The value of approaching bad things

Alexander S. Rich (asr443@nyu.edu)
Todd M. Gureckis (todd.gureckis@nyu.edu)
New York University, Department of Psychology, 6 Washington Place, New York, NY 10003 USA

Abstract

Adaptive decision making often entails learning to approach things that lead to positive outcomes while avoiding things that are negative. The decision to avoid something removes the risk of a negative experience but also forgoes the opportunity to obtain information, specifically that a seemingly negative option is actually positive. This paper explores how people learn to approach or avoid objects with uncertain payoffs. We provide a computational-level analysis of optimal decision making in this problem which quantifies how the probability of encountering an object in the future should impact the decision to approach or avoid it. A large experiment conducted online shows that most people intuitively take into account both their uncertainty and the value of information when deciding to approach seemingly bad things.
Keywords: decision making, approach-or-avoid behavior, value of information, sequential decision making

From talking to a stranger to trying a new restaurant, people often enter into situations with uncertain outcomes. Approaching, sampling, or interacting with unknown options allows highly rewarding situations to be discovered, but also carries the risk of encountering negative outcomes (e.g., an awkward conversation or bad meal). As a result, there is often a tension between approaching new alternatives to discover which are positive and avoiding those that could be negative.

Certain forms of adaptive approach-or-avoid behavior can lead to systematically suboptimal behavior (March, 1996; Denrell, 2007). For example, suppose there is a lecture series that features an engaging speaker on most days but which also has the occasional boring talk. You attend the series for the first time at the start of the semester, happen to hear a boring talk, and decide not to attend in the future. As a result, you gain no information about how good or bad the subsequent talks actually are.

The tendency to avoid alternatives and cease learning about them after an early negative experience is known as the "hot stove effect" and has been used to explain suboptimal behavior and risk-aversion in both individuals and organizations (Denrell & March, 2001; Denrell, 2005). While adaptive (in the sense of changing based on experience), this strategy is suboptimal because it doesn't take into account the value of the information gained by approaching or interacting with an uncertain alternative. After a single boring talk, your uncertainty about the seminar series might remain high. If you go to a second lecture and it is also boring, your uncertainty in your belief should drop and you likely will not attend a third.

But suppose that the second lecture is really good. This new information should help revise your belief about the quality of the series, and increase the chances you will go to a third lecture. If the third is also good, you will likely go to a fourth, and so on until you have enjoyed a whole semester of mostly-good lectures. The critical point is that the potential future payoff of attending that second lecture if you were wrong about the series being boring is far greater than the potential loss from attending a few bad lectures if you were right.

This intuitive example highlights a key aspect of approach-avoid decisions. First, optimal decision making should take into account not just the estimated valence (good or bad) of an option but also the *uncertainty* about that estimate (Berry & Fristedt, 1985; Denrell, 2007; Daw et al., 2006). Second, it can be optimal in some cases to approach seemingly negative outcomes due to the potential gain that could be experienced *in the future*. This is due to the *value of information* which is obtainable from continued approach decisions.

The goal of the present paper is to explore if and how people intuitively utilize these aspects of approach/avoidance decision making. While there is a large literature on the effect of experience dependent sampling on approach/avoid behavior (e.g., Denrell & March, 2001; Niv et al., 2002; Biele et al., 2009; Fazio et al., 2004), there have been fewer empirical tests of the idea that decision makers consider both uncertainty and the value of information when making such decisions (but see Meyer & Shi, 1995).

A computational analysis of approach-avoidance decision making

The scenario we consider here concerns a decision maker who is presented with a single prospect (which might represent an object, a person, a product, or a situation) at each point in time and must decide to either approach or avoid it. Approach decisions result in experience with the prospect which may be either negative or positive. Avoid decisions result in no experience. Such a task corresponds to many real-world decision problems people face in their lives. For scientists this includes which seminar series to attend and which to skip.

Optimal approach-avoidance decision making Consider the case in which a single prospect with Bernoulli payoffs may be approached or avoided an indefinite number of times. When it is approached, it has unknown probability p of yielding a reward of 1, and probability $1-p$ of yielding a reward of -1. Given a subjective belief about the value of p described by a Beta distribution, $Beta(\alpha,\beta)$, the agent's prediction about the probability of reward is $E[Beta(\alpha,\beta)] = \frac{\alpha}{\alpha+\beta}$. The posterior belief about p is updated to $Beta(\alpha+1,\beta)$ after an additional positive outcome, and to $Beta(\alpha,\beta+1)$ after an additional negative outcome. A natural initial setting is $\alpha=\beta=1$ which corresponds a uniform subjective prior distribution for p over the range [0,1]. Figure 1 (left panel)

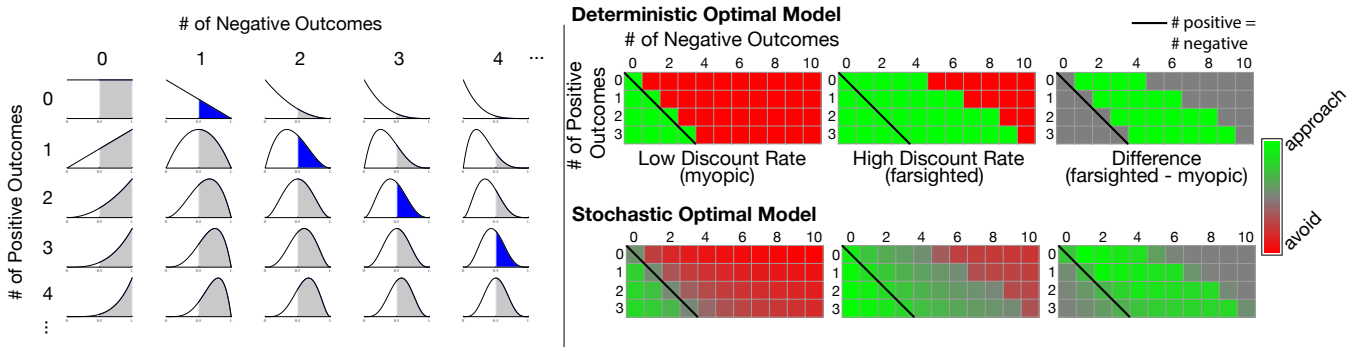

Figure 1: *Left*: Evolution in the posterior belief in the probability of a positive outcome p as a function of experienced outcomes. The model starts with a uniform prior belief at (0,0) in the grid ($Beta(\alpha = 1, \beta = 1)$). The grey region depicts the area under the probability density function for $p > 0.5$ (i.e., mainly positive outcomes) Posteriors with one more negative than positive outcome are highlighted in blue. *Top right*: Optimal choice policies for agents with a low discount rate ($\gamma = 0.0$, left), high discount rate ($\gamma = 0.999$, center) and the difference between the choice policies (right). Grid axes denote different numbers of positive and negative experiences with the prospect. *Bottom right*: An illustration of two stochastic policies ($\gamma = \{0.0, 0.999\}, \tau = 1, \varepsilon = 0.1$). Ultimately the assumptions tend to "blur" the hard line between approach and avoid in the optimal policy and make the probability of approaching slightly higher for negative prospects.

illustrates how posterior belief about the prospect changes as positive and negative experiences accumulate.

Given a current observation set of α positive and β negative outcomes, the total expected value of following the optimal policy is $V(\alpha,\beta) = \max\{Q_{approach}(\alpha,\beta), Q_{avoid}\}$ where

$$\begin{aligned} Q_{approach}(\alpha,\beta) &= E[Beta(\alpha,\beta)][1+\gamma V(\alpha+1,\beta)] \\ &\quad + (1-E[Beta(\alpha,\beta)])[-1+\gamma V(\alpha,\beta+1)] \\ &= \frac{\alpha}{\alpha+\beta}[1+\gamma V(\alpha+1,\beta)] \\ &\quad + \frac{\beta}{\alpha+\beta}[-1+\gamma V(\alpha,\beta+1)] \end{aligned} \tag{1}$$

and $Q_{avoid} = 0$. The optimal decision policy is to approach when $Q_{approach}$ is greater than Q_{avoid}, and avoid otherwise.

The optimal decision strategy depends on a recurrence relation between $V(\alpha,\beta)$, $V(\alpha+1,\beta)$, and $V(\alpha,\beta+1)$ To solve the relation, it is possible to estimate V for all pairs $\alpha+\beta = N$ for some large N and then work backwards towards $V(1,1)$ using dynamic programming (Gittins et al., 2011).

In Equation 1, γ is a free parameter denoting the degree to which future rewards are discounted. When $\gamma \to 0$ the optimal policy cares only about immediate reward and will not consider the value of information possible from approaching. As $\gamma \to 1$, the policy begins to take into account future encounters with the prospect and uncertainty about the current estimate. The dependence of the current value of an action upon the future (weighted by γ) is what helps the model to avoid the "hot stove effect" (but see Denrell, 2007).

To help the reader develop an intuition, Figure 1 (right panel, top) shows the optimal policy of two hypothetical agents that value future rewards to different degrees. Each agent ceases to approach the alternative when the potential future payoffs, weighted by their likelihood, become less valuable than the potential future losses. Because the potential positive outcomes are distributed far out into the future, while the potential negative outcomes–the few more bad trials before avoidance begins–are expected to occur quickly, this threshold changes as a function of how much the model values the future.

Observe also that as the number of positive outcomes experienced by the farsighted agent grows, the region in which a seemingly negative prospect is still sampled becomes wider. For example, with zero positive outcomes the farsighted model stops sampling at five negative outcomes, but at three positive outcomes the model stops sampling after not eight but ten negative outcomes. This is because as the number of total experiences increases, it takes more negative outcomes relative to positive ones to become highly certain that the prospect is negative. This can be seen most clearly by examining the posteriors highlighted in blue in the left panel of Figure 1. In all of these cases, there has been one more negative outcome than positive outcome. But as the total number of trials increases the proportion of the posterior where $p > 0.5$ grows larger, so the uncertainty about whether the prospect is positive increases as well. Thus, one behavioral signature of the optimal model is a region of uncertainty in which the agent continues to sample negative prospects and that grows with the total number of experiences.

Optimal decision making among multiple prospects The logic behind the model just described generalizes to a situation where instead of one prospect, there are multiple, which have independent probabilities of positive outcome p_i and which appear with some base rate or frequency f_i, $\sum f_i = 1$ (e.g., each day a scientist can go or not go to a different seminar, some of which meet more frequently than others). If there is no uncertainty about the identity of the prospect presented on a given trial and assuming that the value of p_i is independently sampled from a uniform distribution for each option, the optimal approach policy for each prospect can be calculated independently.

However, differences in base rate of occurrence between different prospects has a subtle influence on the optimal policy. In particular, the n^{th} future experience with a rare

prospect is further away in time than the n^{th} future experience with a common prospect. One way to account for this difference is through adjustment to the temporal discounting parameter (γ) for each prospect. If the base value of the next trial compared to the current one is γ_b, then the expected value of the next trial for an alternative occurring with frequency f is

$$\gamma_f = \sum_{n=1}^{\infty} f \cdot (1-f)^{n-1} \cdot (\gamma_b)^n \tag{2}$$

which decreases monotonically as f decreases. This means that an optimal agent faced with rare and common prospects will behave as though it values the future less when encountering rare prospects, and will begin avoiding these prospects with less negative evidence (similar to the difference depicted in Figure 1, top right).

Comparing human behavior and the optimal model An optimal agent will approach a prospect on every trial until it determines with sufficiently high certainty that the prospect is negative, and then will never approach again. However, it is plausible that human decision makers behave in a way generally consistent with the model but decide more stochastically. Following typical assumptions in the decision making literature (Luce, 1959; Sutton & Barto, 1998), we assume the model's probability of approach, P_m is

$$P_m(\text{approach}) = \frac{e^{Q_{approach} \cdot \tau}}{e^{Q_{approach} \cdot \tau} + e^{Q_{avoid} \cdot \tau}} \tag{3}$$

Where τ is a parameter that determines how deterministic the participant is in responding, and Q_{approach} and Q_{avoid} are the expected values of approaching or avoiding on the current trial and subsequently following the optimal policy (α and β are implicit in the notation now). In general, the probability of approaching is an increasing function of its value relative to avoiding.

In addition we observed that participants occasionally approached a negative prospect even after avoiding it for several trials. This behavior is not well captured by Equation 3 without an extreme value of τ (near zero) but may plausibly reflect an incorrect assumption about the non-stationarity of reward probability for the prospects (basically "checking" to see if reward probabilities have changed, e.g., Tversky & Edwards, 1966; Knox et al., 2011). To account for these choices, whatever their underlying cause, we let

$$P(\text{approach}) = (1-\varepsilon)P_m(\text{approach}) + \varepsilon \tag{4}$$

such that the model follows the stochastic optimal rule with probability $(1-\varepsilon)$ and approaches regardless of $Q_{approach}$ with probability ε. With $\varepsilon > 0$, the model has a small constant probability of approaching on any trial. Figure 1 (right panel, bottom) illustrates how the form of the optimal policy is generally modified by these additional assumptions.

Two key principles of the optimal model The optimal decision maker just described exposes two key behavioral principles. First, the decision to approach or avoid is strongly influenced by the current uncertainty in the estimate of p associated with a prospect. Second, the optimal decision policy for $\gamma > 0$ takes into account the value associated with future interaction with a prospect. This interacts in an interesting way with the base rate or frequency of the prospect; information gained about a rare prospect is expected to be applicable less frequently and thus has less utility, such that at the same level of overall uncertainty it is more advantageous to approach a frequent prospect than an infrequent one.

Experiment

To investigate the effects of base rate and uncertainty on approach-avoid decisions, we created an online video game called the Mushroom Game. In the game, participants played the role of field biologists cataloging and learning about the edibility of mushroom species growing in different habitats. The base rate of occurrence for different mushrooms was manipulated across the environments and our goal was to see if participants adjusted their approach/avoid behavior in line with the predictions of the optimal model.

Method

Participants One hundred fifty-two participants (65 women and 87 men) age 18 to 66 years ($M = 33.0$, $SD = 10.2$) completed the task via Amazon Mechanical Turk. All participants were paid \$2 for participation with the possibility of earning a bonus that averaged \$1.02 and ranged from \$0 to \$1.60. Participants were instructed on all aspects of the task and were required to pass a quiz demonstrating comprehension of the instructions before entering the experiment. Three participants required more than three tries to pass a quiz on the instructions, and were excluded from all further analyses.

Materials Each participant played the Mushroom Game in two habitats, which shared the same overall structure. Each habitat contained four unique mushroom species which were taken from illustrations of actual mushroom species found online.

Procedure and design The experiment was divided into two "habitats" within which the participant was asked to learn about local mushroom species (e.g., "New England Forest", or "Amazonian Rainforest"). Within each habitat there were four distinct mushroom species, two of which occurred with frequency $4/10$ and two of which appeared with frequency $1/10$. One high-frequency and one low-frequency species were healthy (i.e., rewarding) with probability 0.7, while the other two species were poisonous (i.e., punishing) with probability 0.7. The assignment of base rates, reward probabilities, and the identity of fictitious mushroom species was randomly determined for each participant.

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 by gray dots which appeared on the screen without participant input. Once the entire sample has 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.

Figure 2 shows an example of the interface of the game. The main feature of the interface is the "Field Log", a row of icons representing the local species with dots above each icon to represent the mushrooms that have been observed from that species. These dots effectively form a histogram showing the relative frequency of the species. At the end of the first phase, the screen would look similar to Figure 2 except with only gray dots visible.

In the decision-making phase, participants' goal was to learn which of the habitat's mushroom species were healthy and which

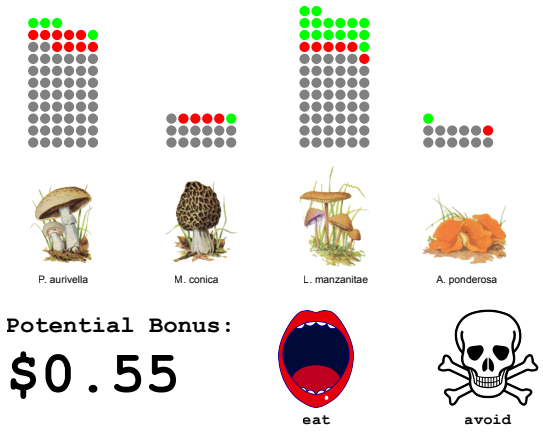

Figure 2: An example of the task interface from the decision phase (see text for description).

were poisonous by eating or avoiding the mushrooms they came across. During each trial of the decision-making phase, participants encountered a randomly selected mushroom and its species was highlighted. The probability a mushroom was selected on any trial was matched to the relative frequencies from the first phase. Participants then chose to either eat or avoid the mushroom by clicking the appropriate button. Eating a mushroom revealed whether the mushroom was healthy or poisonous, and added a green or red dot to the species histogram on top the observation-phase observations (which remained visible). Avoiding a mushroom revealed no information about its healthiness, and added a gray dot like those seen in the first phase to the species histogram. Since each mushroom species had both a non-zero probability of being healthy and a non-zero probability of being poisonous, participants had to sample each species more than once to gain an accurate estimate of its healthiness.

To aid visual estimation of the outcomes and frequencies, the dots were organized on the screen so that grey dots appeared below red dots which in turn appeared below green ones as seen in Figure 2.

Participants earned a cash bonus based on their ability to eat healthy mushrooms while avoiding poisonous ones. Each participant's potential bonus started at \$0.50, and increased by \$0.05 for each healthy mushroom eaten but was reduced \$0.05 for every unhealthy mushroom eaten. Avoiding a mushroom had no effect on the potential bonus. At the end of the experiment, the actual bonus was chosen at random from the potential bonuses earned in the first and second habitat.

Results

Performance Participants generally learned which species were usually healthy and which were usually poisonous, approaching healthy species with probability 0.89 (SD = 0.16) and poisonous species with probability 0.38 (SD = 0.20). The average potential bonus was \$1.08, compared with an average of \$1.36 earned by a deterministic optimal model with $\gamma = 0.995$. We found no difference in the probability of approaching healthy and poisonous species or the bonus earned between participants' first and second habitat and so collapse across habitats for all analyses. Participants' predictions in the "field report" about how many times out of ten they would see each species in the future were on average 0.91 off from the true frequencies (SD = 1.01), showing that they were able to estimate base rates with reasonable accuracy given histogram data.

Model-Based Analysis The predicted relationship between uncertainty, base rates, and approach/avoid decision is complex and multivariate. Thus, the most direct test of our hypothesis is obtainable through model-based analyses of participants' trial-by-trial choice behavior. We consider six different models.

Stochastic optimal model with fixed base-rate (SO-F). The first model behaves in accordance with the optimal model described above but assumes that each of the four mushrooms occurs with probability $1/4$ on each trial. This model effectively represents our null hypothesis that the manipulation of base rate has no effect on participants' choice behavior. This model is endowed with three free parameters per participant (γ, τ, and ε).

Stochastic optimal model with variable base-rate (SO-V). The second model represent the alternative hypothesis that participants adjust their sampling behavior based on the base rate of the mushrooms. This model is identical to the SO-F model but adds one additional parameter: a freely varying base rate parameter f_l representing the (subjective) frequency of each of the low-frequency mushroom. The base rates of the high-frequency mushrooms were set to $(1-2f_l)/2$.

Critically, if participants behave optimally but do not take the difference in the value of information caused by the different base rates into account in their sampling choices, they will be better fit by SO-F. If they do take the difference of base rates into account, they will be better fit by SO-V.

We also tested several alternative models which exhibit similar properties to the optimal model but which should be distinguishable based on behavioral data.

Random choice model. The first model is a baseline which assumes that participants chose to approach all mushrooms with some constant probability p.

Softmax reinforcement learning model (SM) . The SO-F and SO-V models can be contrasted with a standard non-forward-looking reinforcement learning based model (Sutton & Barto, 1998). In this model, the probability of approaching is based solely on the estimate of the value of approaching learned from experience so far and the model is thus susceptible to the hot stove effect (Denrell, 2007). The estimate of the value is updated after each approach according to $Q_{approach} = Q_{approach} + \alpha \cdot \delta$ where $\delta = r - Q_{approach}$, r is the received reward, and α is a learning rate/recency parameter ($0 \leq \alpha \leq 1$) that controls the degree to which the current estimate depends on the most recent rewards. The probability of approaching is again determined by a probabilistic choice rule, given by

$$P_m(\text{approach}) = \frac{e^{Q_{approach} \cdot \tau}}{e^{Q_{approach} \cdot \tau} + 1} \quad (5)$$

To this choice rule we added the same ε parameter as in the optimal model, and used equation 4 to determine the actual approach probability.

This model is an interesting competitor to the optimal model for a few reasons. First, it maintains only a point estimate of the value of each decision option rather than the full

Table 1: Summary of model fits

Name	# Param.	BIC(mean)	% best fit
SO-V	4	174.0	56
SO-F	3	179.4	21
SM-V	5	189.8	7
SM-F	4	189.6	4
SM-0	3	202.7	6
Rand.	1	302.1	7

Table 2: Median(mean) optimal model parameters

Name	γ	τ	ε	f_l
SO-V	0.99(0.91)	0.29(0.39)	0.05(0.09)	0.05(0.08)
SO-F	0.80(0.70)	0.14(0.43)	0.07(0.11)	

uncertainties depicted in Figure 1 (left). Second, it is not sensitive to the future utility of particular outcomes. It can however mimic some aspects of the optimal policy. For example, if the model is given an additional set of parameters corresponding to the initial values of $Q_{approach}$ for each prospect, this can be used to encourage exploratory approach behavior since it may take several negative outcomes to lower the $Q_{approach}$ value past zero ("optimistic initialization" in the RL literature Sutton & Barto, 1998). The softmax model may also account for differences in behavior between base rates by allowing separate optimistic $Q_{approach}$ for high-frequency and low-frequency mushrooms. However, the model provides no a-priori rationale for such parameter differences.

We fit participants to three versions of the softmax model. The SM-0 had initial values for $Q_{approach}$ set to zero for all prospects, and is most similar to the adaptive models described by Denrell & March (2001) and Denrell (2007). The fixed starting Q model (SM-F) allowed the initial $Q_{approach}$ to vary per participant, but set it equal for all mushrooms. The variable starting Q model (SM-V) allowed the initial $Q_{approach}$ to be set to separate values for high-frequency and low-frequency mushrooms. The number of parameters per participant for each of these models is summarized in Table 1.

Model comparison We fit each of the six models to individuals' trial-by-trial choices to maximize the likelihood of participants' approach/avoid decisions. We then calculated the Bayesian Information Criterion (BIC) for each model, which compares the quality of the model fits while penalizing models for their number of free parameters.

Average model BIC, and the proportion of subjects best fit by each model, are shown in Table 1. The SO-V model, which used a probabilistic version of the optimal choice rule and was sensitive to differences in mushroom frequency, provided the best fit for more than half of all participants. The SO-F model provided the best fit for roughly twenty percent of participants, while the three softmax models and the random model best accounted for the final quarter of participants. The mean and median parameters for subjects best fit by the SO-V and SO-F models are listed in Table 2.

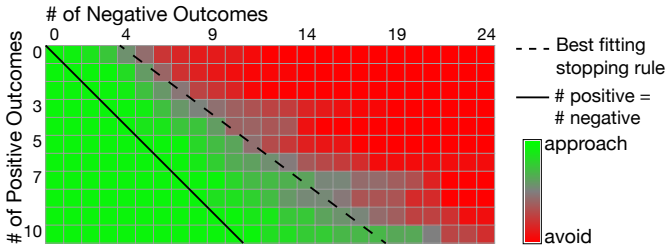

Figure 3: Approximate approach policy of participants for frequent, usually-poisonous mushrooms. To construct this figure, we first determined the positive-negative grid location after the final trial for each experienced mushroom. Then for each square within each row of the grid, we determined the proportion of participants who sampled exactly that many positive outcomes who also sampled *at least* that many negative ones. These proportions were shaded from green (1.0) to red (0.0), providing an estimate of participants' likelihood of continuing to sample in a given grid cell. Linear regression was used to find the best-fitting average stopping rule.

Do people take uncertainty into account? Figure 3 shows choice behavior for the high-frequency, usually-poisonous mushrooms, from which the most data on participants' approach policy towards negative prospects is available (see caption). As can be seen from comparison to the solid "# positive = # negative" line, virtually all participants continue to sample from seemingly negative alternatives for several more negative trials before avoiding the prospect. This is consistent with the generally high discount parameter γ of participants best fit by the optimal models[1].

Comparison of the "# positive = # negative" line to the best-fitting stopping rule shows that the region of uncertainty in which participants continued to sample seemingly negative prospects widened as the number of total experiences grew, as predicted by the optimal model. This observation also gives insight into why the softmax models do not provide a good fit to the data, even with optimistic initial Q values. While initial optimism can explain perseverance through the first few negative outcomes from a prospect, as the number of trials increases, the agent's optimism is washed away by experience. In other words, the point-estimate Q-values regress to the mean of the sample.

Thus if participants were behaving like softmax agents, their stopping rule would be expected to converge to the "# positive = # negative" line as the number of total outcomes increased. The fact that it instead moves farther away suggests that people are not purely adaptive to past experience in their approach decisions, and instead look towards the future and take their uncertainty into account.

Do people take base rate into account? Among the 83 participants best fit by the SO-V model, 78 had $f_l < 0.25$, meaning their behavior was best fit by an optimal model which correctly assumed the low-frequency prospects occurred less often than the high-frequency ones. The mean best-fit f_l (see Table 2) is less than the true frequency of $1/10$,

[1] While the mean γ across the two models is only 0.85, this low value is due to a strong left skew; 47% of these participants had $\gamma > 0.99$, and 71% had $\gamma > 0.9$.

indicating that participants actually responded to the base rate manipulation more strongly than was optimal and undersampled rare items relative to common ones. This behavior is consistent with evidence that people underestimate the probability of rare events when they learn about them through experience rather than description (Hertwig et al., 2004).

The right panel of Figure 4 shows the choice policies of participants best fit by SO-V, along with the difference between their policies for high- and low-frequency prospects and the predicted difference based on SO-V model simulation. As shown at bottom left, participants tended to persist in sampling a negative prospect slightly longer when the prospect occurred frequently. This trend is consistent with participants' low f_l parameter fits, and its shape is similar to the model simulation (bottom right).

Conclusions

In summary, participants were able to modulate their approach behavior in response to their beliefs about the future utility of those interactions. While humans may not *generally* make approach-avoid decisions in an optimal fashion, at the very least they appear to take into account two decision variables (uncertainty and the value of information) in a way consistent with optimal sequential decision making policies. This finding is interesting in light of the large literature on the "hot stove effect" (e.g., Biele et al., 2009; Fazio et al., 2004) and also contrasts with findings which show that exploration behavior in multi-armed bandit tasks is not particularly sensitive to uncertainty (Daw et al., 2006). Interestingly, previous studies of bandit tasks with short, finite horizons have found that people tended to choose uncertain alternatives more when the horizon was longer (Lee et al., 2011; Meyer & Shi, 1995). However, the present experiment is the first, to our knowledge, to show that people are sensitive to the future value of information in a more naturalistic, indefinite-horizon environment.

Unlike the present task, in real-world environments people are often face additional uncertainty about the category membership of individual prospects (e.g., "Is this mushroom the same species as another?"). While research into these kinds of more complex approach-or-avoid problems has thus far only been considered in machine learning (e.g. Guez et al., 2013, 2012), we hope that the current study is a step towards understanding how humans learn through experience-based interactions with their environment.

Acknowledgments. This work was supported by grant number BCS-1255538 from the National Science Foundation and the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023 to TMG.

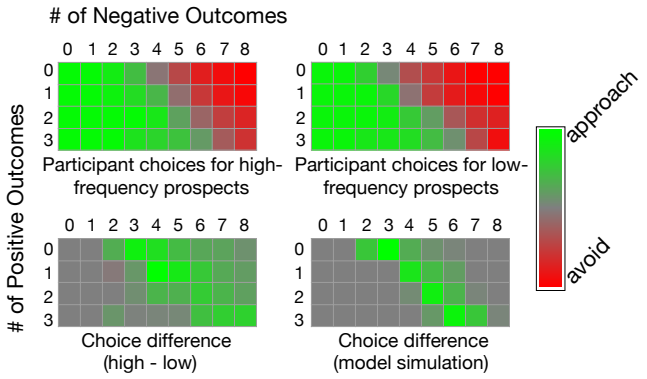

Figure 4: *Top:* Approximate approach policies calculated for high- and low-frequency, usually-poisonous mushrooms, for participants best fit by SO-V. Method of generation matches that described in Figure 3, except only the first 11 trials with each prospect are considered. *Bottom:* Difference in participants' approach policy between high-and low-frequency prospects, with model simulation from SO-V using median best-fit parameters.

References

Berry, D., & Fristedt, B. (1985). *Bandit problems*. London: Chapman & Hall/CRC.

Biele, G., Erev, I., & Ert, E. (2009). Learning, risk attitude and hot stoves in restless bandit problems. *Journal of Mathematical Psychology*, *53*(3), 155–167.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879.

Denrell, J. (2005). Why most people disapprove of me: experience sampling in impression formation. *Psychological review*, *112*(4), 951-978.

Denrell, J. (2007). Adaptive learning and risk taking. *Psychological review*, *114*(1), 177.

Denrell, J., & March, J. G. (2001). Adaptation as information restriction: The hot stove effect. *Organization Science*, *12*(5), 523–538.

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: valence asymmetries. *Journal of personality and social psychology*, *87*(3), 293–311.

Gittins, J., Glazebrook, K., & Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons, Ltd.

Guez, A., Silver, D., & Dayan, P. (2012). Efficient bayes-adaptive reinforcement learning using sample-based search. *Advances in Neural Information Processing Systems*, *25*, 1034–1042.

Guez, A., Silver, D., & Dayan, P. (2013). *Towards a practical bayes-optimal agent.* (Poster presented at The 1st Multidisciplinary Conference on Reinforcement Learning and Decision Making.)

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*(8), 534–539.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2011). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in psychology*, *2*.

Lee, M. D., Zhang, S., Munro, M., & Steyvers, M. (2011). Psychological models of human and optimal performance in bandit problems. *Cognitive Systems Research*, *12*(2), 164–174.

Luce, R. D. (1959). *Individual choice behavior: a theoretical analysis*. John Wiley and sons.

March, J. G. (1996). Learning to be risk averse. *Psychological Review*, *103*, 309-319.

Meyer, R. J., & Shi, Y. (1995). Sequential choice under ambiguity: Intuitive solutions to the armed-bandit problem. *Management Science*, *41*(5), 817–834.

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

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1) (No. 1). Cambridge Univ Press.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology*, *71*(5), 680.