Journal of Mathematical Psychology 53 (2009) 180–193

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Learning in noise: Dynamic decision-making in a variable environment

Todd M. Gureckis [a,*], Bradley C. Love [b]

[a] New York University, United States

[b] The University of Texas at Austin, United States

ARTICLE INFO

Article history:
Received 14 January 2009
Received in revised form
12 February 2009
Available online 2 April 2009

ABSTRACT

In engineering systems, noise is a curse, obscuring important signals and increasing the uncertainty associated with measurement. However, the negative effects of noise are not universal. In this paper, we examine how people learn sequential control strategies given different sources and amounts of feedback variability. In particular, we consider people's behavior in a task where short- and long-term rewards are placed in conflict (i.e., the best option in the short-term is worst in the long-term). Consistent with a model based on reinforcement learning principles [Gureckis, T., & Love, B.C. Short term gains, long term pains: How cues about state aid learning in dynamic environments. *Cognition* (in press)], we find that learners differentially weight information predictive of the current task state. In particular, when cues that signal state are noisy, we find that participants' ability to identify an optimal strategy is strongly impaired relative to equivalent amounts of noise that obscure the rewards/valuations of those states. In other situations, we find that noise and noise in reward signals may paradoxically *improve* performance by encouraging exploration. Our results demonstrate how experimentally-manipulated task variability can be used to test predictions about the mechanisms that learners engage in dynamic decision making tasks.

Millions of Americans suffer from chronic illnesses such as heart disease and diabetes, and must carefully monitor their diet and exercise. However, making healthy lifestyle choices can be difficult. Due to the intrinsic variability in outcomes and observations, decision makers face uncertainty about both the actual state of their health and the costs and benefits associated with various dietary options, complicating the management of illness. For example, daily fluctuations in blood pressure and glucose levels can obscure estimates of actual health. This difficulty in separating signal from noise extends beyond understanding our physical state to the external environment. For example, it is difficult to obtain reliable information on the sodium, fat, and calorie content of foods. Under these circumstances, making effective decisions requires one to manage uncertainty from many sources.

In this paper, we explore how people learn effective decision-making strategies given similar kinds of uncertainty about what is signal and what is noise. Decision-making under uncertainty has played a central role in judgement and decision-making research. However, empirical attempts to understand this ability have often focused on decisions made in static, one-off situations based on verbal descriptions of choice alternatives, as in gambles (Barron & Erev, 2003). More recently, researchers have adopted decision-theoretic approaches to understanding choice behavior in more realistic, online, and dynamic situations (Edwards, 1962, see Busemeyer, 2002 for a recent review). Like the real-world example of managing a chronic illness, participants in these tasks are asked to achieve a particular goal by making a sequence of decisions from one moment to the next based on their ongoing experience (Gureckis & Love, in press; Stanley, Mathew, Russ, & Kotler-Cope, 1989).

In our experiments, we examine human learning in a dynamic decision-making task, called the "Farming on Mars" task, where the experienced reward structure continually evolves in response to the actions of the individual (Gureckis & Love, in press). A key feature of our task is that the strategy that returns the most reward over the course of the experiment requires participants to forego immediately attractive short-term options in favor of a long-term beneficial strategy (Herrnstein, 1991; Herrnstein & Prelec, 1991). Just like the dilemma between deciding whether to eat a healthy meal or indulge in a higher calorie dessert, participants who seek to maximize their long-term well-being must forego immediately attractive alternatives. The extension presented here is to consider how decision maker's ability to uncover a non-obvious reward-maximizing strategy is impacted when relevant outcomes in the task are obscured by variability or noise (i.e., decisions are made with increasing uncertainty about the value of the mean).

* Corresponding address: Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, United States.
E-mail address: todd.gureckis@nyu.edu (T.M. Gureckis).

0022-2496/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2009.02.004

The problem of learning an effective response strategy in our task is one of adaptive control – participants interact with a system and are asked to make a continuous sequence of decisions which guide the dynamics of the system (Brehmer, 1992). We model human behavior in the task using a framework based on Markov Decision Processes (MDPs) and an approximate solution method known as Reinforcement Learning (RL). RL is an agent-based approach to learning through interaction with the environment in pursuit of reward-maximizing behavior (Sutton & Barto, 1998). The focus of RL research is to understand how a situated agent interacting with a responsive environment can arrive at effective strategies, making RL an excellent tool for studying human learning and decision-making in dynamic tasks (Busemeyer & Pleskac, in this issue; Fu & Anderson, 2006; Sun, Slusarz, & Terry, 2005). Interestingly, the RL model that we develop in the later section makes specific predictions about the effects that different sources of noise may have on learning performance, which we then test in our experiments.

0.1. Noise as a Curse and Noise as a Tool

Anyone who has tried talking on a bad telephone connection would likely view noise as a curse that obscures important signals and increases the uncertainty associated with measurement. In most systems, decreasing signal-to-noise ratios results in lower performance (Green & Swets, 1966). Likewise, noise or variability, when interpreted by learners as signal, can guide behavior in surprising (and potentially maladaptive) ways. For example, Skinner (1948) showed how pigeons in an operant conditioning experiment, which were given reinforcement at random intervals, enacted a number of repetitive but "superstitious" behavioral patterns. Similarly, failures of statistical reasoning such as the regression to the mean and the gambler's fallacy, reflect the human tendency to inappropriately view normal variation (i.e., noise) as signal (Kahneman & Tversky, 1973; Tversky & Kahneman, 1971). However, the negative effects of noise are not universal. For example, noise can help overcome local minima such as in systems based on simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983). In addition, in some environments, the inability to effectively separate signal and noise may, paradoxically, lead to improvements in behavior by encouraging alternative strategies or by focusing learners on alternative information sources (Burns, 2004). In this paper, we examine situations where variability or noise actually help people make decisions that maximize their long-term benefit.

In the laboratory, experimenter-manipulated noise is used as a tool to illuminate the structure of human perceptual and cognitive systems, since the way the performance degrades in noise can reveal aspects of system architecture (Gold, Sekuler, & Bennett, 2004; Green & Swets, 1966; Lu & Dosher, 1999; Pelli & Farell, 1999). For example, techniques such as equivalent input noise (Pelli & Farell, 1999) estimate an observer's internal perceptual noise level by systematically degrading external stimuli. We adopt a similar perspective of "noise as a tool" in the experiments that follow. In particular, we parametrically degraded the signal-to-noise characteristics of feedback given to participants in the Farming on Mars task in order to estimate how such variability impacts performance. Before describing our experimental manipulations in more detail, we begin by describing the basic version of the Farming on Mars task and briefly review some previously reported findings using this paradigm.

0.2. The farming on mars task

In the "Farming on Mars" task utilized in our experiments, participants interact with a repeated, two-choice decision-making task presented as a simple video game. The cover story for the game is that two agricultural robots have been sent to the planet Mars in order to establish a farming system capable of generating oxygen for later human inhabitants. Participants are informed that each robot specializes in a different set of farming practices, but that only one robot can be active at a given moment. Participants' job as controller is to select, on each trial, which robot should be employed in order to maximize the total oxygen generated over the entire experiment. Participants indicate which robot should do the farming, and are given feedback about how much oxygen was generated on that trial as a result of their choice.

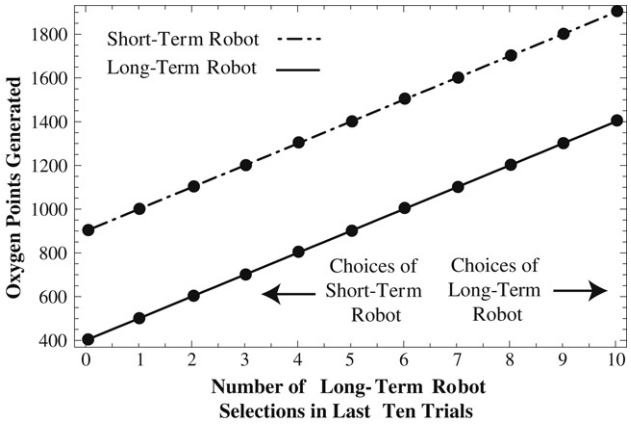

Fig. 1. The payout function for the Farming on Mars task. The horizontal axis is the number of choices out of the last ten in which the Long-Term robot was selected. The vertical axis is the number of oxygen units generated as a result of choosing one of the robots on a trial. The two diagnonal lines show the reward associated with each robot for each state. By design the Short-Term robot is better for every state (i.e., trial), but the best long-term strategy is to exclusively choose the Long-Term robot because the selection of the Short-Term robot transitions the state to the left, whereas selection of the Long-Term robot transitions the state to the right.

Unknown to participants at the start of the task, there is a contingency between recent robot selections and the oxygen points received on the next trial. In particular, the current payoff depends on the relative allocation of responses to one of the two choice options presented on each trial. For expository purposes we will refer to one of the choice options as the "Short-term robot" and the other as the "Long-Term robot" although these labels were not provided to participants in the experiment. Fig. 1A shows an example of the payoff structure used in the task. The horizontal axis in Fig. 1 measures the participant's current allocation to the Long-Term robot over the last ten trials (ranging from 0 to 10). The two diagonal lines describe the function relating the current choice history to the reward at any point in time. The upper diagonal line illustrates the reward received from selecting the Short-Term robot as a function of recent choice history, and the lower diagonal line shows the reward from selecting the Long-Term robot.

Note that the Short-Term robot always generates more oxygen than the Long-term robot on any given trial (i.e., the reward function for the Short-Term robot reward function is larger than for the Long-Term robot in every situation). However, each time the Short-Term robot is selected, the output of both robots is lowered on the following trial (i.e., the state of the system shifts to the left along the horizontal axies in Fig. 1). Selections of the Long-Term robot behave in the opposite fashion. Each time this robot is selected, the output of both robots is *increased* on the next trial (shifted to the right in Fig. 1A). Critically, over the window of the last ten trials, the reward received from repeatedly selecting the Long-Term robot exceeds that from always selecting the Short-Term robot (i.e., the highest point of the Long-Term robot curve is above the lowest point for the Short-Term robot curve in Fig. 1A). As a result, the optimal strategy is to select the Long-Term robot on every trial, even though selecting the Short-Term robot would

earn more on any single trial. In the experiment, participants are not given any relevant information about the differences between the robots, and thus can only arrive at the optimal strategy by interactively exploring the behavior of the system (cf. Berry and Broadbent (1988) and Stanley et al. (1989)).

The structure of the Farming on Mars task borrows from a growing literature looking at choice behavior in situations where actions that lead to long-term rewards conflict with those that yield immediate rewards (Herrnstein, 1991; Herrnstein & Prelec, 1991; Neth, Sims, & Gray, 2006; Tunney & Shanks, 2002). Interestingly, the conclusion from much of this work has been that both humans and other animals often fail to inhibit the tendency to select an initially attractive option even when doing so leads to lower rates of reinforcement relative to other strategies, a phenomena referred to as *melioration*. While melioration is sometimes taken as evidence that people fail to maximize their long-term expected utility (see Tunney and Shanks (2002) for a similar discussion), in practice, it is a significant challenge for learners to discover the optimal strategy in an unknown environment.

0.3. Why is the "Farming on Mars" task so difficult? Effective exploration and the problem of perceptual aliasing

The Farming on Mars task is a challenging task for a number of reasons. First, from the outset of the task, participants may have little sense of how their own actions are influencing the behavior of the system. Effective learning requires exploration of hidden contingencies between the agent's past actions and future prospects. For example, work with a similar task found that manipulations that encourage subjects to "explore" the system by lowering the costs of making early choices can lead to improvements in later performance (Tunney & Shanks, 2002). Similarly, recent work has shown that motivational manipulations can influence the degree to which participants are willing to adopt exploratory strategies in sequential choice tasks, which can in turn translate into improvements in performance in tasks with non-obvious solutions (Worthy, Maddox, & Markman, 2007).

In light of these findings, one hypothesis is that, under certain conditions, performance may actually improve when outcomes in the task are more variable. Imagine trying to find the "globally" best restaurant out of a large set. If you have an enjoyable meal at the first place you visit, you may be less willing to sample other places. However, if each time you visit this restaurant, the quality varies due to random noise, it may coincidently help encourage more extensive sampling of other options (Denrell, 2005). A bad experience one day might lead you someplace else for your next meal. In Experiment 1, we set up an analogous situation in the Farming on Mars task by parametrically varying the amount of variability associated with the reward signal in the task and assessing its impact on participants' ability to discover the reward-maximizing strategy.

Effective exploration is one of multiple challenges facing learners in the Farming on Mars task. A second, but related challenge arises from participants' mental representation of the task. Each time a participant makes a choice in the task, the system changes so that the reward received on the next trial is different than it was on the previous trial. Recognizing how the current situation or "state" of the world is changing as a result of the one's actions is essentially a categorization problem. The learner must identify when the current situation is different, and how to generalize their experience from one situation to the next (Redish, Jensen, Johnson, & Kurth-Nelson, 2007; Veksler, Gray, & Schoelles, 2007). In the standard version of the task, identifying these changes is difficult given that there are few direct cues available indicating to subjects that the world is changing on each trial. As a result, participants must overcome the problem of *perceptual aliasing* in which relevant states in the world are poorly differentiated (Whitehead & Ballard, 1991). Returning to the restaurant example, imagine trying to find the best restaurant when not only the average quality of the food is obscured by randomly distributed noise, but there is also uncertainty about which restaurant one is dining at because all restaurants are identical inside. Failing to differentiate situations associated with different rewards makes learning the true reward structure a difficult task.

The issue of how human learners use cues in the environment to disambiguate distinct states or "situations" remains an active question in reinforcement learning research (Daw, 2003; Gureckis & Love, in press; Redish et al., 2007; Veksler et al., 2007). However, in a recent set of studies (Gureckis & Love, in press), we evaluated how simple visual cues can impact participants' performance in the Farming on Mars task. In one experiment (Gureckis & Love, in press, Experiment 2), displays were augmented with a row of indicator lights. The position of the active light in the display served as a cue about the current state of the system. Participants given cues which correlated with the underlying system state performed better than participants attempting to learn without these cues. In addition, cues which allowed generalization from one situation (i.e., state) to other states performed best. These studies demonstrate the importance of state representation in dynamic and complex task environments. The state representation that the learner adopts may act as a "framework" for effectively structuring, integrating, and generalizing experience. However, to the degree to which participants use information about the current state to structure and integrate their learning experiences, variability on such cues should have a dramatic influence on performance. Variability in the cues that signal the current task state effectively increase the degree of perceptual aliasing by making one unclear at any point in time how state knowledge should be updated given current rewards. In the restaurant example, a diner may know the meal is enjoyable, but be confused about which restaurant served the meal.

0.4. Overview and summary

Previous studies establish that two somewhat similar sources of information may jointly influence performance in sequential decision-making tasks. The first is the structure of the rewards in the task (i.e., the payoff function). The second is information about functionally distinct task states (any cue which can help distinguish or categorize different states or situations). These two sources of information are both functionally and psychologically distinct. State information indicates the agent's place in the overall system and can help link future actions, whereas reward information provides the valuation of those actions. Given that these two sources of information or signals play a distinct role in learning, we predicted that learners would respond differently to noise (i.e, variability) associated with either signal.

The remainder of the paper is structured as follows. We begin by describing the results of two novel experiments. In the first study, participants attempted to uncover a reward-maximizing response strategy given different amounts of noise that obscured the valuation of particular actions. In contrast to the conventional view that noise has a uniformly negative impact on system performance, we find that moderate amounts of noise actually *improves* the ability of participants to find an optimal response strategy by increasing their tendency to explore alternative strategies. In the second study, we compare the effect that different *sources* of noise have on performance in the task. In particular, we compare participants' performance when comparable levels of noise obscure either the reward or state

signals. We find that variability that impairs participants' ability to identify the current task state hurts performance more than noise that obscures the valuation of those states. This result is surprising given that attending to information about the current task state is, in some ways, optional for accomplishing the primary task of maximizing reward. Simulations show that a simple model based on RL principles, including the balance of exploration and exploitation, the appreciation of future outcomes, and the identification of distinct task states, provides an excellent account of behavior across the two experiments. Our results illuminate how experimentally-manipulated task variability can be used to test predictions about the mechanisms that learners engage in dynamic decision making tasks.

1. Experiment 1

Experiment 1 evaluates the impact of variability in the reward signal on task performance. Using the Farming on Mars task, participants were assigned to one of four conditions which were identical with respect to the number of trials and the underlying payoff function, but which differed in the way in which variability was added to the experienced rewards. In each condition, the dependence of current reward on past choices was determined by the number of choices made to the Long-Term robot over the last ten trials (as illustrated in Fig. 1). In the *no-noise* condition, there were no additional sources of variability. In contrast, in the *low-noise*, *medium-noise*, and *high-noise* conditions, on each trial, a normally distributed random noise (with mean equal to zero and standard deviation σ_r) was added to the payoff indicated in Fig. 1. This variability in the experienced rewards obscured the underlying structure of the task. One straightforward prediction is that performance should degrade as the amount of noise increased (consistent with work finding a decrease in performance as the discriminability of two options is lowered, Busemeyer and Myung (1992)). On the other hand, moderate amounts of noise might, in some circumstances, actually help participants adopt appropriate exploration strategies. The inconsistent feedback that arises in a noisy environment might rule out simple hypotheses and encourage sampling of alternatives.

1.1. Method

1.1.1. Participants

Ninety-two undergraduates from New York University and the University of Texas participated for course credit and a small cash bonus which was tied to performance. Participants were randomly assigned to one of the four conditions: the no-noise (NN) condition, a low noise (LN) condition, a medium noise (MN) condition, and a high noise (HN) condition. Twenty-three participants were included in each condition.

1.1.2. Materials

The experiment was run on standard desktop computers using an in-house data collection system written in Python. Stimuli and instructions were displayed on a 17-inch color LCD positioned approximately 47 centimeters away from the participant. Participants were tested individually in a single session. Extraneous display variables, such as which robot corresponds to the left or right choice option, were counterbalanced across participants. In addition, no other relevant information distinguished between the two options other than the rewards that the participant received for their choices.

Table 1
Summary of the Conditions in Experiment 1. The column titled d' Between Reward Curves measures the discriminability of the two reward curves as a function of the increasing variability. Similarly, the column titled d' Between States measures the discriminability of two adjacent states (i.e, between two adjacent points in the same reward curve).

Condition	Noise level	d' Between Reward Curves	d' Between States
No noise (NN)	$\sim N(0, \sigma_r = 0)$	–	–
Low noise (LN)	$\sim N(0, \sigma_r = 70)$	7.14	1.43
Medium noise (MN)	$\sim N(0, \sigma_r = 100)$	5.0	1.0
High noise (HN)	$\sim N(0, \sigma_r = 300)$	1.67	0.33

1.1.3. Design

Participants were given a simple two-choice decision-making task (the Farming on Mars task described above) presented as a simple video game (see Fig. 4 for an example display). At the start of the experiment, participants were presented with instructions on the screen that described the basic cover story and task. Participants were informed that their goal was to maximize the total output from the Mars Farming system over the entire experiment by selecting of one of two robot systems on each trial. Unknown to participants, the number of oxygen units generated at any point in time was a function of their choice history over the previous ten trials. In addition, the payoffs associated with each robot system were manipulated so that one option was better than the other in the long-term, despite appearing worse in the short-term. A graphical depiction of the rewards in the tasks is shown in Fig. 2. If h represents the number of trials in the last 10 which were allocated to the Long-Term robot, then the payoff for any selection of the Long-Term Robot was $400 + 1000 * \frac{h}{10} + N(\mu = 0, \sigma_r)$. Alternatively, the payoff for the Short-Term robot was $900 + 1000 * \frac{h}{10} + N(\mu = 0, \sigma_r)$. At the start of the experiment, we initialized h to 5 (so as to not favor either option). In these rewards equations, $N(\mu, \sigma_r)$ indicates a normally distributed random number with mean μ and standard deviation σ_r. The parameter σ_r varied between conditions. In the NN condition $\sigma_r = 0$, in the LN condition $\sigma_r = 70$, in the MN condition $\sigma_r = 100$, and in the HN condition $\sigma_r = 300$. The effect of the random trial-to-trial noise was to degrade the signal-to-noise characteristics of the reward signal. Table 1 summarizes the difference between conditions in the discriminability of different task states and between the two reward curves while Fig. 2 illustrates the payoff functions in each condition, including the 95% confidence intervals. As is visible in the figure, as the variability increases the discriminability of the two reward curves (i.e., for the Short-term and Long-term options), as well as the transitions between successive task states is lowered.

1.1.4. Procedure

The 500 trials of the experiment were divided into five blocks of 100 trials each. At the end of each block, participants were given a short break and each successive block picked up where the last block left off. In order to maintain motivation, participants were told that they could earn a small cash bonus of $2–5 which was tied to their oxygen generating performance in the task. However, participants were not told how oxygen points would translate into cash rewards, only that generating more oxygen would yield a larger bonus.

On each trial, participants were shown a control panel with two response buttons labeled either *System* 1 or *System* 2. Between these two buttons was a video display where trial-relevant feedback and instructions were presented. Participants clicked one of the two response buttons using a computer mouse. After a selection was made, a short animation (lasting approximately 800 ms) indicated that the response was being sent to the Mars base. Following this animation, the amount of oxygen generated on that trial was shown in numerical terms (i.e., "New Oxygen Added: 800.00"). A short auditory beep was presented when the oxygen

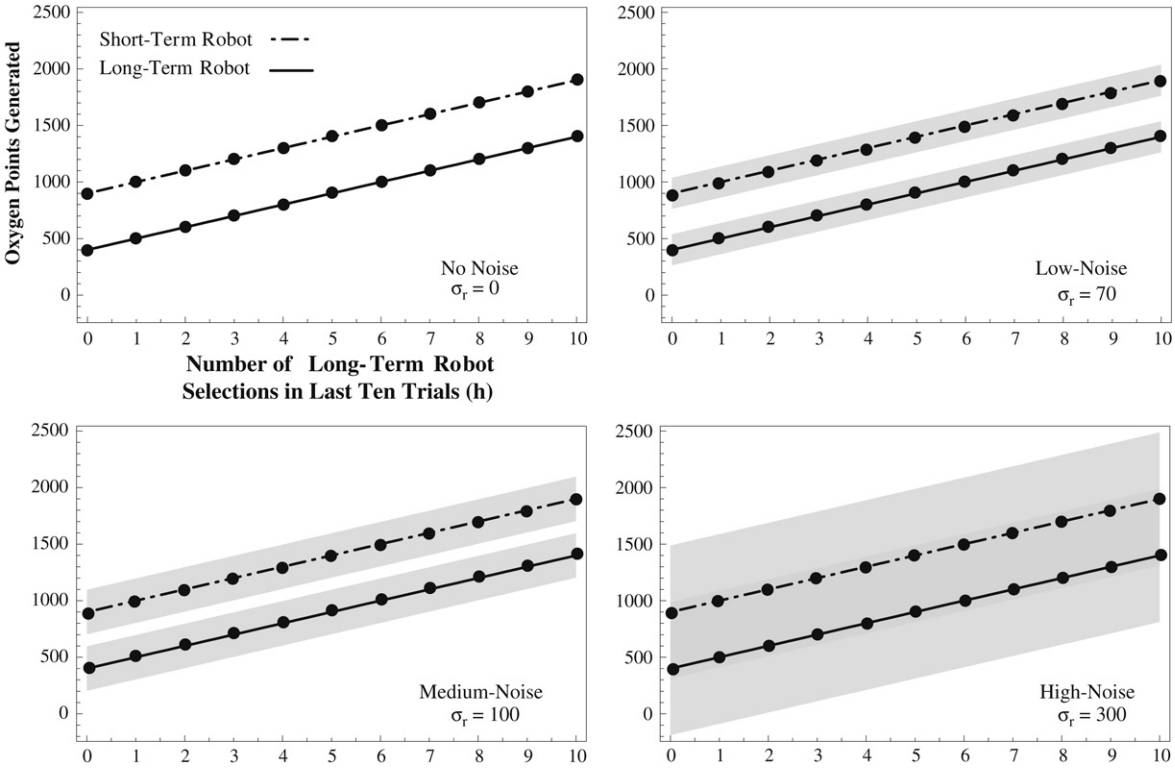

Fig. 2. The reward functions used in each condition of Experiment 1 as a function of the current task state (i.e., the number of Long-term Robot selections in the last 10 trials). The shaded regions show the 95% confidence intervals for each noise condition. See Table 1 for a summary of the effect this noise has on the ability to discriminate the rewards from different states and actions.

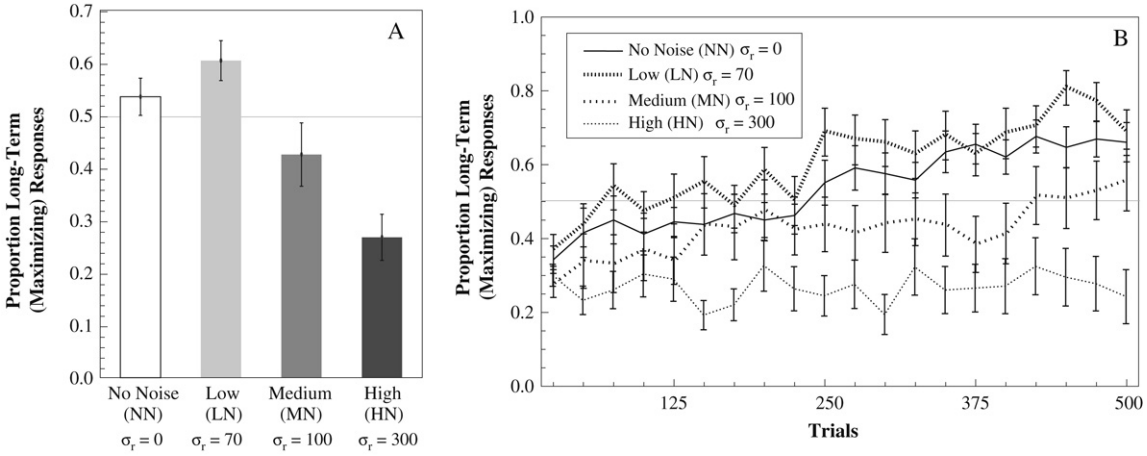

Fig. 3. Overall results of Experiment 1. Panel A shows the average proportion of long-term (maximizing) responses made throughout the experiment as a function of noise condition. Panel B presents the average proportion of maximizing responses considered in blocks of 25 trials for all four conditions. In both figures, the horizontal line at 0.5 shows chance performance. In each graph, error bars are standard errors of the mean.

points display was updated, indicating that the reward for that trial had been received. The points display was shown for 800 ms, after which the screen reset to a "Choose" prompt that indicated the start of the next trial. No information about the cumulative oxygen generated across trials was provided.

1.2. Results

1.2.1. Overall performance

Fig. 3A shows the proportion of trials in which the Long-Term robot (i.e., reward-maximizing response) was chosen in each condition of the experiment. A one-way ANOVA on these proportions revealed a significant effect of condition, $F(3, 91) = 10.42$, $p < .001$. In both the NN and MN conditions, the proportion of Long-Term choices did not significantly differ from chance ($M = .53$, $SD = .17$, $t < 1$ and $M = .42$, $SD = .28$, $t(22) = -1.24$, $p = .23$, respectively). However, in the LN condition, participants chose the Long-Term robot more often than the Short-Term robot, $M = .60$, $SD = .18$, $t(22) = 2.74$, $p = .012$. In contrast, in the HN condition, participants chose the Short-Term option more often, $M = .26$, $SD = .21$, $t(22) = -5.29$, $p < .001$. Planned comparison revealed that the proportion of maximizing responses did not differ between the NN and LN conditions, $t(44) = 1.32$, $p = .19$. However, a significantly larger proportion of maximizing responses was recorded in the LN condition compared to the MN, $t(44) = 2.51$, $p = .016$, and HN conditions, $t(44) = 5.79$, $p < .001$.

1.2.2. Time-course analysis

Fig. 3B shows the proportion of Long-Term choices calculated in non-overlapping blocks of 25 trials at a time. On average, participants in all conditions except the HN condition increased the number of Long-Term choices they made over the course of the experiment. A repeated measures ANOVA on training block (1–5) and condition revealed a main effect of block ($F(4, 352) = 15.66$, $p < .001$), condition ($F(3, 352) = 10.49$, $p < .001$), and a significant interaction ($F(12, 352) = 2.12$, $p = .016$). Planned comparisons confirmed a significant effect of training block in both the NN, LN, and MN conditions ($F(4, 88) = 10.76$, $p < .001$, $F(4, 88) = 9.18$, $p < .001$, and $F(4, 88) = 2.92$, $p = 0.03$, respectively). However, there was no evidence of learning over blocks in the HN condition ($F < 1$). Comparing the proportion of selections allocated to the Long-Term option in the first block of 100 trials compared to the last block of 100 trials revealed a significant increase in both the NN (mean difference $= .26$, $t(22) = 3.72$, $p = .001$), LN (mean difference $= .29$, $t(22) = 5.11$, $p < .001$), and MN (mean difference $= .20$, $t(22) = 3.08$, $p = .005$) conditions, with no difference in the HN condition (mean difference $= .01$, $t < 1$).

1.3. Discussion

In Experiment 1, we assessed the impact that reward variability had on participants' performance in the Farming on Mars task. Overall, we found that as the ratio of signal-to-noise dropped (as in the medium-noise or high-noise condition), participants made fewer selections of the Long-term (reward maximizing) option. In fact, we found little evidence of learning over blocks in the high-noise condition, suggesting that the reward structure of the task was all but eliminated by the task variability. These results are consistent with the view that noise is harmful. However, we found that moderate amounts of noise (as in the low-noise condition) may actually improve performance. Indeed, participants in the low-noise condition were unique in consistently choosing the Long-term option.

Introducing variability in the rewards influenced the way participants allocated their responses with moderate variability in reward outcomes leading to more effective decision making (indeed, participants earned more points on average in the LN condition than in the NN condition $M = 577085$ vs. $M = 593840$, $t(44) = 2.51$, $p < .02$). In contrast to the view that noise always has negative consequences for performance, we found that moderate amounts of variability actually helped participants to adopt more effective response strategies. One hypothesis, that we address in simulations to follow, is that the inconsistent trial-to-trial feedback with noise may have helped to encourage more effective exploration in the task.

2. Experiment 2: Comparing the effects of variability on rewards and states

Experiment 1 evaluated how just one kind of task variability (noise on the reward signal) impacts performance in the Farming on Mars task. The choice to focus on reward variability in Experiment 1 was natural since rewards are primary to the task (the goal of the task is to make points). In Experiment 2, we extend this analysis to consider the impact that variability on task cues (i.e., state cues) has on performance. In this experiment, we provided participants with two pieces of information which could help them discover the best strategy. The first was the magnitude of the reward signal, and the second was an additional "soil indicator" that provided information about the current task state. In this case, the position of the soil indicator correlated with the current task state which is determined by the number of Long-term choices made over the last ten trials. By adding equivalent amounts of trial-to-trial noise to either the reward signal or to the soil indicator, we assessed the differential effect that each signal had on participants' performance.

In previous work (Gureckis & Love, in press), we found that cues which indicated the current task state facilitated the ability of participants to learn the optimal response strategy. According to our account, disambiguating state information aided participants by reducing aliasing of distinct task states. In these previously reported experiments, we also tested a condition where the trial-to-trial dynamics of the state cue information was highly irregular (shuffled-cue condition of Experiment 2, (Gureckis & Love, in press)) but still correlated perfectly with each task state (i.e., the location of the state cue was shuffled such that adjacent system states might not have similar state cues). Shuffled cues helped participants but not to the extent that such cues do when directly aligned with the task structure. Thus, in Experiment 2, we predict that highly variable state information should have a more detrimental impact on performance than would variability in rewards. To the degree that participants use the state cues as a way of structuring their learning in the task, inconsistent information on this signal should severely disrupt updating of the value of particular state-action pairings. In effect, treating noise-as-signal on the state cues can lead to more perceptual aliasing rather than less.

2.1. Method

2.1.1. Participants

One-hundred and eight undergraduates from New York University and the University of Texas participated for course credit and a small cash bonus which was tied to performance. Participants were randomly assigned to either a no-noise (NN) condition, a reward-noise (RN) condition, a state-noise (SN) condition, or a state and reward noise (SRN) condition. Twenty-seven participants were included in each condition.

2.1.2. Materials and design

The materials and general methods in Experiment 2 were the same as in the no-cue condition of Experiment 1. However, two feedback signals were displayed on each trial: the reward (i.e., oxygen point) resulting from the last choice, and the current system state (i.e., soil indicator).

The reward on each trial was determined as follows. If h represents the number of trials in the last 10 which were allocated to the Long-Term robot, then the payoff for any selection of the Long-Term Robot was $20 + 100 * \frac{h}{10} + N(\mu = 0, \sigma_r)$. Alternatively, the payoff for the Short-Term robot was $30 + 100 * \frac{h}{10} + N(\mu = 0, \sigma_r)$. At the start of the experiment, we initialized h to 5 (so as to not favor either option). As in Experiment 1, $N(\mu, \sigma_r)$ indicates a normally distributed random number with mean μ and standard deviation σ_r that was added to the reward function on each trial. In the NN and SN conditions, σ_r was set to zero (i.e., no noise on rewards). In the RN and SRN condition σ_r was set to 12.0. As before, the effect of the random trial-to-trial noise was to degrade the signal-to-noise characteristics of the reward signal. Instead of expressing rewards in numerical terms (i.e.,"800 oxygen points"), rewards were displayed as a vertical bar, the height of which indicated the amount of oxygen generated on a given trial (see Fig. 4). The maximum height of the reward bar was set to 150.0 units. If, due to noise, the reward for a given trial exceeded that value it was clipped. Similarly, negative rewards were clipped at zero. The purpose of this change was to make the nature of the reward signal and state signal similar.

Fig. 4. Example displays from Experiment 1 (left) and 2 (right) are shown. In Experiment 2 an additional display element was added labeled as "soil indicator," which indicated the current task state. In addition, the left panel shows how rewards (oxygen points) were conveyed in Experiment 2.

An example of the "soil indicator" is shown in Fig. 4, left. This indicator consisted of a black horizontal bar in which a brightly colored dot appeared. The position of the dot indicated the current task state (i.e., the value of h above). The entire scale ranged from 0 to 150.0 (ensuring that relative variability was comparable between the reward signal and soil indicator). The position of the dot in this scale was determined by the following equation on each trial: $\frac{h}{10} * 110 + 20 + N(\mu = 0, \sigma_s)$. In the NN and RN condition, σ_s was set to zero (no variability), but in the SN and SRN condition σ_s was set to 12.0. As for rewards, the state cue value was clipped at 0 and 150. Thus, the variability introduced in the state cue was equivalent to that on the reward condition.[1] The direction that the state light moved in response to increasing numbers of Long-term button presses was counter-balanced between subjects. So for some subjects the light moved to the left as h increased, and for others it moved to the left as h decreased.

2.1.3. Procedure

The procedure was identical to Experiment 1.

2.2. Results

2.2.1. Overall performance

Fig. 5A shows the proportion of responses allocated to the Long-Term option (i.e., maximizing responses) over the entire experiment for each condition. A two-way ANOVA on these proportions revealed no effect of noise on rewards ($F(1, 107) = 2.18$, $p = .14$), a main effect of noise on states ($F(1, 107) = 7.82$, $p = .006$), but no interaction ($F(1, 107) = 1.22$, $p = .27$). In all four conditions the proportion of Long-term selections exceeded chance ($t(26) = 10.7$, $t(26) = 7.12$, $t(26) = 8.68$, and $t(26) = 4.59$, for the NN, RN, SN, and SRN condition respectively, all $p < .001$). Planned comparison revealed that the proportion of maximizing responses was higher in the NN and SRN conditions, $t(52) = 3.12$, $p = .003$, and in the RN and SN condition, $t(52) = 2.65$, $p = .01$. However, all other contrasts were not significant. Identical results obtained considering performance only in the last block of 50 trials, $t(52) = 3.03$, $p = .004$ and $t(52) = 2.5$, $p = .02$, respectively.

Optimal responding in the task required pushing the state cue either all the way to the left or all the way to the right (depending on the counterbalancing of this display variable).[2] In the function learning literature this is analogous to learning either a positive or negative correlation between input–output pairs (where input here is the state cue and output is the reward). Given that previous work has shown that learning negative correlations is generally more difficulty (Busemeyer, Byun, Delosh, & McDaniel, 1997), we analyzed our data to assess the impact that the direction of the light had on performance. Collapsing across all four conditions, we failed to find a significant effect of the left-right directionality of the light ($t < 1$). However, there was a trend consistent with this idea ($M = .69$ when it was optimal for the light to move to the left, and $M = .75$ when it was optimal for the light to move to the right). Separate analyses conducted on individual conditions also failed to find a significant effect of the directionality of the light on responding (all $p > .07$). In addition, we failed to find an effect of compatibility between the movement of the light in response to reward-maximizing choices and the left-right arrangement of the maximizing response button on the display ($t < 1$).

2.2.2. Time-course analysis

Fig. 5B shows the proportion of Long-Term choices calculated in non-overlapping blocks of 25 trials at a time. Participants in all conditions increased the number of Long-Term choices they made over the course of the experiment. A 2 × 2 repeated measures ANOVA on noise on reward, noise on states cues, and training block (1-5) and condition revealed a main effect of noise on states ($F(1, 104) = 7.88$, $p = .006$), no effect of noise on reward ($F(1, 104) = 2.17$, $p = .14$), and no interaction ($F(1, 104) = 1.22$, $p = .27$). In addition, the interaction between block and state noise or reward noise failed to reach significance ($F(4, 416) = 2.32$, $p = .056$ and $F(4, 416) = 1.33$, $p = .26$, respectively). Finally, we found no evidence of a three-way way interaction between training block and noise type ($F(4, 416) < 1$). Planned comparisons confirmed that within each condition, there was a highly significant effect of training block on performance, demonstrating learning ($F(4, 104) = 35.14$, $F(4, 104) = 17.25$, $F(4, 104) = 49.97$, and $F(4, 104) = 9.59$, for the NN, RN, SN, and SRN conditions respectively, all $p < .001$). Likewise, comparing the proportion of selections allocated to the Long-Term option in the first block of 100 trials compared to the last block of 100 trials revealed a significant increase in all conditions (mean difference = .31, $t(26) = 8.31$ in the NN condition, mean difference = .30, $t(26) = 10.31$ in the RN condition, mean difference = .24, $t(26) = 5.33$ in the SN condition, and mean difference = .20, $t(26) = 4.41$ in the SRN condition, all $p < .001$). However, a 2x2 ANOVA on these difference scores found only a main effect of noise on state cues ($F(1, 104) = 4.49$, $p < .04$).

[1] Note that because the reward was sampled from either of the two response options, the perceived variability in rewards was likely higher than for states (the reward fluctuated more as a function of the participant's response strategy than did the state cue). However, as we will describe in our later simulations, this likely higher perceived variability for rewards makes the pattern of results we find even more persuasive).

[2] Note, however, that this was only possible in the NN and RN conditions due to the noise on the state cue in the SN and SRN conditions.

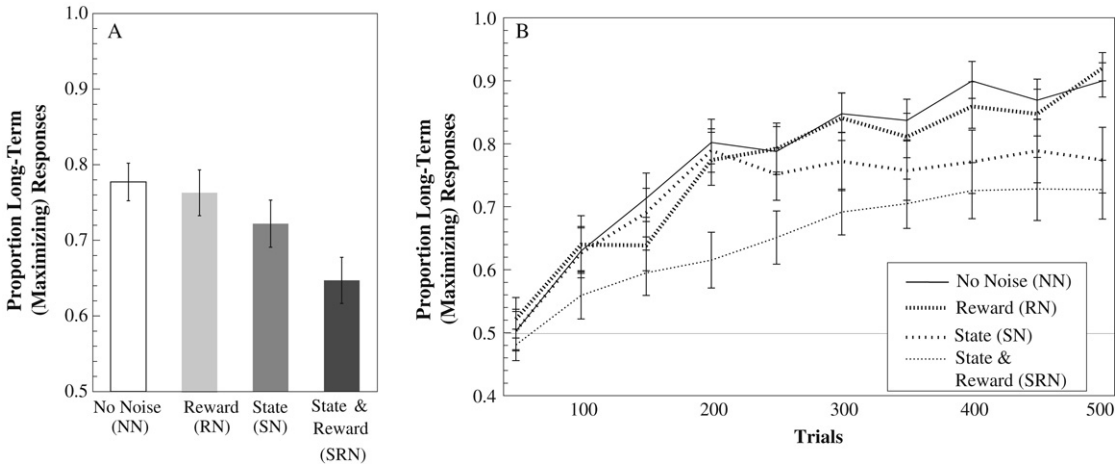

Fig. 5. Overall results of Experiment 2. Panel A shows the average proportion of long-term (maximizing) responses made throughout the experiment as a function of noise condition. Panel B presents the average proportion of maximizing responses considered in blocks of 50 trials for all three conditions. In panel B, the horizontal line at 0.5 shows chance performance. Error bars are standard errors of the mean.

2.3. Discussion

In Experiment 2, we compared the impact that different sources of variability had on participant's performance in the Farming on Mars task. Unlike Experiment 1, participants in this experiment were always presented with information in the display that indicated the current task state. However, we manipulated the degree of variability associated with the reward signal and state cue in order to assess the relative contribution of each source of information. Consistent with previous work (Gureckis & Love, in press), we found that people appear to use the state information to support their decision making. In fact, when the reliability of cues that signaled task state were degraded, performance dropped off more than it did with equivalent amounts of noise on the reward signal. This result is somewhat surprising given that attending to the state cues was, in some sense, optional for solving the primary task of maximizing reward. However, this result is consistent with the idea that perceptual cues that signal the current task state help to resolve perceptual aliasing. When such signals are obscured by noise, the ability of participants to differentiate highly similar states is reduced, leading to worse performance.

One difference between Experiment 1 and 2 is that moderate amounts of noise/variability on the reward signal did not improve performance in Experiment 2. One possibility is that the amount of noise used in Experiment 2 was simply insufficient to encourage additional exploration. However, an important difference between Experiments 1 and 2 is that participants in the RN condition of Experiment 2 also had a consistent state cue on the display. One hypothesis (evaluated in the following simulations) is that state cues not only limit perceptual aliasing, but can help to effectively "smooth-out" variation in the reward associated with particular states. For example, recognizing that one is in a familiar restaurant allows one to recall past experiences and arrive at a better estimate of the value of particular options than an individual who cannot benefit from cues that enable recognition.

To summarize, Experiments 1 and 2 provide three key findings. First, when participants lack explicit perceptual cues about task state, there is a non-montonic relationship between reward variability and task performance. In particularly, moderate amounts of reward noise lead to elevated levels of optimal responding relative to either no variability, or high variability situations (Experiment 1). Second, comparing the effects of matched variability on reward and state cues (Experiment 2) reveals that noise on state cues has a more detrimental effect on performance than does noise on rewards. Finally, we find that when participants have consistent information about task state, there is no performance advantage for moderate variability in the rewards signal (i.e., NN vs. RN in Experiment 2). In the following section, we evaluate these results using a simple model of sequential choice based on reinforcement learning principles (Sutton & Barto, 1998). We begin by describing the basic operation and principles of the model, then turn to the supportive simulations.

3. Model-based analysis

The goal of the modeling is to better understand the mechanisms supporting learning in Experiments 1 and 2. We assume that the goal of the agent is to achieve the most reward possible by exploring the environment and adapting its behavior in a trial-by-trial fashion. In particular, the model attempts to estimate the long-term value of selecting a particular action a in state s, a value we refer to as $Q(s, a)$. These "Q-values" in the model represent an estimate of the future reward the agent can expect to receive given that it selects action a in state s. Ideally, the agent would choose the action in each state which leads to the higher reward. However, these values must be learned from experience, and thus some exploration is warranted. In our task, there are only two actions available to the RL agent for each state, which correspond to selections of either the Short-Term or Long-Term robot. States are not directly observable, but are estimated given the perceptual cues presented in the experiment and response history.

The model's estimate of the Q-value of a particular action on trial t is calculated as a simple linear function of the current perceptual inputs:

$$Q(s_t, a_t) = e_a + \sum_{j=1}^{N} w_{ja}^t \cdot I_j^t \quad (1)$$

where N is the number of inputs, I_j^t is the value of the jth input cue on trial t, w_{ja} is a learned weight from the jth input unit to action a, and e_a is an eligibility trace for action a (described below). The model maps between input cues and the Q-values associated with each state using a simple single-layer network (Widrow & Hoff, 1960). Changing the types of input cues the model is given modulates the ability of the model to learn the appropriate representation of the state structure of the task and ultimately influences its ability to uncover an optimal response strategy.

The model learns the value of state-action pairs (and ultimately adopts a reward-maximizing strategy) through experience using a learning algorithm based on average reward reinforcement learning (Schwartz, 1993; Sutton & Barto, 1998; Tsitsiklis & Van Roy, 1999). Average reward RL assumes the goal of the learning agent is to maximize the average reward received per unit time step. This approach contrasts with models based on temporal-difference methods (such as Q-learning) which treat the agent's goal as optimizing the total discounted reward. In particular, according to discounted reward approaches, the value of each state-action pair, $Q(s, a)$ is given as the total expected discounted future reward:

$$Q(s,a) = E\left[\sum_{k=0}^{T} \gamma^k r_{t+k+1}\right] \tag{2}$$

where γ represents a discounting parameter that determines the weight given the rewards that are delayed into the future (Watkins, 1989). An alternative approach is to assume that the value of each state is the difference between the total reward available from the state-action pair and the average reward:

$$Q(s,a) = E\left[\sum_{k=0}^{\infty} (r_{t+k+1} - \rho)\right] \tag{3}$$

where ρ is an estimate of the average reward per time step encountered so far. The subtraction of the average reward ρ rescales the total cumulative reward available in each state as deviation from average and introduces transients that drive learning. Positive values of $Q(s, a)$ reflect state-action pairs that result in greater than the currently estimated average reward, while negative values are given to state-action pairs that result in lower than average reward. Also note that this approach no longer assumes the learners discount future rewards (i.e., there is no γ parameter). For this reason, average-reward RL algorithms are considered non-discounted.

Overall, discounted reward methods (such as temporal-difference learning or Q-learning) have been used more widely in cognitive psychology. However, there are a number of arguments for why average-reward RL may be a more realistic model of human behavior (Daw, 2003; Daw & Touretzky, 2000; Niv, 2007). For example, a number of authors have argued that the average-reward RL framework more accurately reflects the way in which animals and humans deal with delayed rewards (Daw & Touretzky, 2000; Kacelnik, 1995). In particular, traditional discounted methods assume that reward is exponentially discounted, while considerable work suggests that the discounting function is hyperbolic (Myerson & Green, 1995).[3] Given certain assumptions about the nature of discounting, average reward learning can provide framework similar to the hyperbolic discounting function often found with humans and animals (Kacelnik, 1995) and appears to better fit both behavioral and neuro-biological data (Daw & Touretzky, 2000). In addition, computational analyses suggest that average reward RL converges faster than traditional discounted approaches (Tadepalli & Ok, 1996), a feature which was desirable in our simulations given how quickly participants learn.

In average reward RL, each time an action is selected, the model updates its current estimate of the corresponding Q-value according to the temporal-difference error between the reward received as a result of that action and a current estimate of the future reward available from the following state-action pair according to

$$\delta = [r_{t+1} - \rho + \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{4}$$

where r_{t+1} is the reward received as a result of taking action a_t, ρ is the current estimate of the average reward per unit time step, and $\max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$ is a current estimate of the reward available in the new state, s_{t+1} which results from taking action a_t. This error term, δ, is then used to adjust weights in the model according to

$$w_{ja}^{t+1} = w_{ja}^{t} + \alpha \cdot \delta \cdot I_j^t \tag{5}$$

where w_{ja}^{t+1} is the new value of the weight, w_{ja}^{t} is the old value of the weight, α is a learning rate parameter, and I_j^t is the value of the jth perceptual input.

As show in Eq. (1) the model also assumes that subjects keep a simple memory for recent actions, known as an eligibility trace, and that this memory is used to assist predictions of reward (see Bogacz, McClure, Li, Cohen, and Montague (2007) and Gureckis and Love (in press) for a discussion of eligibility traces). The model stores two eligibility traces, e_a, one for each action. The activation of this additional input to each Q-value was updated on each trial according to

$$e_a = \alpha \cdot \delta \cdot \lambda_a. \tag{6}$$

On each trial, the λ_a for every action decays according to $\lambda_a = \lambda_a * \zeta$ with $0.0 \leq \zeta \leq 1.0$. However, each time a particular action a_i is selected, the trace for only that action is incremented according to $\lambda_i = \lambda_i + 1$ (Bogacz et al., 2007). Eligibility traces simply assert that participants remember which actions they have selected in the recent past, and that this information is used to help credit actions which lead to increased reward. Generally, eligibility traces improve the rate of learning by allowing the error generated on the current trial to propagate backwards in time across more than one previous action.

Finally, the model tracks an estimate of the average reward according to $\rho = \rho + \alpha_{avg} \cdot \delta$ when the currently selected state action pair $Q(s, a)$ is the maximum valued action available from the current state (i.e., the average reward is only updated when the best action is selected). The initial value for ρ was fit as a parameter (see below). The parameter α_{avg} is a learning rate controlling the rate of updating for the average reward.

The probability of selection action a_i on each trial is given by

$$P(a_i) = \frac{e^{Q(s_t, a_i)\cdot\tau}}{\sum_{j=1}^{2} e^{Q(s_t, a_j)\cdot\tau}} \tag{7}$$

where τ is a parameter which determines how closely the choice probabilities are biased in favor of the underlying Q-values (Luce, 1959). In general, the probability of choosing option a_i is an increasing function of the estimated value of that action in that state relative to the other action. However, as $\tau \rightarrow 0$ each option is chosen randomly (the impact of Q-values is reduced). Alternatively, as $\tau \rightarrow \infty$ the model will always select the highest valued option (also known as "greedy" action selection).

In summary, our simple RL model has four parameters: a learning rate for Q-values, α, a learning rate for the average rewards, α_{avg}, a parameter controlling exploratory actions, τ, and a decay parameter, ζ which controls the model's memory for recent actions. In addition, in our simulations we fit three additional parameters: Q_0 which is the initial value of each Q-values in the model, ρ_0 the initial estimate of the average reward, and w_0 the initial value of the learning weights. The addition of these parameters improved the quality of the fit while making explicit our assumptions about prior expectations that participants might bring to the task (in dynamic models such as our initializing weights and initial parameters to zero is not assumption-free). However, additional fits starting these parameters at zero did not change the qualitative results.

[3] See Fu and Anderson (2006) for an approach that explicitly assumes hyperbolic discounting in a discounted-reward paradigm but which cannot be updated iteratively.

3.1. Simulation method

In Experiments 1 and 2, we parametrically manipulated the nature of the task environment that participants experienced and measured the resulting change in behavior. Our simulations followed a similar approach. Rather than fit individual choice sequences using maximum likelihood, we instead investigated the behavior of the model across a wide range of parameters and task environments in order to derive both qualitative and quantitative predictions.[4]

In order to derive predictions, we applied the model to the task in a manner analogous to how participants experienced the task (e.g., the same number of trials). On each simulated trial, the model selected either the Short- or Long-Term robot (stochastically, according to the probabilities given in Eq. (7)), the rewards were delivered, and the current state of the Mars Farming System was updated. Each simulation consisted of 500 trials in the task. The per-block choice proportions from multiple simulations were averaged together. A single set of parameters was fit by minimizing root-mean squared error (RMSE) between the average model performance and human performance curves (averaged across subjects) on a per-block basis (i.e., per 100 trials). The best fit parameters and resulting RMSE for both experiments are show in Table 2.

In order to account for the role that the state cue played in Experiment 2 performance, the model was provided with a representation of the location of the state-cue akin to that given to participants in the task. In particular, a single input unit to the model was activated on each trial, I_j (see Eq. (1)). The value of this input unit coded the position of the "dot" on a scale of 0.0 to 1.0, where in some simulations 0.0 meant that the dot was all the way to the left of the screen, and in others this coded positions all the way to the right (counterbalanced). Noise was added to this cue in the same way as for human participants (i.e., the scale was divided into 150 units and updated according to both the current state of the system and the amount of noise).

In Experiment 1, where these cues were absent, all inputs were set to zero. As a result, the model relies entirely on eligibility traces in order to learn (i.e., since $I_j^t = 0$ for all j, the value of $Q(s_t, a_t)$ in Eq. (1) is simple e_a, Eq. (5) no longer is needed to update learning the predictive learning weights, and the look-ahead term in Eq. (4), $\max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$, was set to zero).[5]

3.2. Experiment 1 simulations

In Experiment 1, participants interacted with the Farming on Mars task in the presence of different amounts of noise on the reward signal. In order to derive model predictions for the effects of noise, we first found best-fit parameters by fitting the model to data from all four conditions tested in our experiment. Using these parameters, we then found the predicted performance of the models across an entire range of noise levels ($0 <= \sigma_r <= 650$), a relationship we refer to as the "noise-performance" function.[6] The results of this parametric analysis are shown in Fig. 6. Panel A compares the fit of the model to the human data for each condition tested. Panel B shows the entire response profile of the model across different levels of reward noise. Interestingly, the model shows near chance performance in the task when $\sigma_r = 0$, and performance gradually increases over the range of $0 < \sigma_r < 50$. After this point, additional variance in the reward signal begins to negatively impact performance such that for σ_r greater 200, performance hovers around 30% maximizing. Finally, as the noise continues to increase on the reward signal, performance begins to drift back to chance (i.e., 50%). Thus, the model, like human participants, predicts a non-monotonic relationships between variability in the reward signal and performance, and this relationship closely matches the one demonstrated by participants.

In order to verify that this relationship was not an artifact of a particular set of parameters, we conducted a number of restricted model fits. First, we found parameters in the model that best fit the data from the no-noise (NN) condition only. Using this parameter set we once again found the model's predicted response proportion across the full parametric range of the reward noise. These parameters are also reported in Table 2. A similar qualitative pattern to the one fit to the full data is shown in Fig. 6B. Finally, we conducted a new parameter fit using only the data from the low noise (LN) condition. Once again the model correctly predicts that performance in the NN condition should be lower than performance in the LN condition, and that a non-monotonic relationship holds where increasing variability results in increased performance. Note, however, that across all three sets of parameters, the qualitative shape of the response functions change. Thus, the specific advantage for increasing amounts of noise appears to be scaled by particular model parameters.

In order to understand the role that particular parameters played in driving model behavior, we conducted a number of parametric parameter analyses. In these simulations, we took the best fit parameters to the full data set found above and plotted the resulting noise-response function as a single parameter was varied over a wide range.[7] As shown in Fig. 7, we considered the effect that different values of each of the four individual parameters had on the noise-response function. The top-left panel of Fig. 7 shows that when the learning rate is close to zero (the horizontal dashed line at 50%), performance is at chance, irrespective of the level of noise. However, as the learning rate increases, the model begins to predict an advantage for moderate amounts of noise on performance. For very large learning rates ($\alpha = 1.0$, the solid line in the figure), the effect does not occur. A similar analysis on the decay rate for eligibility traces (ζ) shows a similar pattern. In this case, the model shows the greatest boost in performance for moderate values of ζ. However, when ζ is set to the upper end of the sampled range, the model no longer predicts a performance advantage for increased noise levels. Interestingly, with respect to this parameter set, the overall shape of the response curves is not greatly effected by the exploration parameter (τ) or the average-reward updating parameter (α_{avg}), except for the obvious boundary cases (e.g., $\tau \rightarrow 0$). The model fits best with relatively low settings of α_{avg}, suggesting an that average reward should be incrementally updated. Overall, the results indicate (across a wide range of model parameters) that the model predicts a non-monontonic relationship between the variability of the rewards and task performance.

[4] We conducted competitive maximum likelihood fits of our model against a number of baseline models. However, we did not find any evidence that the best fit parameters of the model to individual subjects reliably varied between conditions. This matches our expectation that the populations tested in our conditions were all the same, but only the task environment changed.

[5] A separate set of simulations found that the model shows no learning in Experiment 1 without eligibility traces included, but predicts the same qualitative effect in Experiment 2 where state cues were present. Effectively, eligibility traces provide the model with a history of recent action selections in Experiment 1 which proxy the state-disambiguation role that state cues play in Experiment 2 (see Gureckis and Love (in press) for a further discussion of the trade offs between state cues and eligibility traces).

[6] Values of σ_r were sampled at intervals of 25 units over this range in our figures.

[7] In some cases, we used a limited range of parameters that garanteed stable model performance. In addition, we only manipulate individual parameters rather than testing the interactions between parameters.

Table 2
Best-fit parameters and resulting RMSE for Experiments 1 and 2. Parameters are explained in the main text.

	RMSE	α	α_{avg}	τ	ζ	Q_0	ρ_0	w_0
Experiment 1								
All	.04	.44	7.3e-05	.04	.45	1676	625	–
Fit to NN Only	.05	.42	.001	.034	.4	1029	536	–
Fit to LN Only	.14	.34	.002	.036	.17	1298	227	–
Experiment 2								
All	.07	.41	.004	21.9	.12	2.03	.72	3.9

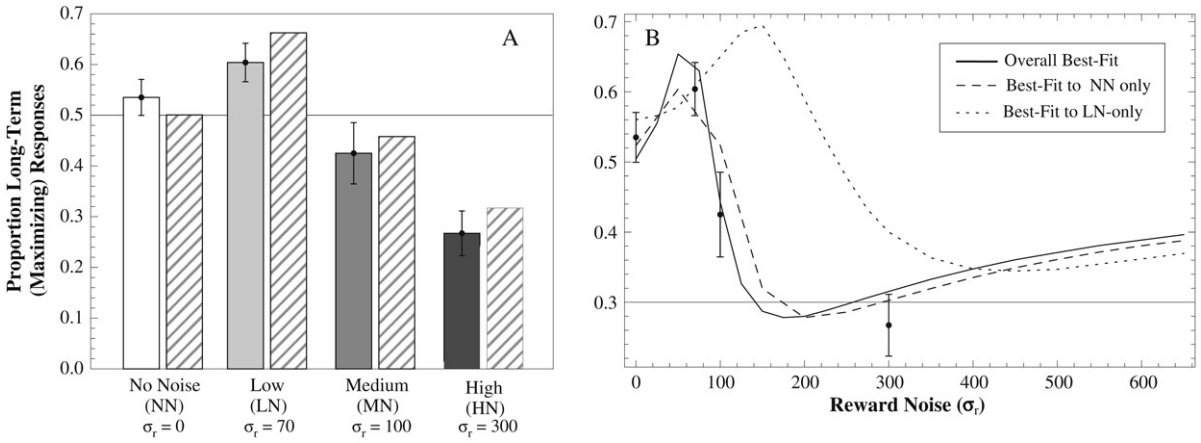

Fig. 6. A comparison of human performance in Experiment 1 with the RL model. Panel A plots the performance for each condition of Experiment 1 along with the predicted response proportions for the model (using the overall best fit parameters). The model, like human participants, shows a non-monotonic relationship between performance and reward noise. Panel B plots this noise-performance relationship over an entire range of σ_r for different parameter fitting procedures. Even though individual parameters change the shape of the response function, all the fits display a performance advantage for higher noise levels relative to no noise at all.

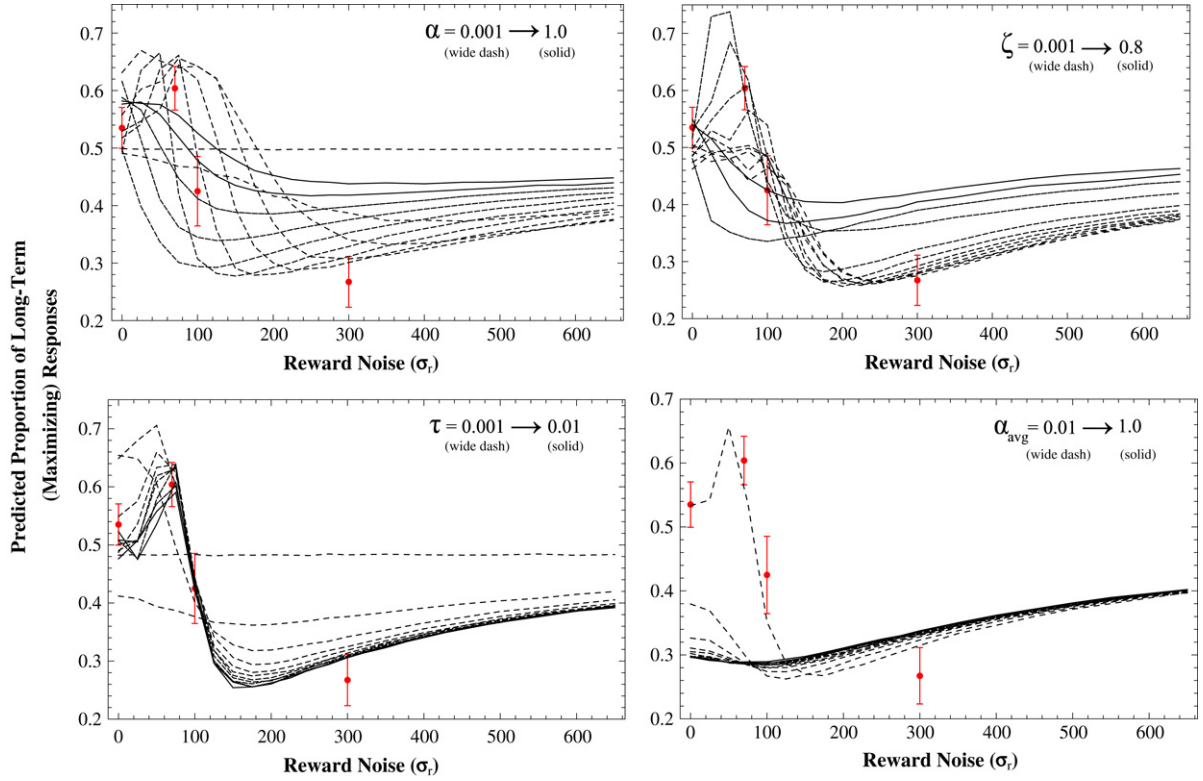

Fig. 7. Parametric model analysis showing the effect of changing different parameters in the model on the noise-response function for Experiment 1. The vertical axis shows predicted performance in the task (i.e., proportion of Long-Term maximizing responses). The horizontal axis shows the standard deviation of the noise (σ_r). The points plotted with standard error bars are the average performance of human subjects in Experiment 1. In these plots, each parameter was moved independently of the others starting with the best-fit set to all of the human data.

The performance advantage for intermediate reward noise levels is a straightforward consequence of the way Q-values are updated in the model. As the level of noise increases, the probability of encountering extreme observations of the value of

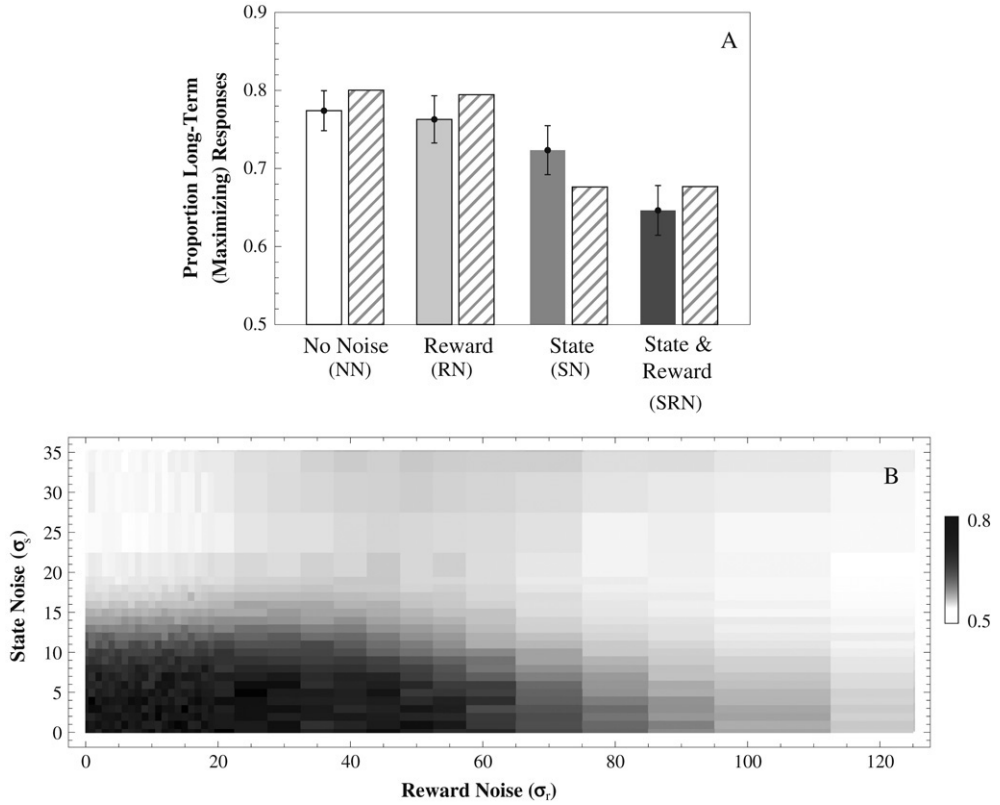

Fig. 8. A comparison of human performance in Experiment 2 with the RL model. Panel A plots the performance for each condition of Experiment 2 along with the predicted response proportions for the model using the best fit parameters. The model shows a main effect of increasing state-cue variability. Panel B plots this noise-performance relationship over a sampled range of both σ_r and σ_s. Dark areas this figure reflect regions of this "task-environment" space where the model predicts increased performance, while white areas show regions closer to chance. The key result is that performance drops off quickly for the σ_s compared to σ_r, consistent with the human results.

any action increases. These extreme values can cause a sudden shift in the estimated value of particular actions. As a result of the sudden fluctuations, the model may switch preferences from one action to the other. Thus, the transient dynamics of weight updates ultimately leads the model to more effectively explore the system. With continued learning, these trial-to-trial fluctuations are eventually "washed out" until the model settles on a effective strategy.

3.3. Experiment 2 simulations

Fig. 8A compares the model and human performance for each of the four conditions tested in Experiment 2. Best-fit parameters were found by fitting the model to the performance curves from all four conditions (see the bottom row of Table 2). Overall, the model captures the main effect that additional noise on the reward signal has little impact on performance, while additional noise on the state cue does. This is interesting given that the levels of variability introduced on each signal were matched (and, if anything, the "effective" variability on reward was higher due to the fact that changing response options introduced additional variability due to the two different reward curves for short- and long-term responses).

In addition, the model no longer shows a non-motonic relationship between performance and reward noise σ_r when σ_s is set to zero. Essentially, performance remains relatively constant across a large range of intermediate σ_r values and declines for higher values. The major difference between these simulations and those of Experiment 1 were the addition of the state cues. Thus, having cues which signal the current task state serve to reduce the impact that moderate amounts of noise have on performance. When these cues are absent (such as in Experiment 1), the model's behavior is predicted to be more sensitive to variations in reward.

Like our simulations for Experiment 1, we characterize the performance of the model for a wide range of task environments (i.e., values for σ_s and σ_r). Fig. 8, panel B shows a density plot that reveals how changing these two task parameters impact performance. Dark colors in this plot indicate cases in which the model predicts performance closer to 80% maximizing, whereas light colors indicate performance closer to chance (50%). Interestingly, increasing noise on the state cue (σ_s) has a much more dramatic effect on performance than does increasing reward noise (σ_r). For example, when $\sigma_r = 0$ for values of $\sigma_s > 15$, performance is close to chance, whereas when $\sigma_s = 0$, an σ_r of almost 80-100 is necessary to show the same level of performance. In addition, the plot provides scant evidence for strong interactions between state and reward noise, although it appears that for higher values of σ_r performance may drop off more quickly for large σ_s. However, these predicted effects occur outside of the range of σ_s and σ_r values tested in Experiment 2.

The prominent role state cues play in the model is quite intuitive. The model does not explicitly take into account the uncertainty associated with particular observations. As a result, all learning is with respect to the data/information that is currently available in the task display (i.e., the input cues). When the state cue fluctuates due to random noise, the model effectively updates the wrong state. This hampers the model's ability to learn the true value of each state. Returning to our restaurant analogy, it is as if the model is mistakenly learning about the quality of one restaurant while eating at another. Interestingly, the impact of reward variability is less pronounced. The model uses a simple linear network to approximate the entire Q-value space. Thus, noisy observations of reward are integrated across all states

simultaneously and variability of any one action is reduced. One key model prediction is that using different kinds of state cues, in particular cues which limit generalization between states, should result in performance being more impacted by reward noise.

As in our simulations of Experiment 1, we attempted to find other parameter sets by fitting the model to a restricted subset of the data. We found that a wide range of parameter values were able to capture performance in a single condition and therefore placed few constraints on the model's predictions for the other conditions. However, even in these simulations, we found that the model generally predicted the same qualitative pattern (i.e., state noise hurt performance more than equivalent amounts of reward noise).

4. General discussion

In this article, we used noise as a tool to elucidate the mechanisms that people use to learn in dynamic decision making tasks. Our results indicate how, in certain task environments, noise can have seemingly paradoxically positive effects on learning and decision-making. In particular, moderate noise can improve performance in some task environments by encouraging exploration (Experiment 1). Additionally, we found that different types of control signals are more or less resistant to the negative effects of noise. In particular, decision making became much more difficult when participants were mislead about the current task state due to noise, but was more robust in the face of similar types of noise on reward signals (Experiment 2).

In Experiment 1, we found that adding intermediate amounts of noise actually improved decision making performance. This finding is interesting given the often times negative relationship between noise and system performance. However, noise is not a pancea: the positive or negative effects of noise are highly tied to the structure of the task. For example, a number of recent studies have shown how increased exploration can lead to more effective decision making in tasks with complex, non-obvious solutions or tasks which require searching a large problem space for a globally optimal solution in a space of deceptive local minima (Maddox, Baldwin, & Markman, 2006; Worthy et al., 2007). In contrast, simpler tasks with more obvious solutions benefit more from rapid exploitation of early strategies (i.e., when the the optimal problem solution is not hidden in a space of sub-optimal strategies, the best outcome would be to exploit this solution early). In these studies, people's tendency to explore was manipulated by changing the motivational framing of the task, however in our experiments noise added to the task was simply part of the task environment. Thus, our results and simulations highlight how even effectively random, non goal-oriented exploration in response to contingencies in the environment can improve performance.

The fact that noise was beneficial in Experiment 1 (where only moderate amounts of noise obscured rewards) and had either no effect or negative consequences in Experiment 2 reinforces the idea that noise or variability can have a range of effects on performance. For example, while we showed that noise could improve performance, the positive effects of reward noise disappear when state cues are present in the task. Given that previous research has shown that state-cues can act to boost task performance (Gureckis & Love, in press), it appears that the positive effect of both increased exploration and better integrated structures do not simply combine. One possible reason is that these two decision-making aids likely influence different aspects of performance. Increased reward variability in the absence of a state cue may encourage exploration which may be critical when people have little insight into the structure of the task. In contrast, the addition of state cues likely gives people information about the structure of the task and encourages exploration in the absence of reward noise. For example, participants may wonder how to move the state cue light from one location to another and, in doing so, gain more information about the reward structure of the task.

The observation that there was no longer a benefit to increased reward noise when participants were given state cues is consistent with the idea that such cues help participants adopt a more robust conceptual "framework" for the task, helping them become more resistant to variability in any particular observation. One might draw an analogy to work in social psychology on the fundamental attribution error (Jones & Harris, 1967) or correspondence bias (Gilbert & Malone, 1995). According to these accounts, observers often attribute to an actor's personality aspects of their behavior that are largely the results of the particular situation the actor is in (for example, believing that a criminal is an intrinsically bad person irrespective of the situation that led them to commit the crime). However, such effects are likely diminished for people who we know better and for whom we may have developed a better model or representation. For example, if you know your friend is a nice person, you don't give up on them given only one time that they let you down (pushing the attribution out to the situation). To the degree that people themselves act as "state-cues" for integrating observations, it makes sense that such representations should help reduce the impact that particular (extreme) observations have on estimates of personality. In effect, the ability to separate signal and noise improves when participants can actually learn a better estimate of the process generating the signal.

Overall, this paper represents a first step toward understanding how task variability can be used as a tool to understand basic learning mechanisms. As mentioned in the introduction, experimenter controlled-noise is often used to reveal the nature of human perceptual and cognitive proceses (Gold et al., 2004; Green & Swets, 1966; Lu & Dosher, 1999; Pelli & Farell, 1999). Likewise, we believe there is considerable potential for using variability to improve our understanding of how people learn adaptive decision making strategies in complex tasks. For example, in our simulations, we assumed that people treated the noise induced in the task more or less veridically. In our model, a strongly deviant outcome was weighted the same as one that fell closer to the mean (and in fact, this fact allowed the noise to have a positive effect in Experiment 1). However, other approaches to adaptive learning explicitly represent and learn the current uncertainty associated with actions, such as the Kalman filter (Daw, O'Doherty, Seymour, Dayan, & Dolan, 2006; Kalman, 1960; Kruschke, 2008) or methods based on partially-observable markov decision processes (POMDPs, Littman, in this issue). The latter approaches suggest that learners should adjust their learning rate on a trial-by-trial basis with respect to the experienced variability in outcomes. While such strategies may be adaptive in certain cases, we show here that the relationship between noise and performance is potentially complex and may vary as a function of the task environment. Manipulations of task variability of the kind considered here are thus likely to give critical insight into how prior beliefs are updated in noisy environments and how the cognitive systems remains robust to such noise.

Acknowledgments

This work was supported in part by NIH-NIMH Math Modeling Post-Doctoral Training Grant T32 MH019879-12 to T.M. Gureckis and AFOSR grant FA9550-04-1-0226, and NSF CAREER grant 0349101 to B.C. Love. Additional data collection and writing were supported by startup funds provided by New York University to T.M. Gureckis. We thank Nathaniel Daw, Yael Niv, A. Ross Otto, and Lisa Zaval for helpful conversations in the development of this work. Author contributions: TG and BL designed research, TG collected and analyzed data, implemented and tested models, and wrote the paper.

References

Barron, G., & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*, 215–233.

Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology*, *79*, 251–272.

Bogacz, R., McClure, S., Li, J., Cohen, J., & Montague, P. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, *1153*, 111–121.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica*, 211–241.

Burns, B. (2004). Heuristics as beliefs and as behaviors: The adaptiveness of the hot hand. *Cognitive Psychology*, *48*, 295–311.

Busemeyer, J. (2002). Dynamic decision making. In N. J. Smelser, & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences*: *Vol. 6* (pp. 3903–3908). Oxford: Elsevier Press.

Busemeyer, J., Byun, E., Delosh, E., & McDaniel, M. (1997). Learning functional relations based on experience with input–output pairs by humans and artificial neural networks. In K. Lamberts, & D. R. Shanks (Eds.), *Studies in cognition series*, *Knowledge, concepts, and categories* (pp. 405–435). Psychology Press.

Busemeyer, J., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, *121*(2), 177–194.

Busemeyer, J., & Pleskac, T. Theoretical tools for understanding and aiding dynamic decision making. *Journal of Mathematical Psychology*, in this issue (doi:10.1016/j.jmp.2008.12.007).

Daw, N. (2003). Reinforcement learning models of the dopamine system and their behavioral implications. *Unpublished doctoral dissertation*. Carnegie Mellon.

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

Daw, N., & Touretzky, D. (2000). Behavioral considerations suggest and average reward TD model of the dopamine system. *Neurocomputing*, *32–33*, 679–684.

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

Edwards, W. (1962). Dynamic decision making and probabilistic information processing. *Human Factors*, *4*, 59–73.

Fu, W., & Anderson, J. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, *135*(2), 184–206.

Gilbert, D., & Malone, P. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38.

Gold, J., Sekuler, A., & Bennett, P. (2004). Characterizing perceptual learning with external noise. *Cognitive Science*, *28*(2), 167–207.

Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York: John Wiley and Sons, Inc.

Gureckis, T., & Love, B.C. Short term gains, long term pains: How cues about state aid learning in dynamic environments. *Cognition* (in press).

Herrnstein, R. (1991). Experiments on stable suboptimality in individual behavior. *The American Economic Review*, *81*(2), 360–364.

Herrnstein, R., & Prelec, D. (1991). Melioration: A theory of distributed choice. *The Journal of Economic Perspectives*, *5*(3), 137–156.

Jones, E., & Harris, V. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*, 1–24.

Kacelnik, A. (1995). Normative and descriptive models of decision making: Time discounting and risk sensitivity. In J. Houk, J. Davis, & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 249–270). MIT Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*(1), 35–45.

Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1983). Optimization by simulated annealgin. *Science*, *220*(4598), 671–680.

Kruschke, J. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning and Behavior*, *36*(3), 210–226.

Littman, M.L. A tutorial on partially observable Markov decision processes. Journal of Mathematical Psychology, in this issue (doi:10.1016/j.jmp.2009.01.005).

Lu, Z., & Dosher, B. (1999). Characterizing human perceptual inefficiencies with equivalent internal noise. *Journal of the Optical Society of America A: Optics, Image Science, and Vision*, *16*(3), 764–778.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Westport, CT: Greenwood Press.

Maddox, W., Baldwin, G., & Markman, A. (2006). A test of the regulatory fit hypothesis in perceptual classification learning. *Memory & Cognition*, *34*(7), 1371–1397.

Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior*, *64*, 263–276.

Neth, H., Sims, C., & Gray, W. (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*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Niv, Y. (2007). *The effects of motivation on habitual instruental behavior. Unpublished doctoral dissertation*. Hebrew University of Jerusalem.

Pelli, D., & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America A*, *16*, 647–653.

Redish, A., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addition, relapse, and problem gamling. *Psychological Review*, *114*(3), 784–805.

Schwartz, A. (1993). A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the 10th international conference on machine learning*. San Mateo, CA: Morgan-Kaufmann.

Skinner, B. F. (1948). 'Superstition' in the pigeon. *Journal of Experimental Psychology*, *38*(2), 168–172.

Stanley, W., Mathew, R., Russ, R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction, and practice in a simulated process control task. *Quartley Journal of Experimental Psychology*, *41A*(3), 553–577.

Sun, R., Slusarz, P., & Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, *112*(1), 159–192.

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

Tadepalli, P., & Ok, D. (1996). Model-based average reward reinforcement learning. *Artificial Intelligence*, 881–887.

Tsitsiklis, J., & Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica*, *35*, 319–349.

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

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *2*, 105–110.

Veksler, V., Gray, W., & Schoelles, M. (2007). Categorization and reinforcement learning: State identification in reinforcement learning and network reinforcement learning. In *Proceedings of the 29th annual conference of the cognitive science society*.

Watkins, C. (1989). Learning from delayed rewards. *Unpublished doctoral dissertation*. Cambridge University, Cambridge, England.

Whitehead, S., & Ballard, D. (1991). Learning to perceive and act by trial and error. *Machine Learning*, *7*(1), 45–83.

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention Record*, *4*, 96–104.

Worthy, D., Maddox, W., & Markman, A. (2007). Regulatory fit effects in a choice task. *Psychonomic Bulletin and Review*.