Psychological Review
2013, Vol. 120, No. 1, 139–154

0033-295X/13/$12.00 DOI: 10.1037/a0030850

Melioration as Rational Choice: Sequential Decision Making in Uncertain Environments

Chris R. Sims
University of Rochester

Hansjörg Neth
Max Planck Institute for Human Development

Robert A. Jacobs
University of Rochester

Wayne D. Gray
Rensselaer Polytechnic Institute

Melioration—defined as choosing a lesser, local gain over a greater longer term gain—is a behavioral tendency that people and pigeons share. As such, the empirical occurrence of meliorating behavior has frequently been interpreted as evidence that the mechanisms of human choice violate the norms of economic rationality. In some environments, the relationship between actions and outcomes is known. In this case, the rationality of choice behavior can be evaluated in terms of how successfully it maximizes utility given knowledge of the environmental contingencies. In most complex environments, however, the relationship between actions and future outcomes is uncertain and must be learned from experience. When the difficulty of this learning challenge is taken into account, it is not evident that melioration represents suboptimal choice behavior. In the present article, we examine human performance in a sequential decision-making experiment that is known to induce meliorating behavior. In keeping with previous results using this paradigm, we find that the majority of participants in the experiment fail to adopt the optimal decision strategy and instead demonstrate a significant bias toward melioration. To explore the origins of this behavior, we develop a rational analysis (Anderson, 1990) of the learning problem facing individuals in uncertain decision environments. Our analysis demonstrates that an unbiased learner would adopt melioration as the optimal response strategy for maximizing long-term gain. We suggest that many documented cases of melioration can be reinterpreted not as irrational choice but rather as globally optimal choice under uncertainty.

Keywords: melioration, rational analysis, Bayesian modeling, sequential decision making

Supplemental materials: http://dx.doi.org/10.1037/a0030850.supp

Daily life consists of an uninterrupted sequence of choices. Many of these choices concern the relative allocation of behavior between competing alternatives. For example, after a long day of work, we may face a choice between exercising versus collapsing on the couch, or cooking a meal versus ordering takeout. Compared to most laboratory tasks on decision making, real-life decisions such as these possess two important properties. First, many real choices have immediate, as well as delayed, consequences. Second, not only do our actions have delayed consequences, they can also interact in nontrivial ways with the perceived value of competing alternatives or even the same alternative on later decisions. For instance, adopting a sedentary lifestyle can not only harm our long-term health and happiness but can also render the immediate prospect of physical exercise more unappealing.

How do humans navigate decisions with delayed and indirect consequences? Despite its simplicity, this question encompasses over a century of research in human and animal learning (Thorndike, 1911) and is encountered at multiple levels of analysis—from understanding the computations of individual neurons (Gold & Shadlen, 2007) to studying higher level faculties like mental representations and goal-directed cognition (Daw & Frank, 2009). Lying dormant in much of this work is the key assumption that human choice is at its core largely consistent with the framework of rational choice theory (M. Friedman & Savage, 1948; Von Neumann & Morgenstern, 1944). According to this framework, the driving force behind our actions is the overall maximization of expected utility—that is to say, we seek our betterment as we each have defined it.

This article was published Online First December 10, 2012.

Chris R. Sims, Department of Brain and Cognitive Sciences, University of Rochester; Hansjörg Neth, Center for Adaptive Behavior and Cognition (ABC), Max Planck Institute for Human Development, Berlin, Germany; Robert A. Jacobs, Department of Brain and Cognitive Sciences, University of Rochester; Wayne D. Gray, Department of Cognitive Science, Rensselaer Polytechnic Institute.

This work was supported, in part, by Office of Naval Research Grant N000140710033 and Air Force Office of Scientific Research Grant FA9550-06-1-0074 to Wayne D. Gray, as well as National Science Foundation Grant DRL-0817250 to Robert A. Jacobs. Wayne D. Gray's work was also supported by the Alexander von Humboldt Stiftung and the Max Planck Institute for Human Development. We would like to thank Lael Schooler for providing comments that greatly improved the manuscript.

Correspondence concerning this article should be addressed to Chris R. Sims, Center for Visual Sciences, Meliora Hall, University of Rochester, Rochester, NY 14627. E-mail: csims@cvs.rochester.edu

Although intuitively appealing, the notion that behavior is ultimately governed by utility maximization is not without its critics. Over the past several decades, numerous instances have been documented where human behavior seemingly deviates from the predictions of rational choice theory (for a review, see Shafir & LeBoeuf, 2002). Not content to build a psychological theory of human choice on what was viewed as an unsound foundation, Richard Herrnstein and colleagues (Herrnstein & Vaughan, 1980) instead advocated an *empirical* basis for describing and predicting choice behavior. This alternative, known as melioration theory, asserts that human (and nonhuman animal) choice is governed by a myopic tendency towards alternatives with higher local rates of reward (Herrnstein, 1982; Herrnstein & Prelec, 1991; Herrnstein & Vaughan, 1980; Vaughan, 1981). Critically, melioration deviates from rational choice by ignoring the consequences of actions on future utility. In basing choice on the consideration of local rates of reward rather than the global maximization of utility, melioration has been characterized as a kind of *temporal myopia* (Herrnstein, Loewenstein, Prelec, & Vaughan, 1993; Herrnstein & Prelec, 1991).

In this article, we critically reexamine human performance in a repeated choice task known as the Harvard game (Rachlin & Laibson, 1997) that directly pits the predictions of melioration theory against rational choice accounts of behavior. Previous empirical results from this paradigm have widely been taken to indicate either generic irrationality (Herrnstein, 1991) or fundamental impulsivity (J. R. Gray, 1999; Kudadjie-Gyamfi & Rachlin, 2002; Otto, Markman, & Love, 2012; Tunney & Shanks, 2002; Warry, Remington, & Sonuga-Barke, 1999) in human choice, consistent with melioration theory. A key feature of this paradigm, common to many complex decision environments, is that actions have both immediate and delayed consequences, and selecting one alternative has nontrivial effects on the utility of competing alternatives on future choices. However, an additional feature of this paradigm is that participants must learn the consequences of their actions through experience.

As we demonstrate, a critical factor in judging the rationality of any behavior lies in one's understanding of the relationship between actions and their consequences. The economist Frank Knight (1921) distinguished between the concepts of *risk* and *uncertainty*. Risk is involved whenever outcomes are not guaranteed but the relevant factors and possible outcomes are known and can be quantified. By contrast, uncertainty implies that not all of the possible consequences of an action are known, or even knowable, before a decision is made. The optimal decision strategy under risk may be very different from the optimal decision strategy under uncertainty (cf. Gigerenzer, Hertwig, & Pachur, 2011). As we argue here, previous examinations of melioration theory have failed to differentiate these two concepts and, as a result, have drawn unfounded conclusions about the rationality of human choice behavior.

From the perspective of an experimenter who has accurate and complete knowledge of the dependencies inherent in the task environment, human behavior in the Harvard game exemplifies suboptimal choice under risk. However, for an experimental participant, the relationship between actions and consequences is uncertain and must be learned from experience. In the face of this uncertainty, it is not immediately evident that human behavior is suboptimal.

Historically, melioration as a theory of behavior has relied on two key assumptions: (a) The empirical occurrence of meliorating behavior is evidence for generic irrationality in human decision making, and (b) the origin of this behavior is a myopic tendency to favor alternatives with higher immediate rates of reinforcement at the expense of overall utility. In this article, we unmask both of these assumptions to show that, quite to the contrary, melioration frequently represents optimal choice behavior, even when optimality is defined as global, rather than local, utility maximization. In other words, what appears to be irrational choice under risk is in fact rational choice under uncertainty.

Melioration and Maximization

Melioration was originally proposed as the behavioral mechanism to explain the matching law (Herrnstein, 1961, 1979), which states that, at equilibrium, the relative proportion of responses made to an alternative will equal the proportion of rewards received from that alternative. The matching law, or its subsequent extension known as the generalized matching law, has been successful in accounting for human and nonhuman animal choice behavior across hundreds of experiments (for a review, see Davison & McCarthy, 1988). Whereas the matching law characterizes an aggregate property of behavior, melioration purports to describe the local, dynamic process that results in matching. Given the widespread occurrence of approximate matching behavior across species, matching has been hypothesized to reflect an innate decision strategy (Gallistel, 2005), and Herrnstein (1991) offered the speculative possibility that "people are, in fact, always following the principle of melioration and that, when they are being rational, they are being rational only incidentally" (p. 364).

Melioration theory has been offered as an explanation for phenomena as diverse as impulsivity and self-control (Herrnstein, 1981), delayed reinforcement (Chung & Herrnstein, 1967), and natural selection (Dawkins, 1999; Vaughan & Herrnstein, 1987). In the area of skill acquisition and training, Yechiam, Erev, Yehene, and Gopher (2003) argued that melioration may lead novice typists to abandon their touch-typing training, as methods such as visually guided typing have higher immediate payoff, despite lower global utility. Similarly, W. D. Gray and Fu (2004) found that small differences in the effort required to access information can lead to behavior that is locally efficient but globally suboptimal. In the clinical domain, melioration has been proposed as a tool for understanding the cognitive aspects of delinquency (J. Q. Wilson & Herrnstein, 1985) and addiction (Herrnstein & Prelec, 1992; Heyman, 1996; Heyman & Dunn, 2002). Thus, the impact of melioration theory has been broad, and its implications are of practical importance across a wide range of domains.

In most simple decision environments without delayed or indirect consequences, melioration theory predicts behavior that is similar to or indistinguishable from global utility maximization. To experimentally distinguish between global maximization and melioration, Herrnstein and others have devised decision environments in which both mechanisms are pitted directly against each other. Figure 1 illustrates the mechanics of one such environment, known as the Harvard game (Rachlin & Laibson, 1997). In this environment, the participant must make a series of choices between two alternatives, which we refer to as the *maximization* and *melioration* alternatives (the alternatives are not labeled as such for

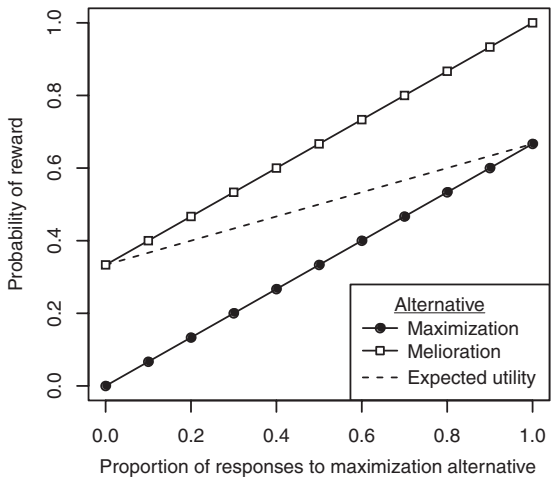

Figure 1. Reward contingencies of a decision environment designed to discriminate between global maximization and melioration. The participant faces a repeated choice between two alternatives. Irrespective of the choice history, the local probability of reward for choosing the meliorating alternative always exceeds that for choosing the maximizing alternative. For both alternatives, the probability of reward varies as a function of the proportion of choices in the recent history made to the maximizing alternative. The dashed line indicates the overall average reward associated with adopting any mixture of maximization and melioration choices. The highest point along this line yields the globally optimal strategy. Thus, the maximal payoff is achieved by *always* choosing the locally inferior alternative.

the participants in the experiment). After each choice, the participant may receive a small monetary reward. The participant's task is simply to collect as much money as possible over a series of choices in this environment.

The *y*-axis in Figure 1 plots the probability of receiving a monetary reward for choosing either of the two alternatives. The *x*-axis plots the proportion of choices among the previous 10 that were allocated to the maximization alternative. Importantly, the probability of receiving a reward for either alternative depends on this history of choices. If the participant has chosen the maximization alternative on the last 10 choices, then another maximization response will have a 66% chance of reward, but a melioration response leads to a reward with 100% certainty. In fact, the melioration response *always* yields a higher immediate probability of reward. But the more this alternative is chosen (moving to the left along the abscissa), the worse *both* alternatives become on future choices. In other words, the highest global reward is achieved when the participant always chooses the alternative that has the lower immediate probability of reward.

If the goal is to maximize total monetary winnings, the optimal strategy in this environment is to select the maximizing alternative on every trial (ignoring an end-of-game effect where the optimal strategy shifts to melioration for the last few trials). By contrast, melioration theory predicts that people will choose the alternative with the highest immediate reward probability but lowest global utility. When human participants have been tested in this environment (Gureckis & Love, 2009b; Herrnstein, 1991; Neth, Sims, & Gray, 2006; Tunney & Shanks, 2002), the modal finding is that behavior settles into a stable suboptimal pattern approximately conforming to the predictions of melioration theory. Indeed, humans appear to suffer from the same apparent irrationality as pigeons performing the same task (Herrnstein, 1982; Herrnstein & Vaughan, 1980; Vaughan, 1981).

Given the provocative result that humans are apparently no more rational than pigeons, a large number of empirical and theoretical studies have attempted to either disprove or further elucidate the role that melioration plays in human choice. One approach has been to explore a richer space of utility functions to account for the observed behavior (Staddon, 1992) by assuming that subjects seek to maximize some quantity other than total accumulated reward. However, assuming rational maximization of utility functions that severely discount or neglect the long-term consequences of actions only blurs the distinction between rational choice and melioration theory. Experimentally, it has been demonstrated that individuals recovering from drug addiction were more likely to favor the meliorating alternative in the Harvard game compared to control subjects (Heyman & Dunn, 2002). Similarly, individuals who rated higher on a psychological assessment of impulsivity were found to be more likely to meliorate (Otto et al., 2012). Numerous other studies have also concluded that melioration in the Harvard game results from impulsivity in choice or failures of self-control (J. R. Gray, 1999; Kudadjie-Gyamfi & Rachlin, 1996, 2002; Warry et al., 1999).

While the empirical results described so far generally support the theory of melioration, here we reinterpret these results by emphasizing another aspect of the Harvard game that has received little attention, namely, that people playing this game must *learn* the contingencies between actions and their consequences. It is clear that learning this relationship requires some amount of cognitive effort, and it may be the case that meliorating individuals attempt to maximize global utility but fail due to imperfect memory, lack of attention or understanding, or other cognitive limitations. In support of this view, it has been found that adding perceptual cues to disambiguate the state of the task environment or consequences of actions can facilitate global maximization (Gureckis & Love, 2009b; Herrnstein et al., 1993; Otto, Gureckis, Markman, & Love, 2009; Stillwell & Tunney, 2009). Worthy, Otto, and Maddox (2012) recently demonstrated that adding a concurrent working memory load leads to increased preference for local versus global rewards, again implicating a cognitive constraint on performance.

Similarly, if the primary impediment to globally optimal performance is learning, then providing subjects with more information about the task environment (in terms of additional instructions or hints about the reward contingencies) should increase the likelihood of maximizing behavior. While several studies have manipulated the amount of explicit information given to participants regarding the reward contingencies (Herrnstein et al., 1993; Kudadjie-Gyamfi & Rachlin, 1996; Warry et al., 1999), in no case has the occurrence of meliorating behavior been fully extinguished. As one study concluded, "even under conditions when all the factors favored global choice, participants did not respond optimally in terms of the global contingency. Perhaps this indicates a fundamental tendency to choose immediate gratification, at least occasionally, despite 'knowing better'" (Warry et al., 1999, p. 71). In summary, despite considerable experimental and theoretical effort, a basic question remains yet unanswered: Is human behav-

ior in sequential choice tasks fundamentally at odds with the normative ideals of rational choice theory?

On the Rationality of Melioration

The goal of the present article is straightforward: Whereas all previous experiments on melioration theory have examined human choice behavior and found varying degrees of suboptimality as assessed from the perspective of the experimenter, none have addressed the question of whether experimental participants should rationally have been *expected to meliorate,* even assuming the goal of maximizing global utility. This missing piece of information is of critical importance, since the long-standing claim that meliorating behavior is evidence for generic irrationality requires that a rational agent would not also meliorate. To date, this analysis is lacking.

Our approach is therefore to examine what an optimal learner should rationally believe about the relationship between choices and future rewards in a melioration experiment. Since the problem facing human participants is essentially one of uncertainty regarding the structure of the environment, our analysis takes the form of a Bayesian learning model, which we label the *rational learner model.* This approach builds on the rational analysis framework (Anderson, 1990), which states that a greater understanding of human cognition can be gained by examining the structure of the external environment as well as the goals of the cognitive system and proposes that cognition is intricately adapted to this structure in achieving its goals. In the present application, it is assumed that a primary goal of the cognitive system in an uncertain choice environment is discovering how actions relate to their consequences. The structure of the environment is such that each decision is unique, as the same environmental context is never exactly repeated. This lack of stable context is a serious impediment to generalizing past experience to future decisions. However, by abstracting across irrelevant features, it is possible for the cognitive system to gain an understanding of a task environment in terms of its functionally equivalent states (Redish, Jensen, Johnson, & Kurth-Nelson, 2007; R. C. Wilson & Niv, 2011). A key piece of the learning challenge for our model is therefore to discover this latent structure in the environment.

Previous theoretical results have demonstrated that adopting an incorrect representation of the task environment will lead to meliorating behavior under a fairly broad class of learning and decision rules (Sakai & Fukai, 2008). However, to date, no one has examined whether a rational decision maker could, in principle, *learn* an appropriate representation of the task environment in a melioration experiment. Without this key piece of information, it is not clear whether documented instances of melioration reflect irrationality in human decision making under risk or whether they point to a rational agent acting optimally in the face of significant environmental uncertainty. To address this limitation, our rational learner model was designed with the explicit goal of inferring the structure of the task environment, without placing strong restrictions on what structures are possible a priori. Thus, the model is free to entertain incorrect hypotheses of the task environment, just as human participants in a melioration experiment are free to do. One limitation of the present analysis is that it is primarily concerned with whether people's behavior is rational given what they have observed. We do not explicitly address the question of whether people optimally explore a decision environment (Steyvers, Lee, & Wagenmakers, 2009). Importantly, the rational learner model is also not intended as a process-level model of human learning (in fact, it is rather implausible in this regard). Rather, it is intended as a bound on what could potentially be learned by human participants, assuming the absence of cognitive limitations. If a fully rational learning agent meliorates, then the claim that meliorating behavior indicates irrationality must be discarded.

In the next section, we briefly describe the results from an experiment using the Harvard game in which humans predominantly fail to learn the globally maximizing decision strategy. After presenting the basic results, we derive our rational learner model and apply it to the empirical data from each individual. To preview our results, the analysis uncovers some rather surprising facts about our data. These results include the finding that even an unbiased rational learner could be led to believe that melioration will be of higher long-term value than the supposedly optimal strategy, despite extensive experience in the task environment. Based on these results, we conclude that melioration can be interpreted not as irrational behavior under risk but instead as rational choice under uncertainty.

Experiment

Method

Participants. Twelve undergraduate students from Rensselaer Polytechnic Institute (Troy, New York) participated in the experiment in exchange for monetary compensation. The amount of money earned by each participant depended on his or her performance in the task.

Apparatus. On each trial of the experiment, participants faced a choice between two alternatives. The alternatives were presented as two buttons labeled *Left* and *Right* on a computer screen. After each decision, the task interface indicated whether or not the choice resulted in a reward, as well as the participant's cumulative winnings for the experiment. Rewards were given probabilistically such that the reward on any trial was either 0 or 2 cents. The probability of receiving a reward depended on the participant's choice on the most recent trial, as well as his or her previous 9 choices in the experiment, as illustrated in Figure 1. In general, the probability of receiving a reward for one of the alternatives (the *melioration* alternative) was always higher than the other (the *maximization* alternative). Specifically, the probability of receiving a reward following a choice of the maximization alternative was given by $p(\text{reward}|\text{max}) = (2/3) \times (n_{\text{max}}/10)$, while the probability of reward for the melioration alternative was $p(\text{reward}|\text{mel}) = 1/3 + (2/3) \times (n_{\text{max}}/10)$, where n_{max} indicates the number of choices made to the maximization alternative during the previous 10 choices. Since the choice history is ill-defined for the first 10 choices of the experiment, we initialized each participant's history with a sequence of 10 alternating meliorating and maximizing choices.

Over the course of the entire experiment, consistently choosing the maximization alternative would earn the participant an expected cumulative reward of $10.67, while consistently meliorating would earn an expected payoff of $5.33. The maximization and melioration alternatives were randomly mapped to the left and

right buttons for each participant. Participants indicated their response on each trial using a standard computer mouse, and there were no time constraints on their decision.

Procedure. Before beginning the experiment, each participant received instructions on the task interface. Participants were told that the amount that they won in the experiment depended on their choices and that they could win between approximately $5 and $11. Beyond this, participants were not informed about the specific dependencies between choices and outcomes but were instructed to maximize their earnings over the course of 800 experimental trials. Each participant completed all 800 trials in a single session lasting approximately 45 min. At the end of the experiment, participants were paid the amount of money gained during the study.

Results and Discussion

On average, participants chose the maximizing alternative on 46.4% of trials, and earned an average of $7.72 (minimum = $6.42, maximum = $9.92). Figure 2 shows a raster plot of the history of choices made by each participant. Black line segments indicate choices allocated to the maximizing alternative, while white indicates melioration. As can be seen, participants exhibited large individual differences in their allocation of choices. As a first characterization of participants' choice behavior, a binomial test was performed in order to determine if each participant favored the maximization or melioration alternative at a greater-than-chance level. The results indicated that seven of the 12 participants demonstrated a significant bias towards melioration, while just three participants favored the maximizing alternative (all ps $< .01$; corrected for multiple comparisons using Holm's method). To examine whether mean choice behavior changed over the course of the experiment, trials were grouped into blocks of 100, and a one-way within-subject analysis of variance was performed on the proportion of maximizing choices, using block number as a factor. The results of this analysis indicate that no significant shifts in overall choice behavior occurred across all participants, $F(7, 77) = 1.17$, $p = .34$, *ns,* though clearly at least two individuals (11 and 12) increased their allocation to the maximizing alternative. Figure 3a shows the mean proportion of maximization and mean reward rate for each participant, where each plot marker indicates a different participant. It is notable that even the participant showing the strongest degree of melioration (indicated by the left-most plot marker in Figure 3a) exhibits *undermatching,* or behavior closer to indifference as compared to a complete bias towards melioration. Undermatching has been shown to be fairly common in studies of choice behavior (Baum, 1979) and, as we demonstrate later, is also predicted by a rational learner model for this task.

In summary, the majority of participants failed to exhibit the globally optimal choice strategy. Indeed, seven of the 12 participants favored the melioration alternative at above-chance levels, and the overall proportion of maximization did not increase across blocks of the experiment. This finding echoes results in other studies that have used a similar or identical paradigm (Gureckis & Love, 2009a; Tunney & Shanks, 2002).

A Rational Learner Model

In this section, we derive a *rational learner model* for the inference problem facing individuals in uncertain sequential decision environments such as the Harvard game. The goal for our analysis is to determine, for a given person's history of choices and rewards, what he or she should rationally believe about the task environment with regard to the relationship between past choices and future rewards. This analysis allows us to examine whether an individual's decisions are rational on the basis of the evidence available to the participant, rather than judging optimality from the perspective of the experimenter. If all the evidence available to a participant suggests that melioration is a better long-term alternative, then the claim that consistent melioration consti-

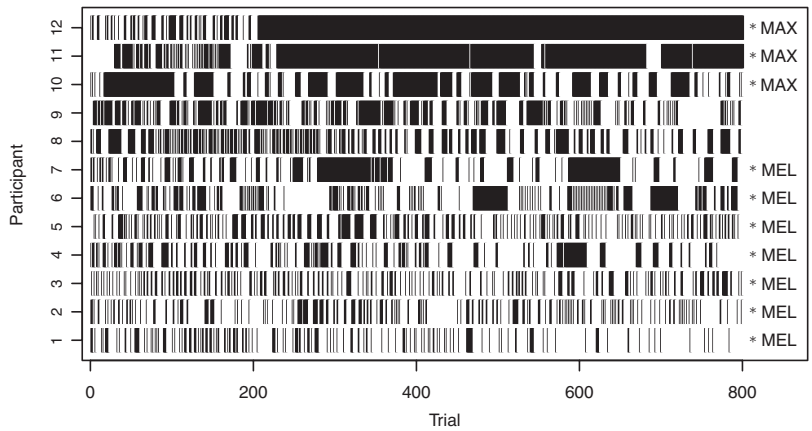

Figure 2. Raster plot of choices made by each participant in the experiment. Each row of the plot indicates the entire choice history for a single participant. Black line segments indicate choices allocated to the maximizing alternative, while white indicates melioration. The participants' data are ordered from bottom to top in ascending order of total number of maximizing choices made during the experiment. An asterisk to the right of a participant's data indicates that his or her proportion of maximizing choices differed significantly from chance at the $p < .01$ level, with the direction of bias indicated as MEL (bias towards melioration) or MAX (bias towards maximization).

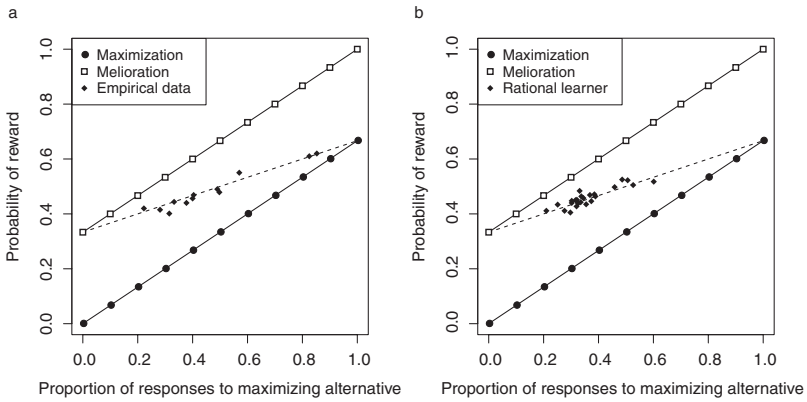

Figure 3. a: Average proportion maximization and average reward rate for each participant, overlaid over the reward contingencies. b: Performance of 24 simulated rational learners.

tutes evidence of generic irrationality in human decision making becomes untenable.

The learning challenge facing participants in the experiment is discovering the unknown function that relates past choices to future rewards. The process of learning this function further involves three subproblems that must each be overcome by the learner. First, participants are not told that reward probabilities depend on the past 10 trials but must somehow infer this *history window* from experience. Second, even assuming knowledge of the relevant history window, there are a large number of different possible choice histories of length 10 but only a small number of functionally distinct *states* in the environment. In particular, all choice sequences that have five maximizing choices in the most recent 10 are functionally equivalent. For example, the two choice sequences of length 10—{XXXXXLLLLL} and {XLXLXLXLXL}, where X = maximization, L = melioration—would both lead to a situation where the probabilities of reward on the very next choice are identical. Thus, these two very different choice sequences correspond to just a single underlying state of the task environment, and a rational learner should somehow infer this from its limited experience. As we show, the process of learning which choice histories map onto which states of the environment is essentially a problem of categorization, and the implementation of our learning model closely mirrors existing rational models of categorization (Anderson, 1990; Sanborn, Griffiths, & Navarro, 2010). Finally, even if the participant knows how different choice histories map onto different states of the task environment, he or she must still learn the reward probabilities for each state (i.e., the numerical values plotted in Figure 1). Only by overcoming all three of these learning challenges will a decision maker be able to accurately predict the consequences of each choice in the environment.

Since the learning problem is characterized by uncertainty regarding the environment, the optimal solution takes the form of a Bayesian learning model. We formalize the learning problem as one of inferring a posterior distribution over three quantities: *w,* the relevant history window for the task environment; *f,* the function that maps each choice history of length *w* onto one of a discrete number of states; and θ, which indicates the probability of obtaining a reward for pressing either of the alternatives in each possible state of the environment. The posterior distribution over these three quantities can be written using an application of Bayes' rule:

$$p(\theta,\ f,\ w|X) = \frac{p(X|\theta,\ f,\ w) \times p(\theta|f,\ w) \times p(f|w) \times p(w)}{p(X)}. \tag{1}$$

The denominator is a normalizing constant to ensure that probabilities sum to 1 and can be ignored in the present case. The first term in the numerator of Equation 1 indicates the likelihood function for the observed data *X,* given knowledge of the three unknown quantities. Given a sequence of binary choices and binary outcomes, as well as knowledge of the reward probabilities and underlying states of the environment, the likelihood of the observed data is defined by

$$p(X|\theta,\ f,\ w) = \prod_a \prod_s \left[(\theta_{a,s})^{NS_{a,s}} (1 - \theta_{a,s})^{NF_{a,s}}\right], \tag{2}$$

where $\theta_{a,s}$ indicates the reward probability associated with choosing alternative *a* (MEL or MAX) in state *s.* The observed data *X* can be compactly summarized by the count of the number of successes (NS) and number of failures (NF) for each alternative in each state, where successes and failures indicate rewarded and non-rewarded outcomes.

The remaining three terms in the numerator of Equation 1 define the prior belief distribution over the reward probabilities, the prior over functions assigning choice histories to states, and the prior over possible history windows. For the reward probabilities, $p(\theta|f, w)$, the model assumes a uniform prior distribution in the range (0, 1) for each alternative in each state. Similarly, the prior probability over history windows is assigned a uniform prior over the range of 0 . . . 10 past choices. A history window of 0 trials is equivalent to assuming that outcomes do not depend at all on past choices, while a history window of 10 trials captures the true structure of the environment. Adopting a uniform prior over this range represents the minimal set of possible history windows that a learner would need to consider to accurately learn the task. The only remaining term to specify is the prior distribution over functions *f* that map choice histories onto functionally distinct states of the environment.

As mentioned above, the challenge of inferring the state structure of the environment bears striking similarity to a categorization problem: On each trial, the rational learner will be acting from the starting point of a particular choice history and must decide if this history is equivalent to an already experienced state or if the present situation is functionally distinct from all past experiences. However, rather than categorizing objects by their perceptual features, the present approach categorizes choice sequences by their expected utility in the task environment. Another way of looking at this problem is deciding whether the two choice histories {XXLL} and {LLXX} should be assigned to the same category (assuming a relevant history window of four previous choices). If the true environmental structure is such that reward probability depends only on the count of maximizing choices, then the two sequences are equivalent and should be assigned to the same state, despite the fact that each individual choice differs in the two sequences. However, a priori, it is just as likely that reward probability depends not on the count of maximizing choices but rather on some other property of the choice sequence. In the absence of prior knowledge of the task environment, a rational learner should consider all possible functions that map choice sequences onto functionally equivalent states. This is captured by defining a prior distribution over the space of possible functions *f*, where the input to this function is a choice sequence and the output is the state to which that choice sequence belongs.

One possibility for the learner is to treat each history as a unique state of the task environment. However, under this assumption, the learner is unlikely to experience the same state twice and thus cannot generalize his or her past experience to the present situation. Therefore, the prior over functions implemented in the rational learner model instead assumes that each newly encountered choice history is likely to be functionally equivalent to a previously encountered state. However, the model still maintains the possibility that the choice history should instead be assigned to a separate state, so that the number of inferred states can grow over time as the learner acquires more evidence and gains finer distinctions between choice histories. In implementation, this prior over functions is defined by what is known as a Dirichlet process (Ferguson, 1973; Neal, 2000), which has also been used to develop rational models of human categorization (Anderson, 1990; Sanborn et al., 2010) and word segmentation (Goldwater, Griffiths, & Johnson, 2009). Formally, the prior probability of a particular function *f* under the Dirichlet process model is given by

$$p(f|w) = \frac{\Gamma(1+\alpha)}{\Gamma(n+\alpha)} \times \alpha^{S(f)-1} \times \prod_{s=1}^{S(f)} \left(n_s^{(f)} - 1\right)!, \qquad (3)$$

where $S^{(f)}$ indicates the total number of distinct states assumed by the particular function *f*, $n = 2^w$ is the total number of choice histories, and $n_s^{(f)}$ indicates the number of different choice histories that are assigned to state *s* under the given function. The Dirichlet process introduces one parameter, $\alpha \geq 0$, which in essence determines the strength of generalization in the model. Large values of α favor assigning each choice history to a unique state, whereas small values favor assigning many choice histories to the same state, such that experience acquired from previous choice histories will generalize to each newly encountered history. In the present implementation of the model, this parameter was set to $\alpha = 1$ as this value favors a moderate degree of generalization appropriate for the true reward function.

Note that the choice of a Dirichlet process prior over the space of possible reward functions defines one possible rational learner model, but other choices are possible. Alternative choices might place stronger or weaker prior probability on different functions and thus lead to different predictions for rational behavior. Clearly, priors that assign high probability to the true reward function would lead to improved learning performance. While we believe the present approach to be both general and principled, additional research would be necessary to fully characterize the consequences of this assumption.

We can apply the rational learner model defined by Equations 1–3 and a given participant's actual sequence of choices and rewards to infer what could rationally be known about the task environment on the basis of a limited set of observations. While computing the full posterior distribution given by Equation 1 is intractable, recent Monte Carlo sampling techniques enable Bayesian inference for this model (Gilks, Richardson, & Spiegelhalter, 1998; Neal, 2000). Details regarding the inference procedure are provided in the Appendix. In the next section, we analyze human performance according to the rational learner model.

Model Results

For each trial of the experiment, the rational learner model was provided with the actual sequence of choices and outcomes for an experimental participant and then used to infer what each participant could rationally believe about the task environment given his or her particular task experience. As a starting point, we first examine whether participants had enough information to infer the correct history window (the number of previous choices that determine the probability of reward).

Figure 4 shows the posterior probability for different history windows in the range of 0 . . . 10 previous choices, estimated according to the rational learner model. The results shown are averaged across all 12 participants. The width of each band in Figure 4 reflects the posterior probability for each history window. On the first trial of the experiment, all history windows have equal prior probability, and so,

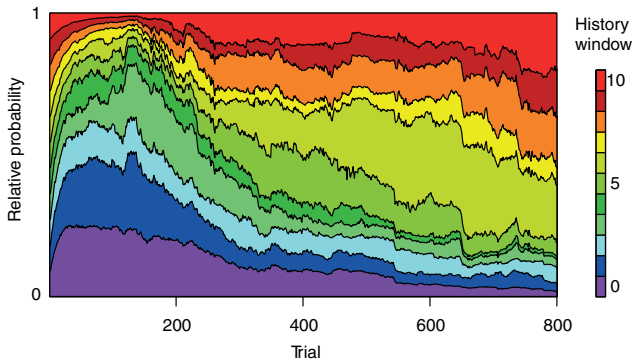

Figure 4. Posterior probability of different history windows, averaged across participants. Each band indicates the average posterior probability of a different history window (ranging from 0 through 10 previous trials). The width of the band indicates relative probability (total probability sums to 1). The true history window for the experiment is 10 previous trials (band at top of plot).

each band is of equal width. As participants accrue evidence in the experiment by choosing actions and observing outcomes, the relative probability of different history windows changes. With limited experience in the task (<200 trials) there is little evidence to support a complex dependency between rewards and past choices, and so, the simplest hypothesis, $w = 0$, has the highest posterior probability. As participants acquire more evidence, there is more data to support complex models of the task environment. However, even by the end of the experiment, there is insufficient evidence to conclude with certainty that rewards depend on the history of the previous 10 choices.

To explore in further detail what each individual might rationally believe about the task, we used the rational learner model to infer the expected value associated with adopting a range of 11 different behavioral strategies in the experiment. A given strategy chose the maximization alternative on each trial independently with probability $= i/10$, where i could range from 0 through 10. This defines a range of strategies that includes pure maximization ($i = 10$, or choosing maximization with probability 1) through pure melioration ($i = 0$). The predicted value of each strategy was determined as its expected reward rate (average probability of reward per trial) over a sequence of 100 choices. The model was also used to assess the predicted reward rate for the participant's subsequent 100 choices in the task (i.e., predicting the value of the empirical choices before they are taken). When fewer than 100 trials remained in the experiment, the predicted values were extrapolated to the end of the experiment.

Figure 5 shows the inferred value functions for each participant after completing half of the experiment (400 trials). Each panel shows the predicted long-term reward rate (defined as the average

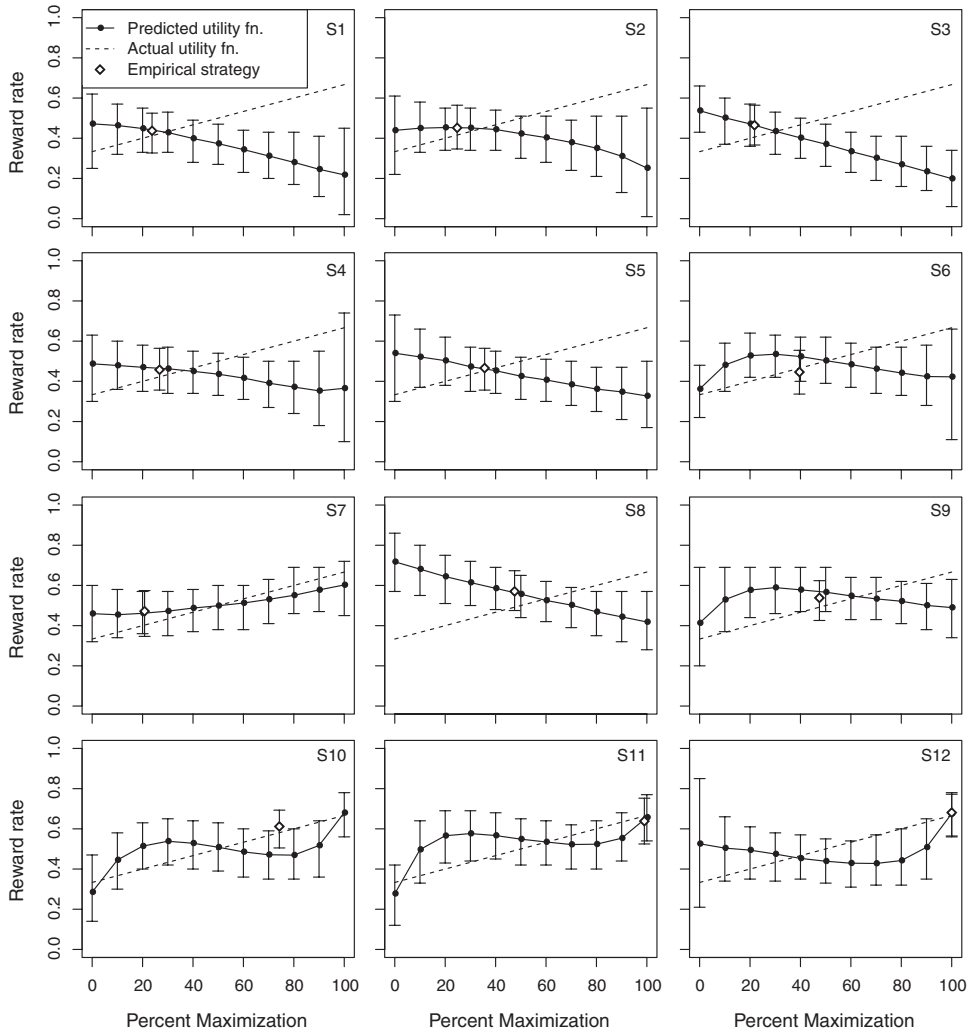

Figure 5. Inferred value functions for each participant after 400 trials of experience in the task environment. Each panel shows the predicted reward rate associated with different behavioral policies (varying across the *x*-axis). The dashed line indicates the true utility function. The inferred value of the empirical policy for each subject is indicated by a diamond-shaped marker. Error bars correspond to the 95% highest density credible interval around the expected value. The error bars stem from the rational learner model's uncertainty regarding the true reward structure of the environment. (fn. = function; S = subject.)

number of rewards per trial over a span of 100 trials) for a single participant across the range of behavioral strategies. The dashed line indicates the true value function for the task, while the solid line indicates the subjective value inferred by a rational learner. The predicted value of the empirical policy is indicated by a diamond-shaped marker. What is immediately apparent from Figure 5 is that for many participants, the value function inferred by the rational learner is markedly different from the true value function. While in reality maximization has the highest global reward rate, for approximately half of participants, the task environment has provided misleading evidence that maximization has the *lowest* global value, while a strategy of pure melioration would be expected to maximize global winnings.

Participants 1 and 12 reflect the experimental participants with the least and most maximizing choices in the experiment, respectively. According to the standard interpretation of melioration theory, Participant 12 exhibited rational behavior, while Participant 1 was fundamentally irrational, due to a myopic tendency to focus on immediate rewards. The results in Figure 5 prove this interpretation to be misguided. For both participants, the empirical sequence of choices actually adopted in the task would be predicted by a rational learner to be near optimal in terms of maximizing global utility. In other words, both participants were rational in their choices given their experience.

While Figure 5 considers human performance at a single point in the experiment (after completing 400 trials), it is also possible to examine the rationality of empirical behavior over the course of the entire experiment. The line with square markers in Figure 6a plots the average reward rate predicted for the choices made by participants. This line represents what a rational learner would predict for the outcomes of participants' subsequent 100 choices, given their observed sequence of choices and outcomes up to a given trial. For comparison, the predicted values of pure melioration and maximization strategies are also shown. With limited experience in the task ($<$200 trials), a strategy of exclusive melioration would be expected to have the highest overall utility. Over time, as participants acquire more evidence in the experiment, the predicted value of melioration decreases, while the value of maximization increases. Importantly, however, the actual sequence of choices made by participants would be predicted (by a rational learner) to be superior to pure maximization. Thus, the environment provided insufficient evidence for participants to recognize that their behavior was suboptimal relative to the supposedly rational maximization strategy.

If participants allocated their choices in a rational manner, then those who observed more evidence that maximization had higher value should also exhibit more maximizing behavior in their subsequent choices. To assess whether this in fact was the case, the rational learner model was used to compute the relative value of maximization (defined as the predicted value of maximization minus the predicted value of melioration) for each participant on each trial of the experiment. The relative value of maximization was then correlated with the percent of maximizing choices observed for each participant over the subsequent 100 trials. If participants are in fact sensitive to their experience in a manner approximating the rational learner model, then a positive correlation would be expected. Figure 6b shows the results of this analysis. For most of the experiment, the correlation remains between 0.5 and 0.9, indicating that the rational learner model offers a fairly accurate prediction of participants' subsequent choices. That is to say, participants who had more evidence to rationally favor the maximization option did in fact maximize more strongly.

Strategic Exploration of the Task Environment

The results presented so far were obtained by using the rational learner model to infer what each participant should believe about the task environment given his or her own past sequence of choices and observed outcomes. However, it may be misleading to claim that behavior is rational given what participants observed because participants largely determined what they observed through their own sequence of choices. This feature is true of any decision environment, and so, optimal behavior requires that participants must simultaneously balance exploration of the task environment with choices that exploit the knowledge gained. While we have shown that participants' choices were largely optimal given what they observed, it may be the case that they were suboptimal in their exploration of the environment.

To break any potential circularity between past observations and predicted value, we performed an additional simulation in which the choices made by the rational learner model were not tied to the choices made by experimental participants. Instead, the rational learner model was allowed to generate its own experience by

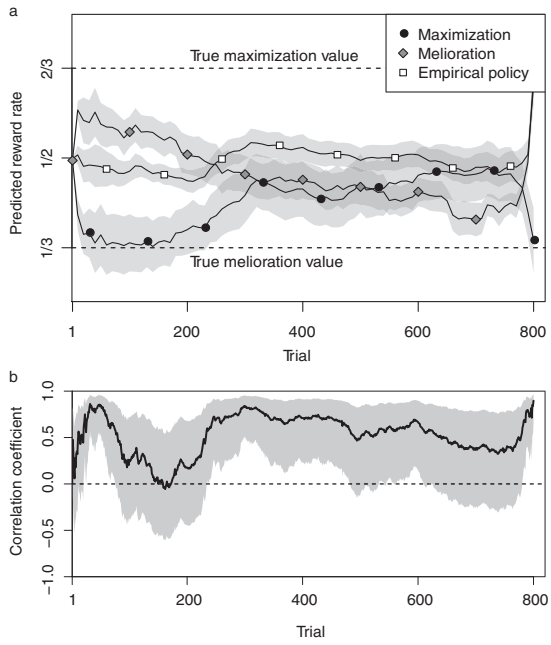

Figure 6. a: Predicted value of maximization, melioration, and the empirical sequence of choices made by participants, estimated according to the rational learner model. Shaded region indicates ± 1 standard error around the mean. Predicted value is computed on every trial; the plot markers are added as a visual aid to help distinguish between the different lines. b: Correlation between empirical percent maximization and the relative value of maximizing versus meliorating, as inferred by the rational learner for each participant on each trial of the experiment. The solid line plots the Pearson correlation coefficient, the shaded region indicates the 95% confidence interval.

selecting the maximization or melioration alternative on each trial (in essence, the model was treated as a participant and run through the experiment in the same manner as human participants).

On each trial, the rational learner model based its decision on the inferred value function given its past experience. Specifically, the model computed the predicted value of a range of different choice strategies. Each such strategy chose the maximization response with probability (i/10), where i ranged from 0 through 10. This defines a space of strategies that includes pure maximization ($i = 10$) and pure melioration ($i = 0$), as well as a number of intermediate behavioral allocations. Since optimal exploration of the present decision environment represents a computationally infeasible problem, the model selected between these strategies using a heuristic approach known the softmax decision rule. This same decision rule has been used in numerous existing models of human choice where individuals must balance exploration and exploitation in uncertain environments (e.g., Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Fu & Anderson, 2006; Gureckis & Love, 2009a; Yechiam & Busemeyer, 2005). Accordingly, the probability of selecting a behavioral allocation strategy a_i that maximizes on [100 $\times$ (i/10)]% of its choices is given by

$$P(a_i) = \frac{e^{U(a_i)/\tau}}{\sum_{j=0}^{10} e^{U(a_j)/\tau}}. \quad (4)$$

The function $U(a_i)$ indicates the inferred reward rate for strategy a_i, and τ is a noise parameter governing the trade-off between exploration and exploitation. If on a given trial the selected strategy was $a_i = 90\%$ maximization, the model chose the maximization alternative with probability 0.90 and chose the melioration alternative with probability 0.10. Figure 3b shows the results of running 24 simulated participants through the experiment, using a noise parameter $\tau = 0.01$. The behavior of the rational learner model closely resembles human participants: Both settle on a behavioral allocation that favors melioration over maximization, though the model appears to exhibit less between-subject variability than humans. It is noteworthy that both the human participants and the model demonstrate undermatching (Baum, 1979), or a behavioral allocation closer to indifference than predicted by the matching law. In the case of the rational learner model, this is largely controlled by the noise parameter in the softmax decision rule: Increasing noise tends to predict behavior closer to indifference, while decreasing noise increases the strength of meliorating behavior. Critically, the model accounts for the existence of apparent meliorating behavior by attempting to maximize global, rather than immediate, reinforcement.

One possible criticism of this result is that both human subjects and the rational learner model infrequently selected the maximizing alternative. Given this fact, one might argue that it is not surprising that they did not discover the true value of the maximization strategy. First, this argument points to a fundamental challenge inherent in any complex decision environment—that exploration requires selecting suboptimal actions according to current beliefs—but does not point to irrationality on the part of either human participants or the rational learner model. Second, and less obvious, is the fact that even if participants exhibited a strong bias towards maximization, it is unlikely that they would have acquired an accurate assessment of the relative value of maximizing and meliorating within the span of 800 trials. To demonstrate this counterintuitive fact, we conducted simulations in which the rational learner model selected between the two alternatives, but with a biased choice allocation, so that the maximization alternative was chosen on either 25% of trials (bias towards melioration) or 75% of trials (bias towards maximization) regardless of the inferred value for each alternative. For comparison, an unbiased choice allocation (50% independent probability of maximization on each trial) was also evaluated. Furthermore, we extended the simulation to one million trials to assess the time course of learning.

Figure 7 shows the results of this analysis. As expected, when choice allocation is biased towards melioration (see Figure 7a), the predicted value of maximization is well below that of melioration, and the same holds true even when choice allocation is unbiased (see Figure 7b). More surprisingly, for an unbiased allocation strategy, it would take a rational learner more than 20,000 trials to

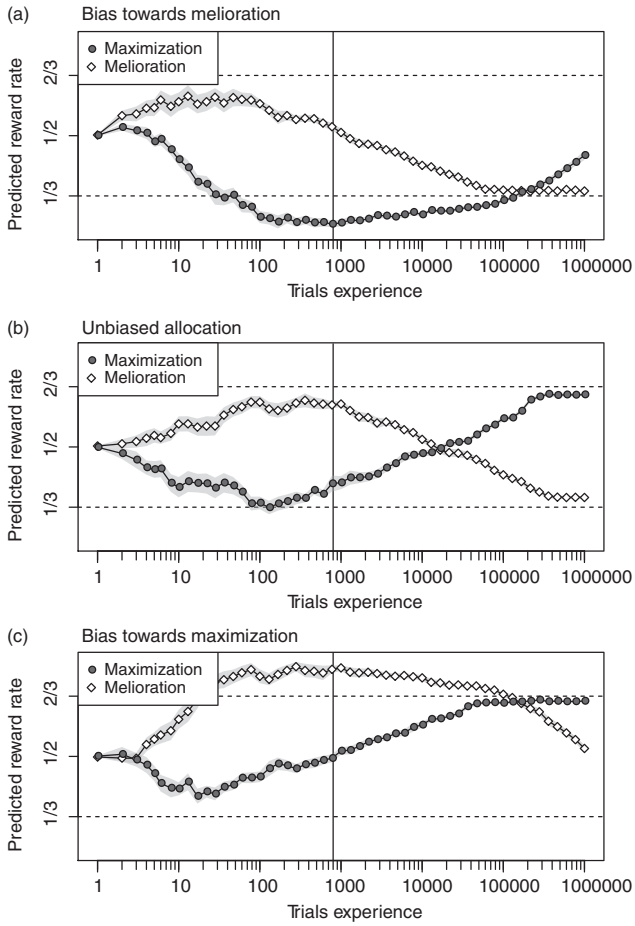

Figure 7. Inferred value of maximization and melioration for the rational learner model as a function of amount of experience in the task environment. a: Values inferred given a choice allocation biased towards melioration. b: Unbiased choice allocation. c: Values inferred given a choice allocation biased towards maximization. The shaded region indicates ± 1 standard error, computed using 24 simulated agents. The dashed lines indicate the actual reward rate for strategies of pure melioration and pure maximization. The vertical line indicates 800 trials of experience, the amount available to human participants in the experiment.

infer that maximization is preferable to melioration, an amount of experience that would require a human participant close to 24 hr in the laboratory to reach. Even when choice allocation is instead biased towards maximization (see Figure 7c), this pattern is *not* reversed. Rather, in this case, the predicted value of meliorating would be even higher. The reason is that on the (infrequent) trials when the melioration alternative is selected, the probability of reward is close to 1. In other words, the more one maximizes, the more the environment pulls behavior back towards melioration.

Evaluation of Alternative Choice Allocation Strategies

As an additional index of the difficulty of learning the environmental contingencies, we computed the mean squared error (MSE) between the actual reward function and the inferred value function (e.g., computing the MSE between the dashed and solid lines in Figure 5). We then examined how MSE in the inferred value function changed as a function of experience with the task. As the model acquires more evidence and learns the true reward structure, the MSE would be expected to decrease towards zero. This analysis also allows us to examine how different strategies for exploring the environment influenced the rate of learning. A 75% bias to maximization or melioration represents two such strategies. We also tested a choice strategy where consecutive choices were highly correlated: This strategy repeated the choice from the previous trial with probability 0.75 and switched to the other alternative with probability 0.25. Additionally, we examined performance of the win–stay, lose–shift (WSLS) heuristic (Robbins, 1952): When the previous trial was rewarded, the same choice is repeated; otherwise, the next choice is made to the other alternative. As applied to the present paradigm, adopting WSLS would predict an allocation of around 35% maximizing responses; the question is whether the resulting observations would be sufficient to indicate to participants the suboptimality of this strategy. For a baseline of comparison, we also examined a completely random (50% independent probability of maximization on each trial) allocation of choices.

The results of this analysis are shown in Figure 8a. Perhaps surprisingly, three of the choice allocation strategies would be expected to hinder understanding of the task dynamics compared to a completely random allocation of behavior (specifically, 75% bias to melioration, 75% bias to maximization, and the WSLS strategy). Only the correlated choice strategy would be expected to

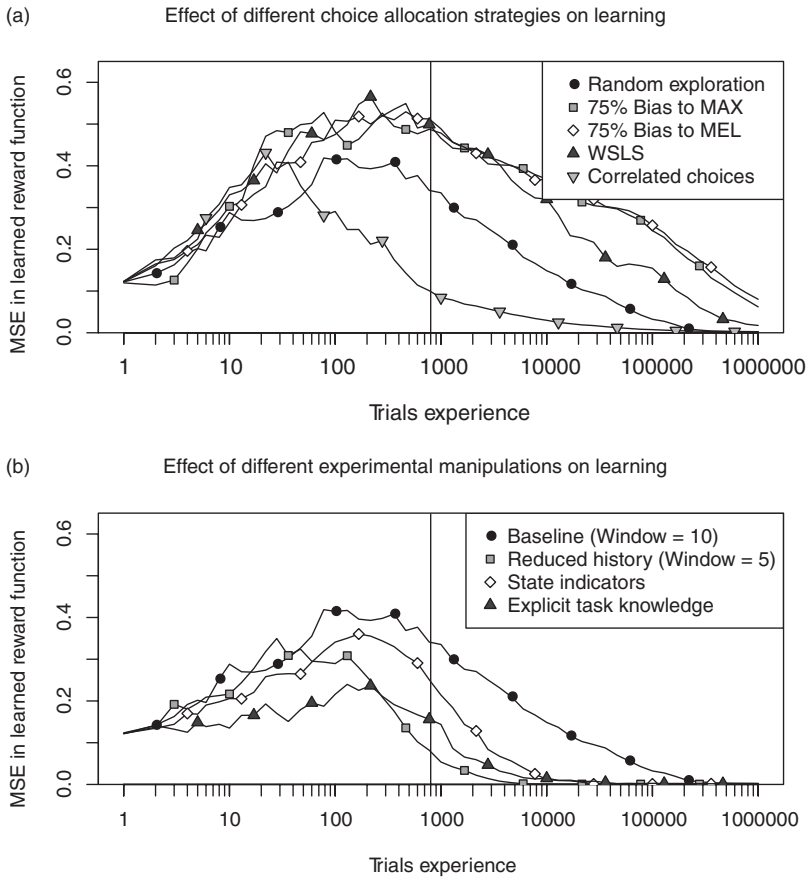

Figure 8. Mean squared error (MSE) comparing learned value functions to the true reward contingencies of the task as a function of amount of experience in the task environment. a: Evaluation of different choice allocation strategies. MAX = maximization; MEL = melioration; WSLS = win–stay, lose–shift. b: Evaluation of different manipulations to the task environment. All curves are averaged over 24 simulated agents. The vertical line indicates 800 trials of experience.

learn the true reward contingencies more quickly. In fact, this strategy would be able to learn an accurate model of the reward structure within a time frame comparable to the amount of experience the experimental participants were given in the task. However, there is no feature of the task environment that would indicate to a participant that a correlated choice strategy is likely to be successful. Thus, if an individual happened to exhibit a tendency to repeat a single choice for many trials, he or she might learn that maximization had a higher value than melioration (perhaps accounting for the two participants who exhibited stable maximization in the experiment).

Intriguingly, the finding that strong correlations in choice behavior could lead to improved learning sheds light on a recent finding from Otto et al. (2012) that participants who rated higher on a psychological assessment of impulsivity were more likely to favor the meliorating alternative. Otto et al. attributed this result to the hypothesis that participants rating higher on impulsivity were more likely to favor immediate reinforcement over higher long-term gains (a melioration theory account). However, the present results suggest an interpretation that may have nothing to do with a myopic view of rewards: If impulsive subjects are not myopic but rather are more variable in their behavior (for any reason, even if they are attempting to maximize long-term gains), they would be predicted to learn a meliorating strategy. This is due to the fact that behavior closer to random (independent) choice favors melioration more strongly than autocorrelated choices distributed to both alternatives.

Evaluation of Manipulations to the Task Environment

The results presented so far directly challenge the long-standing assumption that the empirical observation of meliorating behavior indicates any form of irrationality in human choice. However, a large number of additional experimental manipulations have been conducted within the context of the basic melioration paradigm. Experimentally, it has been found that maximization increases when rewards depend on a smaller span over previous trials (Herrnstein et al., 1993). Figure 8b (square plot markers) shows that when the history window is reduced from 10 to five previous trials, the rate of learning an accurate model of the task environment increases by two orders of magnitude. Other experimental manipulations have examined how adding cues to the underlying state of the task improves learning (Gureckis & Love, 2009b; Herrnstein et al., 1993; Otto et al., 2009), for example, by adding indicator lights that correspond to the number of maximizing choices over the relevant choice window. Consistent with the empirical finding that maximization increases in this case, the rational learner model predicts an improvement in performance when the task is modified to give labels that disambiguate the state that a particular choice history should be assigned to (see Figure 8b, diamond plot markers).

It has also been demonstrated empirically that providing participants with explicit hints regarding the dynamics of the task also improves performance (Herrnstein et al., 1993; Kudadjie-Gyamfi & Rachlin, 1996; Warry et al., 1999). We tested this manipulation by running a rational learner model that assumed a priori that rewards depended on the number of maximizing choices in the recent history but was not informed that the relevant history span was 10 previous trials. This manipulation also leads to improved learning performance (see Figure 8b, triangle plot markers), although even in this case it takes several thousand trials before the error in the learned value function approaches zero. This demonstrates two important facts: First, even when given substantial knowledge of the task, the learning challenge facing participants is far from trivial. Second, it may be the case that human participants could outperform the rational learner model presented here without invalidating the present results. In defining the rational learner, it was important that it not be biased a priori towards the correct structure of the environment. However, it is well known that simple but biased heuristics often outperform more sophisticated (but unbiased) decision strategies (Gigerenzer & Brighton, 2009; Katsikopoulos, Schooler, & Hertwig, 2010). More specifically, experiments in function learning indicate that people possess a bias favoring linear relationships with a positive slope (Kalish, Griffiths, & Lewandowsky, 2007); such a bias would be expected to favor learning a correct model of the task environment (where reward probability is in fact a linear function of number of maximizing choices). Similarly, human participants might assume a restricted feature space compared to the rational learner model. Memory limitations (e.g., Stevens, Volstorf, Schooler, & Rieskamp, 2011) might constrain people to consider only the tally of maximizing and meliorating choices and thus improve their performance in the task. A careful assessment of the impact of cognitive biases and constraints on performance is an important area for future research.

Summary and Conclusions

In this article, we have considered two basic but competing accounts of the organizing principles of human decision making: rational choice theory and melioration. According to rational choice theory, humans act in a manner that seeks to maximize the overall achievement of subjective utility. By contrast, melioration theory asserts that the driving force underlying decision making is not the attempt to maximize global utility but rather a process of continually shifting behavioral preferences towards alternatives with higher local rates of reward. The implications of the debate between melioration and rational choice theory are both important and widespread, impacting fields as diverse as training and education, criminal justice, and the treatment of substance abuse and addiction.

An important piece of evidence in this debate comes from a simple experimental paradigm, known as the Harvard game, in which participants must make a sequence of repeated choices between two alternatives. Compared to most laboratory studies of human decision making, this experiment contains two important properties that are common in real-life decisions: First, the utilities of competing alternatives are not independent, and second, the consequences of actions can be delayed in time without any obvious indication of this fact. In the past, the results using this paradigm have been interpreted as evidence that humans are either fundamentally impulsive (J. R. Gray, 1999; Kudadjie-Gyamfi & Rachlin, 2002; Otto et al., 2012; Tunney & Shanks, 2002; Warry et al., 1999) or generically irrational (Herrnstein, 1991) in terms of their allocation of behavior.

Historically, the interpretation of these results as evidence against rationality and in favor of melioration theory has relied on two key assumptions: (a) The empirical occurrence of meliorating

behavior is evidence for generic suboptimality in human decision making, and (b) the origin of this behavior is temporal myopia, or the exclusive consideration of short-term rewards at the cost of long-term optimality. The primary contribution of this article is to demonstrate that both of these assumptions are false. It is not the case that melioration necessarily reflects the behavior of an irrational individual who only considers immediate rewards. Rather, with limited experience in an uncertain environment, melioration defines the response strategy that would be predicted by a rational agent attempting to maximize global utility.

From the perspective of the experimental designer, humans may indeed appear to act suboptimally. However, for the experimental participant, the task is more complicated. At the outset of the typical melioration experiment, the participant may know that current and future rewards depend on past choices, but this does little to narrow the vast space of possible relationships between actions and outcomes. Thus, the problem for a participant in such an experiment is not merely one of choice but fundamentally also one of learning. The distinction between these two perspectives is extremely critical in terms of what one may infer about the rationality of human choice. If crucial information about the contingency between choices and outcomes needs to be learned under uncertainty, it may be premature to question the rationality of human choice in principle. In other words, "irrational choices arising from incomplete learning do not imply the need to modify standard choice theory" (D. Friedman, 1988, p. 942).

Critics may point out that the present results only address human choice behavior in a particular laboratory decision task. In the real world, many individuals continue to smoke cigarettes despite being aware of the negative long-term health consequences. Our results cannot be taken as evidence for the rationality or irrationality of smoking. Instead, we would emphasize that the claim that addiction is explained by melioration theory (e.g., Heyman, 1996) relies, in large part, on the misinterpretation of empirical data from simpler laboratory tasks such as the one considered here. Our contribution is to show that this misinterpretation stems from a failure to differentiate human choice under risk from choice under uncertainty.

Although the experimental paradigm presented here was designed to discriminate between economic rationality and melioration theory, an alternative framework for understanding the present results is that of ecological rationality (Goldstein & Gigerenzer, 2002; Todd, Gigerenzer, & the ABC Research Group, 2012). Our results have demonstrated that learning the true structure of an uncertain decision environment on the basis of limited experience may be both computationally and cognitively infeasible. When optimality is out of reach due to constraints of learning, finding satisficing solutions, or solutions that are good enough (Simon, 1955), may be the next best thing. Similarly, a narrow view from the laboratory obscures the fact that human choice in any one context is not a closed system and may be adapted to other environments.

Knight (1921) emphasized the distinction between problems involving risk and problems involving uncertainty: In the former case, the outcomes of actions may not be certain, but at the least, the relevant factors and consequences are enumerable and can be quantified. Similarly, Savage (1954) employed the concept of a *small world*: a decision environment in which the consequences of actions are understood in terms of probability distributions and numerical utilities over possible outcomes. In small worlds, the calculus of expected utility theory forms the basis of rational behavior. However, Savage was careful to point out the limits of this approach. Outside of small worlds (such as the typical laboratory study of human choice), not all possible outcomes and probabilities can be considered. In complex environments, people must use decision strategies that ignore some possibilities. The rationality of such behavior is based on a match between the capacities of the agent and the broader structure of the environment. As a result, mechanisms of choice like melioration, decision strategies like WSLS (Robbins, 1952), and various cognitive biases (Tversky & Kahneman, 1981) and simple heuristics (Gigerenzer et al., 2011) might only appear to be irrational when studied outside of their ecological niche (Marewski & Schooler, 2011).

Relation to Other Models of Learning and Choice

We believe that our analysis is the first to address the phenomenon of melioration from the perspective of what each individual should rationally believe about the decision environment. Acuña and Schrater (2010) considered the problem of learning the structure of a decision environment in a simpler setting, but our analysis extends their work to more complex environments in which there are unknown sequential dependencies between past actions and future outcomes.

In recent years, several reinforcement learning models have been proposed as process-level explanations of human performance in melioration tasks (Gureckis & Love, 2009a, 2009b; Neth et al., 2006). What have these models contributed to our understanding of the mechanisms of human choice, and how do they differ from the current analysis? At an abstract level, reinforcement learning is designed to provide an approximately optimal solution to a broad class of environments known as Markov decision problems (Sutton & Barto, 1998), and thus, any reinforcement learning model might be called an approximately optimal model of choice. By manipulating parameters of the learning equations or adding mechanisms such as eligibility traces, the full range of behavior from exclusive melioration through strong preference for global maximization can be predicted by such models (Neth et al., 2006). One important difference between our rational learner model and existing reinforcement learning models applied to melioration theory lies in their assumption of a prespecified or constrained mental representation of the task environment. These models therefore address the question of learning to act optimally for an existing representation of the task but do not address the more difficult challenge of learning the environmental structure. By contrast, we have emphasized that a critical learning problem lies in the fact that an appropriate understanding of the task environment is neither fixed nor given to the learner.

We note that reinforcement learning in general is not restricted in this manner. Model-based reinforcement learning (Lee, Seo, & Jung, 2012; Sutton & Barto, 1998) explicitly addresses the challenge of learning a predictive model of the environmental structure. For example, Camerer and Ho (1999) developed a simple learning model for competitive games that combines model-free reinforcement learning and belief learning. Model-based reinforcement learning has also been offered as an explanation for how people learn the perceptual consequences of actions in a sequential action task (Yakushijin & Jacobs, 2011). The rational learner

model developed here can also be considered a model-based reinforcement learning system, the first applied to the domain of melioration.

Perhaps most relevant to the present case is the model developed by Redish et al. (2007), in which the identification of unique states of the environment is also viewed as a categorization problem. Whereas the model developed by Redish et al. proposes a neural network model of learning to differentiate between states, the rational learner model developed here is based on a normative solution to this problem (following existing rational models of categorization; Anderson, 1990; Sanborn et al., 2010). Compared to other models of categorization, the present approach is simpler in that the notion of similarity between different choice sequences depends only on their utility. The choice of a Dirichlet process prior in our model defines one possible rational learner model, but other choices are possible. Alternative choices might place stronger or weaker prior probability on different functions and thus lead to different predictions for rational behavior. It seems likely that human participants make stronger assumptions about the environment, even if the task instructions given to participants do not warrant such assumptions. For instance, experiments on human function learning have shown that people possess an a priori bias towards assuming positive linear relationships among variables (Kalish et al., 2007). Such a bias would be expected to improve learning performance in the experiments presented here, since reward probability was in fact a linear function of the number of maximizing or meliorating choices in the history. Carefully documenting how such prior assumptions influence choice behavior in various environments represents an important but difficult challenge.

Also related to the present approach is the idea that decision makers may use reinforcement learning to select from a repertoire of different strategies for accomplishing any given task (Erev & Barron, 2005; Rieskamp & Otto, 2006). Competing strategies may differ in their assumptions regarding the structure of the environment or relevant informational variables for basing judgments, and successful behavior requires learning which strategy is best matched to the current environmental structure. This view of learning mirrors the present analysis by demonstrating that, in complex decision environments, learning involves more than merely assigning utilities to the physical actions available to the actor.

Conclusion

The results of the rational learner model indicate that not only would participants in the Harvard game have little evidence to favor the correct state representation of the task but many would be completely rational if they inferred an incorrect representation of the task, in which meliorating is believed to provide higher long-term rewards. In the face of significant uncertainty regarding the structure of the decision environment, there is no automatic equivalence between apparent meliorating behavior and globally suboptimal choice. One individual may rationally meliorate, while another may irrationally maximize. In many cases, people who exhibit meliorating behavior have a rational reason for doing so. Arriving at a point where these possibilities can be meaningfully distinguished requires a dramatic shift in how choice behavior in uncertain environments is studied. Beyond the concrete task environment examined in this article, the rational learner model exemplifies a methodology that demonstrates how the learnability of an uncertain decision environment can be assessed. It is our hope that rational analysis from an individual's perspective will play an increasingly important role in studying the mechanisms of human choice.

References

Acuña, D. E., & Schrater, P. (2010). Structure learning in human sequential decision-making. *PLoS Computational Biology, 6*(12), Article e1001003. doi:10.1371/journal.pcbi.1001003

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Baum, W. M. (1979). Matching, undermatching, and overmatching in studies of choice. *Journal of the Experimental Analysis of Behavior, 32,* 269–281. doi:10.1901/jeab.1979.32-269

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis, 1,* 121–143. doi:10.1214/06-BA104

Camerer, C., & Ho, T.-H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica, 67,* 827–874. doi:10.1111/1468-0262.00054

Chung, S.-H., & Herrnstein, R. J. (1967). Choice and delay of reinforcement. *Journal of the Experimental Analysis of Behavior, 10,* 67–74. doi:10.1901/jeab.1967.10-67

Davison, M., & McCarthy, D. (1988). *The matching law: A research review*. Hillsdale, NJ: Erlbaum.

Daw, N. D., & Frank, M. J. (2009). Reinforcement learning and higher level cognition: Introduction to special issue. *Cognition, 113,* 259–261. doi:10.1016/j.cognition.2009.09.005

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006, June 15). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876–879. doi:10.1038/nature04766

Dawkins, R. (1999). *The extended phenotype: The long reach of the gene* (rev. ed.). Oxford, England: Oxford University Press.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112,* 912–931. doi:10.1037/0033-295X.112.4.912

Ferguson, S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics, 1,* 209–230. doi:10.1214/aos/1176342360

Friedman, D. (1988). Monty Hall's three doors: Construction and deconstruction of a choice anomaly. *American Economic Review, 88,* 933–946.

Friedman, M., & Savage, L. J. (1948). The utility analysis of choices involving risk. *Journal of Political Economy, 56,* 279–304. doi:10.1086/256692

Fu, W.-T., & Anderson, J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General, 135,* 184–206. doi:10.1037/0096-3445.135.2.184

Gallistel, C. R. (2005). Deconstructing the law of effect. *Games and Economic Behavior, 52,* 410–423. doi:10.1016/j.geb.2004.06.012

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1,* 107–143. doi:10.1111/j.1756-8765.2008.01006.x

Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior*. New York, NY: Oxford University Press.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1998). *Markov chain Monte Carlo in practice*. London, England: Chapman & Hall.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30,* 535–574. doi:10.1146/annurev.neuro.29.051605.113038

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109,* 75–90. doi:10.1037/0033-295X.109.1.75

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112,* 21–54. doi:10.1016/j.cognition.2009.03.008

Gray, J. R. (1999). A bias toward short-term thinking in threat-related negative emotional states. *Personality and Social Psychology Bulletin, 25,* 65–75. doi:10.1177/0146167299025001006

Gray, W. D., & Fu, W. T. (2004). Soft constraints in interactive behavior: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science, 28,* 359–382.

Gureckis, T. M., & Love, B. C. (2009a). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology, 53,* 180–193. doi:10.1016/j.jmp.2009.02.004

Gureckis, T. M., & Love, B. C. (2009b). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition, 113,* 293–313. doi:10.1016/j.cognition.2009.03.013

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

Herrnstein, R. J. (1979). Derivatives of matching. *Psychological Review, 86,* 486–495. doi:10.1037/0033-295X.86.5.486

Herrnstein, R. J. (1981). Self-control as response strength. In C. M. Bradshaw, E. Szabadi, & C. F. Lowe (Eds.), *Recent developments in the quantification of steady-state operant behavior* (pp. 3–20). Amsterdam, the Netherlands: Elsevier.

Herrnstein, R. J. (1982). Melioration as behavioral dynamism. *Quantitative Analyses of Behavior, 2,* 433–458.

Herrnstein, R. J. (1991). Experiments on stable suboptimality in individual behavior. *Learning and Adaptive Economic Behavior, 81,* 360–364.

Herrnstein, R. J., Loewenstein, G. F., Prelec, D., & Vaughan, W., Jr. (1993). Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making, 6,* 149–185. doi:10.1002/bdm.3960060302

Herrnstein, R. J., & Prelec, D. (1991). Melioration: A theory of distributed choice. *Journal of Economic Perspectives, 5,* 137–156. doi:10.1257/jep.5.3.137

Herrnstein, R. J., & Prelec, D. (1992). A theory of addiction. In J. T. Elster & G. Loewenstein (Eds.), *Choice over time* (pp. 331–360). New York, NY: Russell Sage Foundation.

Herrnstein, R. J., & Vaughan, W. (1980). Melioration and behavioral allocation. In J. E. R. Staddon (Ed.), *Limits to action: The allocation of individual behavior* (pp. 143–175). New York, NY: Academic Press.

Heyman, G. M. (1996). Resolving the contradictions of addiction. *Behavioral and Brain Sciences, 19,* 561–610. doi:10.1017/S0140525X00042990

Heyman, G. M., & Dunn, B. (2002). Decision biases and persistent illicit drug use: An experimental study of distributed choice and addiction. *Drug and Alcohol Dependence, 67,* 193–203. doi:10.1016/S0376-8716(02)00071-6

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review, 14,* 288–294. doi:10.3758/BF03194066

Katsikopoulos, K. V., Schooler, L. J., & Hertwig, R. (2010). The robust beauty of ordinary information. *Psychological Review, 117,* 1259–1266. doi:10.1037/a0020418

Knight, F. (1921). *Risk, uncertainty, and profit*. Boston, MA: Houghton Mifflin.

Kudadjie-Gyamfi, E., & Rachlin, H. (1996). Temporal patterning in choice among delayed outcomes. *Organizational Behavior and Human Decision Processes, 65,* 61–67. doi:10.1006/obhd.1996.0005

Kudadjie-Gyamfi, E., & Rachlin, H. (2002). Rule-governed versus contingency-governed behavior in a self-control task: Effects of changes in contingencies. *Behavioural Processes, 57,* 29–35. doi:10.1016/S0376-6357(01)00205-4

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience, 35,* 287–308. doi:10.1146/annurev-neuro-062111-150512

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118,* 393–437. doi:10.1037/a0024143

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics, 9,* 249–265.

Neth, H., Sims, C. R., & Gray, W. D. (2006). Melioration dominates maximization: Stable suboptimal performance despite global feedback. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 627–632). Hillsdale, NJ: Erlbaum.

Otto, A. R., Gureckis, T. M., Markman, A. B., & Love, B. C. (2009). Navigating through abstract decision spaces: Evaluating the role of state generalization in a dynamic decision-making task. *Psychonomic Bulletin & Review, 16,* 957–963. doi:10.3758/PBR.16.5.957

Otto, A. R., Markman, A. B., & Love, B. C. (2012). Taking more, now: The optimality of impulsive choice hinges on environment structure. *Social Psychological and Personality Science, 3,* 131–138. doi:10.1177/1948550611411311

Rachlin, H., & Laibson, D. I. (Eds.). (1997). *The matching law: Papers on psychology and economics by Richard Herrnstein*. New York, NY: Russell Sage Foundation.

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review, 114,* 784–805. doi:10.1037/0033-295X.114.3.784

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135,* 207–236. doi:10.1037/0096-3445.135.2.207

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society, 58,* 527–535. doi:10.1090/S0002-9904-1952-09620-8

Sakai, Y., & Fukai, T. (2008). When does reward maximization lead to matching law? *PLoS ONE, 3*(11), Article e3795. doi:10.1371/journal.pone.0003795

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117,* 1144–1167. doi:10.1037/a0020511

Savage, L. J. (1954). *The foundations of statistics*. New York, NY: Wiley.

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology, 53,* 491–517. doi:10.1146/annurev.psych.53.100901.135213

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69,* 99–118. doi:10.2307/1884852

Staddon, J. E. R. (1992). Rationality, melioration, and law-of-effect models for choice. *Psychological Science, 3,* 136–141. doi:10.1111/j.1467-9280.1992.tb00013.x

Stevens, J. R., Volstorf, J., Schooler, L. J., & Rieskamp, J. (2011). Forgetting constrains the emergence of cooperative decision strategies. *Frontiers in Psychology, 1,* Article 235. doi:10.3389/fpsyg.2010.00235

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology, 53,* 168–179. doi:10.1016/j.jmp.2008.11.002

Stillwell, D. J., & Tunney, R. J. (2009). Melioration behavior in the Harvard game is reduced by simplifying decision outcomes. *Quarterly Journal of Experimental Psychology, 62,* 2252–2261. doi:10.1080/17470210902765999

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

Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. New York, NY: Macmillan. doi:10.5962/bhl.title.55072

Todd, P. M., Gigerenzer, G., & the ABC Research Group. (2012). *Ecological rationality: Intelligence in the world*. New York, NY: Oxford University Press. doi:10.1093/acprof:oso/9780195315448.001.0001

Tunney, R. J., & Shanks, D. R. (2002). A re-examination of melioration and rational choice. *Journal of Behavioral Decision Making, 15,* 291–311. doi:10.1002/bdm.415

Tversky, A., & Kahneman, D. (1981, January 30). The framing of decisions and the psychology of choice. *Science, 211,* 453–458. doi:10.1126/science.7455683

Vaughan, W. (1981). Melioration, matching, and maximization. *Journal of the Experimental Analysis of Behavior, 36,* 141–149. doi:10.1901/jeab.1981.36-141

Vaughan, W., & Herrnstein, R. J. (1987). Stability, melioration, and natural selection. In L. Green & J. H. Kagel (Eds.), *Advances in behavioral economics* (Vol. 1, pp. 185–215). Norwood, NJ: Ablex Publishing.

Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Warry, C. J., Remington, B., & Sonuga-Barke, J. S. (1999). When more means less: Factors affecting human self-control in a local versus global choice paradigm. *Learning and Motivation, 30,* 53–73. doi:10.1006/lmot.1998.1018

Wilson, J. Q., & Herrnstein, R. J. (1985). *Crime and human nature*. New York, NY: Simon & Schuster.

Wilson, R. C., & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in Human Neuroscience, 5, Article 189*. doi:10.3389/fnhum.2011.00189

Worthy, D. A., Otto, A. R., & Maddox, W. T. (2012). Working-memory load and temporal myopia in dynamic decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38,* 1640–1658. doi:10.1037/a0028146

Yakushijin, R., & Jacobs, R. A. (2011). Are people successful at learning sequences of actions on a perceptual matching task? *Cognitive Science, 35,* 939–962. doi:10.1111/j.1551-6709.2011.01176.x

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin & Review, 12,* 387–402. doi:10.3758/BF03193783

Yechiam, E., Erev, I., Yehene, V., & Gopher, D. (2003). Melioration and the transition from touch-typing training to everyday use. *Human Factors, 45,* 671–684. doi:10.1518/hfes.45.4.671.27085

Appendix

Details of the Bayesian Inference Procedure

The challenge for the rational learner model is to infer a posterior distribution over three quantities, as described by Equation 1 in the main text: the relevant history window, w; the function assigning choice sequences to unique states of the environment, *f*; and the reward probabilities associated with each state, indicated by θ.

If the history window is known, then the conditional posterior distribution $p(\theta, f|X, w)$ is equivalent to a straightforward application of a Dirichlet process mixture model with a conjugate prior (Neal, 2000). To approximate this posterior distribution, Markov chain Monte Carlo algorithms (Gilks et al., 1998) can be applied to generate a large set of samples from this distribution. In particular, our implementation follows the collapsed Gibbs sampler algorithm described by Neal (2000, Algorithm 3).

This leaves the challenge of computing the full posterior distribution, $p(\theta, f, w|X) = p(\theta, f|X, w)p(w|X)$. Determining the posterior distribution over history windows, $p(w|X)$, requires computing the marginal likelihood, or normalizing constant for the Dirichlet process. While direct calculation of this quantity is intractable, Blei and Jordan (2006) developed a variational approximation algorithm that computes a lower bound on the marginal likelihood for Dirichlet process mixture models. This algorithm was used to compute an approximation to $p(w|X)$ for each history window in the range of 0 through 10 previous choices.

For each possible history window, 1,000 samples were obtained from the posterior distribution $p(\theta, f|X, w)$, using a collapsed Gibbs sampling algorithm. These samples were then resampled according to $p(w|X)$ for each history window $w = 0 \ldots 10$, resulting in a set of 1,000 samples distributed according to the joint posterior distribution over the three quantities of interest. These posterior samples were then used to compute the expected value for various decision strategies (e.g., pure melioration or pure maximization) given the observed evidence *X*.

Received August 25, 2011
Revision received August 31, 2012
Accepted September 25, 2012 ■