org/10.1037/a0038199

Yechiam, E., & Busemeyer, J. R. (2006). The effect of foregone payoffs on underweighting small probability events. *Journal of Behavioral Decision Making, 19,* 1–16. http://dx.doi.org/10.1002/bdm.509

Zhang, S., & Yu, A. J. (2013). Forgetful Bayes and myopic planning: Human learning and decision-making in a bandit setting. *Advances in Neural Information Processing Systems, 26,* 2607–2615.

Received March 28, 2016
Revision received November 18, 2016
Accepted December 6, 2016 ■