Psychonomic Bulletin & Review
2009, ?? (?), ???-???
doi:10.3758/PBR.

Navigating through abstract decision spaces: Evaluating the role of state generalization in a dynamic decision-making task

A. Ross Otto
University of Texas, Austin, Texas

Todd M. Gureckis
New York University, New York, New York

and

Arthur B. Markman and Bradley C. Love
University of Texas, Austin, Texas

Research on dynamic decision-making tasks, in which the payoffs associated with each choice vary with participants' recent choice history, shows that humans have difficulty making long-term optimal choices in the presence of attractive immediate rewards. However, a number of recent studies have shown that simple cues providing information about the underlying state of the task environment may facilitate optimal responding. In this study, we examined the mechanism by which this state knowledge influences choice behavior. We examined the possibility that participants use state information in conjunction with changing payoffs to extrapolate payoffs in future states. We found support for this hypothesis in an experiment in which generalizations based on this state information worked to the benefit or detriment of task performance, depending on the task's payoff structure.

In many real-world situations, short-term rewards conflict with long-term benefits. Consider the case of global warming, for which a group's difficulty in changing its behavior reflects a considerable difference in immediate payoffs between long-term beneficial and long-term detrimental actions. That is, the long-term detrimental action (unrestricted pollution) results in greater immediate reward (higher industrial output, greater comfort) than does the long-term beneficial option, which involves receiving smaller immediate rewards (lower industrial output, reduced comfort) but contributing to a larger overall pattern of reward in the long term (greater overall quality of life). Effective decision making in real-world situations involves not only weighing the costs and benefits of particular actions, but also understanding how actions in the past influence the costs and benefits of future actions. It should be noted that this class of problem differs from those used in studies of delay discounting in humans (e.g., Myerson & Green, 1995), in which single decisions are made on the basis of explicit instructions and it is made clear at what point in time the larger, delayed rewards will be received (Rachlin, 1995).

In this article, we examine how information about the state of the world affects decision making in dynamic tasks that require valuing either long-term or short-term rewards. Unlike static and one-shot decision-making problems in which the payoff contingencies are not influenced by participants' behavior (e.g., Ido & Barron, 2005), in our task, the possible payoffs associated with each choice change as a function of participants' recent choices. Thus, participants' behavior in the task can effectively influence the state of the task environment, which, in turn, has consequences for future rewards. In our experiments, we manipulated the information that people had about the current task state in order to study the relationship between their mental representation of the task and their ability to adopt effective decision strategies.

For example, in Figure 1A, each curve represents the payoff from one of two choice options in a repeated-choice task. The horizontal axis represents the current *state* of the task environment, and the vertical axis represents the payoff from selecting either choice. In all states, one option (which we call *long-term decreasing* [LT-D]) always yields a higher payoff than does the other option (called *long-term increasing* [LT-I]). Note that the current task state is defined as the number of LT-I choices made over the last 10 trials. Increases in the proportion of LT-I selections in one's history shift the current state of the system rightward on the horizontal axis (increasing the payoffs for both choices), whereas increases in the proportion of LT-D selections move the state leftward (decreasing the payoffs for both choices). Thus, options that lead to

A. R. Otto, rotto@mail.utexas.edu

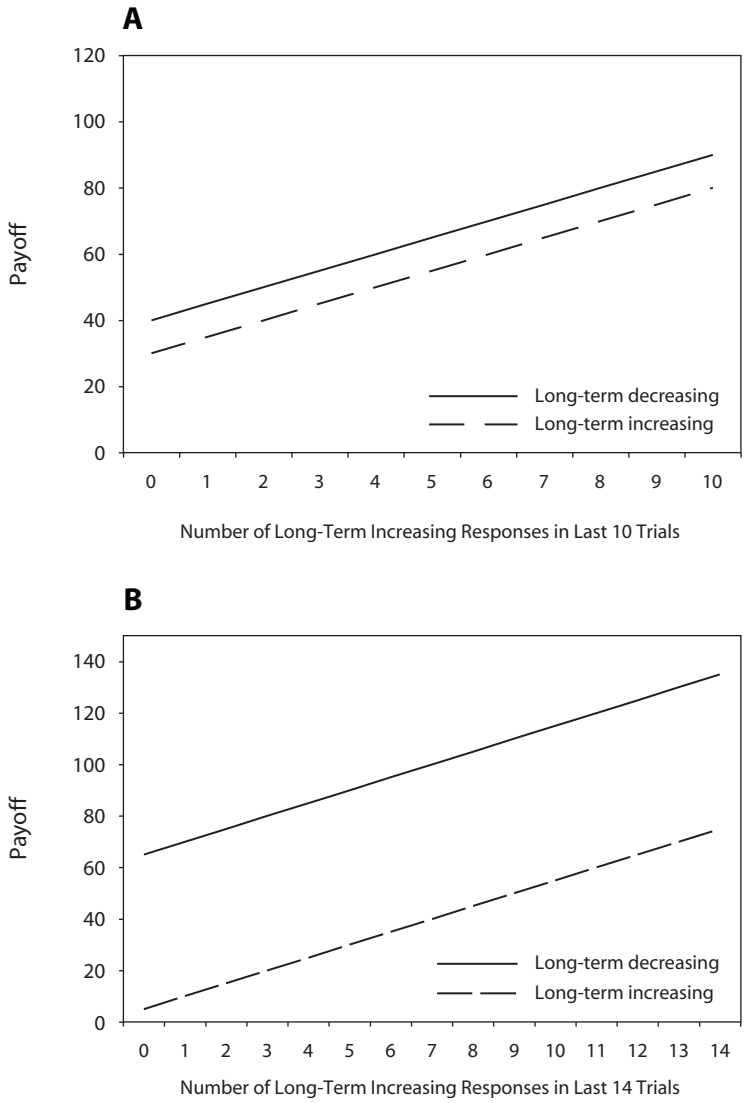

Figure 1. (A) Payoff functions for two choices as a function of response allocations over the previous 10 trials: Payoff functions in the close together payoff curves condition (similar to the one used in Gureckis & Love, in press). Of particular interest is the fact that the highest point of the long-term increasing (LT-I) curve is higher than the lowest point of the long-term decreasing (LT-D) curve. Thus, the optimal strategy is to choose the LT-I option on every trial. (B) In contrast, consider the payoff structure when the LT-D choice always generates higher immediate payoffs than does the LT-I choice but the global minimum of the LT-D payoff curve is greater than the global maximum of the LT-I payoff curve. Among pure response strategies, consistently choosing LT-D is the optimal pattern of response. Critically, optimal behavior in the two payoff structures depicted in panels A (i.e., *close together*) and B (i.e., *far apart*) requires different patterns of choices. The two reward curves were tested in the experiment.

larger immediate payoffs negatively affect future payoffs, whereas options that are less immediately attractive lead to larger future payoffs.

Consider a participant who has made only LT-D choices for 10 trials in a row, effectively making the task state 0. The payoffs from the LT-I and LT-D choices would be 30 and 40, respectively. If he or she makes one LT-I choice at this point, the task state would change to 1, since only 1 out of 10 of the last trials were LT-I choices. Consequently, the LT-I and LT-D choices would result in payoffs of 45 and 35, respectively. The payoffs associated with the choices fluctuate with the recent choice behavior of the participant.

In this dynamic payoff structure, a payoff-maximizing response pattern requires forgoing the LT-D choice and continually making LT-I choices (because the equilibrium point for the LT-I option is higher than that for repeated selections of the LT-D option). However, this strategy is not apparent to participants at the outset and must be learned

through experience. Prior research using similar payoff structures suggests that under certain task environment conditions, people eventually learn the optimal reward-maximizing response strategy (Herrnstein, Loewenstein, Prelec, & Vaughn, 1993; Tunney & Shanks, 2002). One question of interest in the literature concerns the sort of information that facilitates globally optimal responding in these tasks (Neth, Sims, & Gray, 2006).

Gureckis and Love (in press) pointed out that one challenge participants face in this class of tasks is forming an appropriate representation of the state of the task environment. Each time a participant makes a selection, the environment state can change so that the payoff for the chosen option changes on the next trial. Thus, it is not transparent to participants whether the task environment itself is changing or whether choice payoffs are simply fluctuating over time. In the standard version of this task (e.g., Herrnstein et al., 1993, Experiment 3; Tunney & Shanks, 2002, Experiment 2), it is difficult to recognize these changes without information that specifies the current state of the environment. One may view the problem as one of *perceptual aliasing*, wherein the decision-maker confounds environment states that it must distinguish in order to solve the task (Whitehead & Ballard, 1991).

Perceptual aliasing is a common problem that arises in spatial navigation tasks in which observations often fail to differentiate between multiple locations that an agent may actually occupy (Stankiewicz, Legge, Mansfield, & Schlicht, 2006). Historically, this literature has emphasized the importance of *landmarks*, defined as salient contextual cues associated with particular states or locations in the environment that serve as anchors or reference points to guide decision making and planning (O'Keefe & Nadel, 1978). There is evidence that landmarks play a prominent role in navigation performance for both humans and animals in spatial tasks (Cartwright & Collett, 1982; Siegel & White, 1975), but little research has been done to examine the role of landmarks in other types of sequential decision spaces. In the class of dynamic decision-making tasks considered in this article, the states are more abstract, representing the individual's recent choice history, as opposed to concrete spatial locations within the environment. Landmark-type information may bear influence on decision-making performance in these tasks.

Gureckis and Love (in press) found that an accurate representation of the task environment state facilitated optimal decision-making behavior, consistent with reinforcement-learning (RL) models (Sutton & Barto, 1998). Key to their behavior manipulations was the presentation of perceptual state cues (akin to visually salient landmarks) that indicated the current task state. In Gureckis and Love's study, participants saw an array of 11 lights arranged horizontally across the screen, only 1 of which was active. The location of the active light perfectly correlated with the current state of the task environment (i.e., the number of LT-I responses made over the last 10 trials). This manipulation is similar to that employed by Herrnstein et al. (1993, Experiment 1), which showed a marginally beneficial effect on participants' ability to maximize long-term payoffs when a cue was provided indicating the task state. However, in Gureckis and Love's study, the impact of two different types of perceptual cues was evaluated: In one condition, the currently active cue moved unidirectionally with the changing state from one end of the cue array to the other, whereas in another condition, the cue positions mapped randomly to different task states. Although both types of cues made clear that each choice changed the current task state—thereby helping participants overcome the problem of perceptual aliasing—the authors found that only unidirectional cues significantly improved participants' ability to make long-term payoff-maximizing responses, as compared with participants who did not have any disambiguating perceptual information.

Gureckis and Love (in press) suggested that observing the covariance between changing payoffs and the systematic movement of the state cue led participants to generalize experience acquired in one state to other, not-yet-experienced states, akin to extrapolating the slopes of the payoff curves. That is, participants who successfully learned that the payoffs were greater in State 2 (i.e., the number of LT-I responses over the last 10 trials was two) than in State 1 might have extrapolated this relationship to predict greater payoffs in States 9 and 10. Consequently, these participants were better able to move systematically through the decision space, adopting near-optimal response patterns. Simulations of an RL model—utilizing a simple linear network to estimate action values for each choice—revealed that supplying the model with consistent state information afforded extrapolation of rewards in unexperienced states and, subsequently, improved performance. Furthermore, this model provided the best account of participants' behavior, in comparison with other contemporary RL models (Bogacz, McClure, Li, Cohen, & Montague, 2007; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006). The authors concluded that local state information helped participants adopt optimal response patterns by reducing the perceptual aliasing inherent in abstract decision spaces.

One possible explanation of Gureckis and Love's (in press) findings is that participants provided with cues will engage in more systematic exploration of the state space and realize that there are two *fixed points* associated with the end points of the cue arrangement, where the payoffs no longer change. Through explicit comparison of the payoffs associated with these end point landmarks (namely, the minimum of LT-D and the maximum of LT-I, which are located at States 0 and 10, respectively), participants will settle on a strategy of consistent LT-I choices, since the maximum of the LT-I option yields greater payoffs than does the minimum of the LT-D curve. Alternatively, Gureckis and Love's model predicts that participants will extrapolate the gradient of the payoff function, performing actions that guide them "upward" to the optimal state. In other words, participants generalize from the local signal of rising payoffs, leading them to make repeated LT-I choices until they reach the global maximum of the LT-I payoff function. This local strategy is effective because the optimal strategy in the task requires repeated LT-I choice.

The *fixed points* and *generalization* accounts can be dissociated. If the payoff curves are spaced farther apart, so that LT-I is no longer a globally optimal choice (as shown in Figure 1B), participants following the payoff gradient should actually do worse overall when provided with cues, because chasing the rising rewards is not the globally maximizing strategy.

The experiment reported here elucidates the candidate mechanisms—specifically, the fixed points view and the generalization view assumed by Gureckis and Love (in press)—responsible for human choice patterns in dynamic decision-making environments. Using a task environment similar to that in Gureckis and Love, we manipulated the properties of the payoff curves to create situations in which linear extrapolation from one state to the next either did or did not lead participants to make consistently optimal choices. To foreshadow, our results suggest that participants indeed utilize payoff estimates obtained from generalization, a strategy that can lead to suboptimal performance, depending on the reward structure.

In the present experiment, we examined how participants use consistent state information to guide their choices. The experimental procedure was similar to the one used by Gureckis and Love (in press; Experiment 2). Payoff curves varied between participants. In one case, the curves for the LT-D and LT-I responses were placed close together so that the optimal strategy was to always choose the LT-I option (see Figure 1A). In the other, we spaced the curves far apart (without changing their slope), so that the optimal strategy was to actually choose the LT-D option (Figure 1B). By comparing the choice allocations of participants for these two payoff structures with and without these landmark-like cues, we can determine whether generalization about payoff function gradients is indeed the mechanism by which these cues drive choice behavior.

If consistent state cues assist participants with systematic exploration of the decision space according to the fixed points view, state cues should lead participants to make repeated LT-I choices in the *close together* payoff structure (Figure 1A) and repeated LT-D choices in the *far apart* payoff structure (Figure 1B). In other words, since comparison of payoffs at fixed points reveals the optimal long-term response strategy to participants, the fixed points hypothesis predicts that participants with state cues will make significantly more optimal decisions than will those without state cues in both payoff structures.

On the other hand, if choices are driven by generalization from unidirectional cue movement and payoff gradients, then "following" the positive slope of the LT-I curve should lead participants with state cues to make repeated LT-I choices in both payoff structures. That strategy is optimal for the close together payoff structure. In contrast, this strategy is suboptimal in the far apart payoff structure, where the minimum payoff of the LT-D choice exceeds the maximum of the LT-I choice. According to this view, participants with state cues should make a greater proportion of LT-I choices in both payoff structure conditions, because the state information will promote generalization/extrapolation about payoff curve slopes. Crucially, the generalization view predicts that participants will perform *less* optimally with cues than without cues in the far apart payoff structure.

METHOD

Participants

A total of 104 undergraduates at the University of Texas at Austin participated in this experiment for course credit plus a small cash bonus tied to performance on the task. The participants were randomly assigned to one of four conditions that varied in both the payoff structure (close together or far apart) and the presence or absence of state cues (cues vs. no cues). Twenty-six participants were assigned to each condition.

Materials

The experimental stimuli and instructions were displayed on 17-in. monitors. The participants read a cover story about extracting oxygen from the Mars atmosphere and were told that their goal was to maximize overall long-term extraction by pressing one of two buttons on each trial, corresponding to two systems for oxygen extraction. The participants were informed that the specific oxygen-extracting properties of the two systems were unknown but that they should learn the best strategy. It was also explained that each decision could temporarily change the quality of the atmosphere.

Procedure

The experiment consisted of 500 trials. At the start of the experiment, the number of LT-I responses over the last 10 trials (i.e., the state) was initialized to five. On each trial, the participants were presented with a control panel with two buttons labeled "Robot 1" and "Robot 2," shown in Figure 2. Using the mouse, the participants clicked one of the buttons to indicate their choice, after which a chirping sound was played and the payoff was presented visually, using an 11×11 grid of blue dots. The number of visible dots indicated the number of oxygen points extracted on that trial, also shown in Figure 2.

As was described above, in the close together payoff structure condition (Figure 1A), the number of oxygen points generated for selections of the LT-I option was $30 + 50 * (h/10)$, whereas the payoff for selecting the LT-D option was $40 + 50 * (h/10)$, where h in both equations represents the number of LT-I choices made by the participant over the last 10 trials. In contrast, in the far apart reward curves shown in Figure 1B, the payoff for LT-I was defined by $5 + 50 * (h/10)$ and the payoff for LT-D was $65 + 50 * (h/10)$.

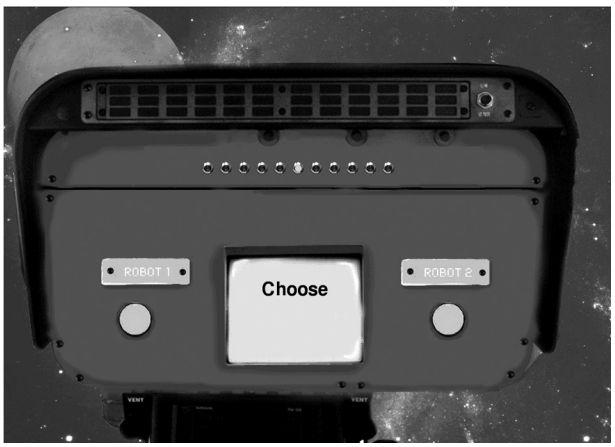

Figure 2. Screenshot of "Farming on Mars" experiment in the cues condition with an example immediate trial payoff.

The participants in the cues condition saw a display that included a row of 11 dots, 1 for each possible state, above the two choice buttons, as shown in Figure 2. The position of the active dot indicated the current state *h*, ranging from 0 to 10 (akin to its position on the horizontal axis in Figures 1A and 1B). Each time the participant made a choice, the position of the active dot was updated to reflect the state of the task environment. The presence and function of these state cues were not mentioned in the task instructions. This row of dots was not present for the participants in the no-cues condition.

The mapping of response buttons to choices and the direction of cue movement (leftward or rightward as *h* increased) were counterbalanced across participants. At the end of the trials, the participants were paid a cash bonus commensurate with their cumulative payoffs.

RESULTS

The main dependent measure was the proportion of trials for which participants made LT-I responses, depicted in Figure 3A. In short, our results favor Gureckis and Love's (in press) generalization view over the fixed points view. A 2 (payoff structure) × 2 (presence of state cues) ANOVA revealed a main effect of cue presence [$F(1,102) = 20.03$, $p < .001$], a main effect of distance between payoff curves [$F(1,102) = 12.53$, $p < .001$], and a significant interaction [$F(1,102) = 3.97$, $p < .05$]. Among the close together payoff curve groups, the participants with state cues ($M = .72$, $SD = .054$) made significantly more optimal LT-I responses than did the participants without state cues ($M = .57$, $SD = .043$) [$t(50) = 2.12$, $p < .05$]. In the far apart payoff curve condition, those with state cues ($M = .62$, $SD = .057$) made significantly more suboptimal LT-I responses than did those without state cues ($M = .31$, $SD = .051$) [$t(50) = 4.11$, $p < .01$]. Among participants with state cues, the participants' response proportions were not significantly different between the far apart and close together payoff curve conditions [$t(50) = 1.25$, $p = .216$].

Optimality of responding was measured by calculating the proportion of each participant's cumulative payoff to the maximum possible cumulative payoff under an optimal pure response strategy (strictly LT-I responses in the close together condition and strictly LT-D responses in the far apart condition). These average cumulative payoff proportions are depicted in Figure 3B. A 2 (distance between reward curves) × 2 (presence of state cues) ANOVA on this measure revealed a significant interaction [$F(1,102) = 11.69$, $p < .001$] and a main effect of distance between reward curves [$F(1,102) = 33.39$, $p < .001$], with no significant main effect of state cue presence. Within the close together payoff curve condition, the participants with state cues ($M = .852$, $SD = .027$) garnered significantly more total reward than did the participants without state cues ($M = .779$ $SD = .021$) [$t(50) = 2.14$, $p < .05$]. Among the participants in the far apart payoff curve condition, the participants with state cues ($M = .895$, $SD = .010$) earned significantly less than did the participants without state cues ($M = .947$, $SD = .008$) [$t(50) = -4.102$, $p < .001$]. As predicted by the generalization view, the participants in the far apart payoff structure responded less optimally with cues than without cues.

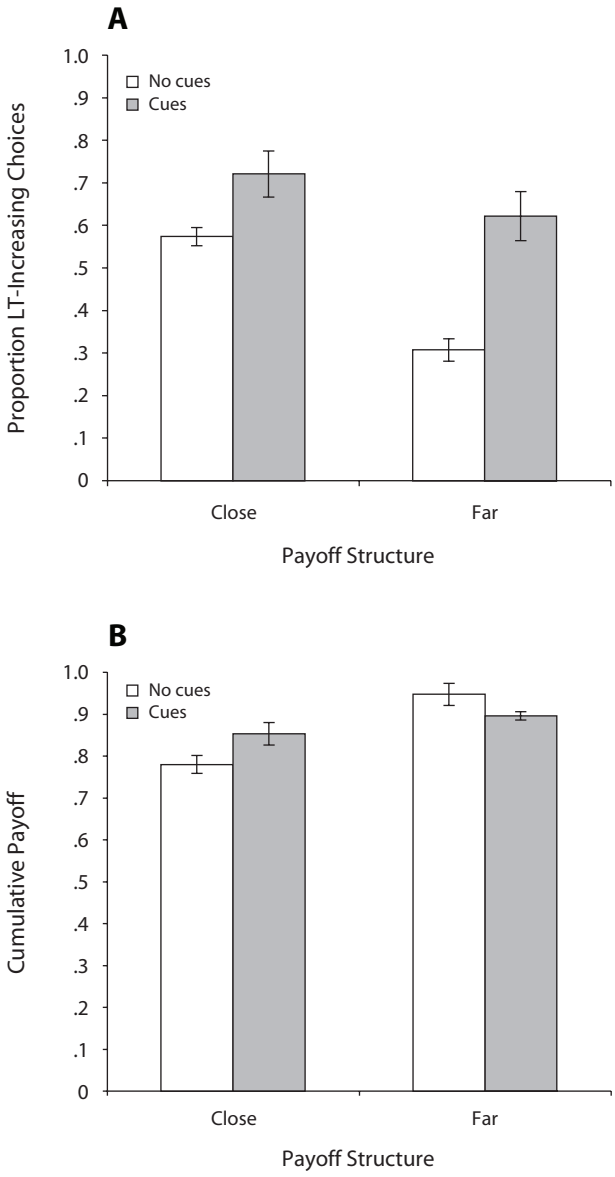

Figure 3. Overall results of Experiment 1. (A) Average overall proportion of long-term increasing (LT-I) choices as a function of condition. (B) Average proportion of cumulative payoff to maximum possible payoffs as a function of condition. Error bars are standard errors of the means.

In addition, we evaluated the extent to which state cues facilitated systematic exploration of the state space. Whereas the fixed points view predicts that participants with state cues should visit both fixed points (namely, the minimum and maximum of the LT-D and LT-I payoff curves, respectively) and remain there long enough to observe that the payoffs stop changing, the generalization view predicts that state cues will simply drive participants toward the maximum of the LT-I curve. As a proxy, we considered a fixed point as *visited* if a participant had spent at least 10 consecutive trials in that state. The proportion of participants in each condition who visited both

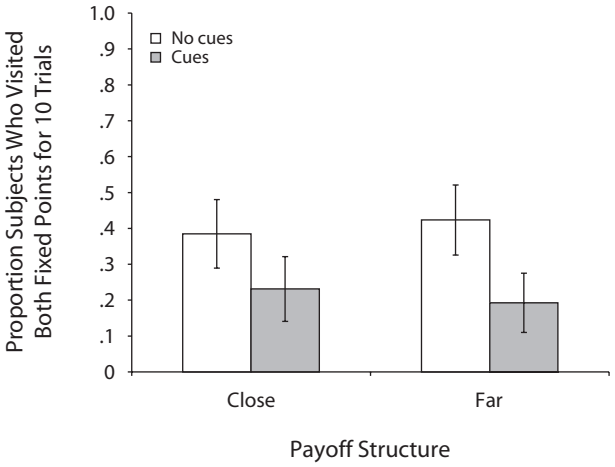

Figure 4. Proportions of participants who spent 10 or more consecutive trials at both the minimum of the long-term decreasing (LT-D) payoff curve and the maximum of the long-term increasing (LT-I) payoff curve. Error bars represent standard errors of proportions.

fixed points is depicted in Figure 4. A smaller proportion of the participants with state cues (31% and 23% of the participants in the close together and far apart conditions, respectively) visited those points than did the participants without state cues (38% and 53% of the participants in the close together and far apart conditions, respectively). A log-linear model was fit to the data, revealing a significant effect of state cues on this measure [$\chi^2(1,N = 104) = 4.18, p < .05$]. This trend in exploration behavior supports the generalization hypothesis over the fixed points view.

DISCUSSION

In this article, we examined the mechanism by which landmark-like state cues drive choice behavior in a dynamic decision task. The results of the experiment reported demonstrate how consistent cues reflecting the task environment state can lead participants either toward or away from globally optimal response patterns. State cues led to optimal choices when the payoff curves were close together but to suboptimal choices when they were far apart.

According to Gureckis and Love (in press), these cues afford generalization about local changes in rewards, which leads to suboptimal responding with the far apart payoff curves and optimal responding with the close together payoff curves. If state cues simply facilitate explicit comparison between the payoffs at fixed end points, we would expect to see optimal patterns of choice in the far apart condition. Instead, participants with state cues adopt the strategy observed by Gureckis and Love—generalizing about the payoff curve gradients and following local changes in reward—regardless of the positioning of the payoff curves. This extrapolation behavior can make locally inferior options with rising payoffs attractive even in cases in which such options are globally suboptimal.

A number of previous studies in which the impact of additional information (i.e., beyond immediate choice payoffs) on choice behavior in dynamic decision-making environments has been examined deserve mention. Warry, Remington, and Sonuga-Barke (1999) found that providing the expected payoffs of the choices on the next trial facilitated optimal choice behavior when the difference in immediate payoffs between payoff curves was large. In contrast, Neth et al. (2006) found that even prospective feedback reflecting participants' expected total earnings—emphasizing the *global* suboptimality of their choices—did not alter participants' ability to make globally optimal choices. Both of these manipulations were attempts to influence globally optimal responding, using global patterns of feedback. In contrast, our results show that local information (i.e., information about the current state) can have a strong impact on behavior. Our study shares some similarities with Herrnstein et al. (1993), which reported that a simple cue reflecting the current state of the task environment improved participants' ability to make payoff-maximizing responses. The present work extended these investigations, elucidating the mechanisms by which people, for better of for worse, utilize local state information to infer global solutions; specifically, we found that local state information facilitated generalization about payoff gradients. Surprisingly and counterintuitively, the efforts documented above were not able to elicit global changes in behavior, even with *global* feedback.

It is well documented that humans make use of landmark information to guide them through spatial decision spaces (e.g., Siegel & White, 1975). However, the present study evaluated the role of these cues in other, more abstract decision spaces. We found that humans use the structure of such cues to make inferences about unseen, future rewards, which can sometimes lead to suboptimal performance. One can easily conceive of a more complex decision environment in which the optimal response strategy is more complex than the "pure" strategy (i.e., repeatedly select the LT-I option) tested here. In these situations, adopting a strategy of following rising rewards—especially in the face of consistent state information—may be a kind of heuristic.

AUTHOR NOTE

This research was supported by start-up funds from New York University to T.M.G., AFOSR Grant FA9550-07-1-0178 and NSF CAREER Grant 349101 to B.C.L., and NIMH Grant MH077708 and AFOSR Grant FA9550-06-1-0204 to W.T.M. and A.B.M. Correspondence concerning this article should be addressed to A. Ross Otto, Department of Psychology, University of Texas, Austin, TX 78712 (e-mail: rotto@mail.utexas.edu).

REFERENCES

BOGACZ, R., MCCLURE, S. M., LI, J., COHEN, J. D., & MONTAGUE, P. R. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, **1153**, 111-121. doi:10.1016/j.brainres.2007.03.057

CARTWRIGHT, B. A., & COLLETT, T. S. (1982). How honey bees use landmarks to guide their return to a food source. *Nature*, **295**, 560-564. doi:10.1038/295560a0

DAW, N. D., O'DOHERTY, J. P., DAYAN, P., SEYMOUR, B., & DOLAN, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, **441**, 876-879.

GURECKIS, T. M., & LOVE, B. C. (in press). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*. doi:10.1016/j.cognition.2009.03.013

HERRNSTEIN, R. J., LOEWENSTEIN, G. F., PRELEC, D., & VAUGHN, W. (1993). Utility maximization and melioration: Internalities in individual choice. *Journal of Behavioral Decision Making*, **6**, 149-185. doi:10.1002/bdm.3960060302

MYERSON, J., & GREEN, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior*, **64**, 263-276. doi:10.1901/jeab.1995.64-263

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.

O'KEEFE, J., & NADEL, L. (1978). *The hippocampus as a cognitive map*. Oxford: Oxford University Press, Clarendon Press.

RACHLIN, H. (1995). Self-control: Beyond commitment. *Behavioral & Brain Sciences*, **18**, 109-159.

SIEGEL, A. W., & WHITE, S. H. (1975). The development of spatial representations of large-scale environments. *Advances in Child Development & Behavior*, **10**, 9-55.

STANKIEWICZ, B. J., LEGGE, G. E., MANSFIELD, J. S., & SCHLICHT, E. J. (2006). Lost in virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception & Performance*, **32**, 688-704. doi:10.1037/0096-1523.32.3.688

SUTTON, R., & BARTO, A. G. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

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

WARRY, C. J., REMINGTON, B., & SONUGA-BARKE, E. J. S. (1999). When more means less: Factors affecting human self-control in a local versus global choice paradigm. *Learning & Motivation*, **30**, 53-73. doi:10.1006/lmot.1998.1018

WHITEHEAD, S. D., & BALLARD, D. H. (1991). Learning to perceive and act by trial and error. *Machine Learning*, **7**, 45-83. doi:10.1023/A:1022619109594

(Manuscript received January 16, 2009;
revision accepted for publication May 21, 2009.)