The Impact of Perceptual Aliasing on Exploration and Learning in a Dynamic Decision Making Task

Lisa Zaval (lz2261@columbia.edu)
Columbia University, Department of Psychology
416 Schermerhorn, 1190 Amsterdam Ave., New York, NY 10027 USA

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

Abstract

Perceptual aliasing arises in situations where multiple, distinct states of the world give rise to the same percept. In this study, we examine how the degree of perceptual aliasing in a task impacts the ability of human agents to learn reward-maximizing decision strategies. Previous work has shown that the presence of perceptual cues that help signal distinct states of the environment can improve the ability of learners to adopt an optimal decision strategy in sequential decision making tasks (Gureckis & Love, 2009). In our experiments, we parametrically manipulated the *degree* of perceptual aliasing afforded by certain perceptual cues in a similar task. Our empirical results and simulations show how the ability of the learner improves as relevant states in the world uniquely map to differentiated percepts. The results provide further support for the model of sequential decision making proposed by Gureckis & Love (2009) and highlight the important role that state representations may have on behavior in dynamic decision making and learning tasks. **Keywords:** perceptual aliasing, dynamic decision making, reinforcement learning

Introduction

A crucial problem facing both human and artificial learners is correctly perceiving and interpreting the current state of the environment. For instance, imagine a traveler staying in an unfamiliar hotel, with each floor and exit decorated identically. Based on perceptual cues alone, this guest may experience difficulty navigating towards his room, since each floor is effectively indistinguishable. In order for navigation to be successful, the traveler must overcome the problem of *perceptual aliasing*, in which relevant "states" or situations in the world map to a single percept (Whitehead & Ballard, 1991; McCallum, 1993). In this example, that current state is the location of the traveler in the building, and the percept is the various cues available that might indicate this location. Note that environments may be aliased along a continuum from the perspective of any individual. For example, suppose that only every other floor in the building is decorated identically. In this case, the guest will be able to differentiate at least half the floors, and his ability to navigate might be somewhat improved. This example can be extended to cases where each floor of the hotel is uniquely decorated, such that salient perceptual cues indicate the traveler's location at any moment. Across these cases, the decision-making ability of the learner is expected to improve as the potential confusion is reduced, and relevant states in the world become mapped to differentiated percepts.

In this paper, we examine how the degree of perceptual aliasing in a task environment impacts the ability of humans to learn effective decision strategies in a dynamic task environment. A growing body of work suggests that human trial-and-error learning shares a similar computational foundation with algorithms developed in the reinforcement learning (RL) literature (see Dayan & Daw, 2008 for a review). However, less work has examined how the identification and categorization of distinct task states might interact with these learning and decision-making processes to determine human performance.

Previous Work

Our work builds upon previous studies of behavior in the "Farming on Mars" task (Gureckis & Love, 2009b, 2009a; Otto, Gureckis, Love, & Markman, 2009). In this task, participants make repeated selections between two "robots" presented on a computer screen. Selection of each robot results in a certain number of "oxygen" points. Participants' goal is to maximize the total amount of oxygen generated over the entire experiment. One robot (the "Short-term" option) always returns more points than the other (the "Long-term" option). However, unknown to participants at the start of the task, the experienced reward structure (i.e., payoff for selecting either robot) continually changes in response to the recent choice history of the participant. In particular, a dynamic is set up so that when the immediately attractive alternative is selected (i.e., the Short-term option), the long-term expected value of both robots is generally lowered on the following trial (Figure 1 illustrates the payout function used in previous Farming on Mars task experiments). Conversely, selections of the immediately worse option (the Long-term option) cause the expected value of both options to increase (in particular, the payoff for each option depends on the number of selections of the Long-term option over the last nine trials). As a result, the optimal reward-harvesting strategy is to learn to choose the option that

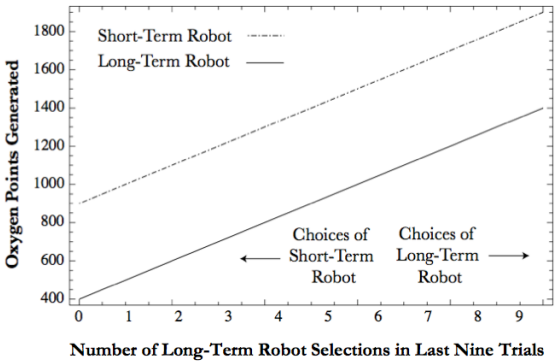

Figure 1: Illustrative payout function of the Farming on Mars Task. The horizontal axis in the figure represents the number of selections out of the last nine in which the Long-Term robot was chosen. The upper diagonal line measures the reward earned from choosing the Short-Term robot as a function of recent choice history, while the lower line illustrates the reward produced from Long-Term selections.

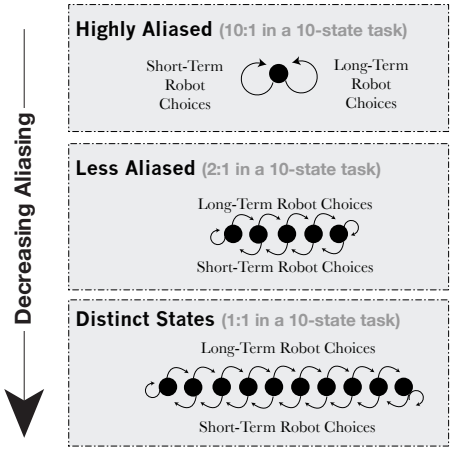

Figure 2: Degrees of perceptual aliasing. At the top is an example of a highly aliased environment where multiple distinct states maps onto a single percept (many-to-one). At the other extreme, distinct perceptual information disambiguates all states (one-to-one). Intermediate levels maps a subset of states to a single percept.

appears worse on each individual trial, since this strategy leads to the greatest cumulative reward.

Critically, performance in the task requires an appropriate balance of *exploration* (in order to discover the hidden contingencies) as well as *exploitation* of choice options known to be rewarding. In addition, a key observation about this task is that there are multiple distinct "states" of the environment (which correspond to the number of Long-term robot selections over the previous trials). When participants fail to recognize this structure, and the fact that the state of the system is changing as a function of their past response history, it becomes difficult to learn the reward-maximizing strategy. Consistent with this, Gureckis & Love (2009a,b) found that providing participants with simple perceptual cues that readily aligned with the state structure of the task improved their ability to learn the reward maximizing strategy. In their experiment, participants' display screen was augmented with a horizontal row of ten indicator lights which served as a cue indicative of the current state of the system. Participants who were given cues that correlated with the underlying task state performed better than participants attempting to learn without these cues. Further, results revealed that cues which supported generalization from one situation to the next had a more beneficial effect on performance relative to cues that effectively limited such generalization (see also Otto, et al., 2009). Gureckis & Love suggested that associating separate perceptual cues with each task "state" could reduce perceptual aliasing and facilitate more effective learning in the same way that appropriate state representations help artificial learning agents based on Q-learning (Sutton & Barto, 1998; Watkins, 1989).

The Present Studies

The present studies were designed to test a key prediction of Gureckis & Love's RL model. As anticipated by our example of the traveler in an unfamiliar hotel, the perceptual aliasing of states in the environment to distinct percepts can vary along a continuum (see Figure 2). At one extreme, every state in the world could map to the same percept (a many-to-one relationship). At the other extreme, each state in the world could map to a distinct percept (a one-to-one relationship). Intermediate cases exist where only a subset of distinct environmental states are perceptually aliased. One possibility is that any time distinct states are poorly differentiated, performance in situations such as the Farming on Mars task should suffer. Alternatively, it is possible that learners may still be able to acquire effective decision strategies when the representation of the task suggested by perceptual cues and the true structure of the task misalign, given that this misalignment takes a particular form. In other words, learners may not need to have a completely accurate representation of the task environment in order to still acquire a near-optimal reward-maximizing strategy. Indeed, this later hypothesis is what is predicted by Gureckis & Love's RL model which can still find optimal policies in some cases given misleading or inaccurate cues about the structure of the task. In the following experiments, we explore how various types of misalignment between perceptual information and task state information influences human learning. In particular, we are interested in how misalignments between perception of the world and the actual structure of contingencies influence learning and exploration behavior. Understanding the nature of this process is important since it is unlikely that human learners have completely accurate informa-

tion about the state structure of the environment at all times.

Experiment 1

In Experiment 1, each subject was randomly assigned to one of four conditions in the Farming on Mars task. Participants in each condition were given different types of perceptual cues which suggested a different interpretation of the nature of the task. Besides the type of cues displayed, each condition was identical with respect to the payoff function and task dynamics. The overall manipulation (providing different types of perceptual cues to learners in the task) parallels the approach in Gureckis & Love (2009).

In one condition (the *no-cue* condition), participants were given no additional cues as part of the display, and thus had to rely on memory and non-perceptual cues in order to uncover the optimal task strategy (c.f., Bogacz, McClure, Li, Cohen, & Montague, 2007). In the second condition (the *two-cue* condition), the interface screen was augmented with a simple cue consisting of two lights. At any point in time, only one of these lights was active, and a shift between the two cues indicated a change in the underlying task system. The position of the activated light was determined by the number of times the Long-term robot was selected over the previous nine trials of the experiment (this condition reflects a many-to-one situation with 5 states mapping to each percept). In the third condition (the *five-cue* condition), a circle of five lights (see Figure 3) was presented on the interface. The indicator lights were organized in a consistent array along the circle, such that the active light moved one position either clockwise or counterclockwise as the task state was updated. The five lights were mapped onto the underlying task system using a "modulus" rule, resulting in two distinct task states mapping to each percept. In the final condition, a display of ten lights was employed, such that each light corresponded exactly to a distinct numerical state in the underlying task system (one-to-one mapping).

Consistent with Gureckis & Love (2009a), we predicted that providing participants with light cue arrays which readily align with the underlying state of the system will limit the aliasing of functionally distinct states, and improve subjects' ability to learn the reward maximizing strategy. Thus, we predict that conditions where perceptual cues limit this aliasing (i.e., the ten-state condition) will result in better overall performance. In addition, we expect that participants' induced representation of the task will strongly influence the strategies they use to balance exploration and exploitation in the task.

Methods

Participants One hundred and ninety-two New York University undergraduates participated for course credit and a small cash bonus based on task performance. A total of 12 participants were dropped from the analysis for responding with the same button on more than 95% of the trials. The remaining participants were randomly assigned to one of four conditions: the *no-cue* condition (N = 44), the *two-cue* condition (N = 45), the *five-cue* condition (N = 45), and the *ten-cue* condition (N = 46).

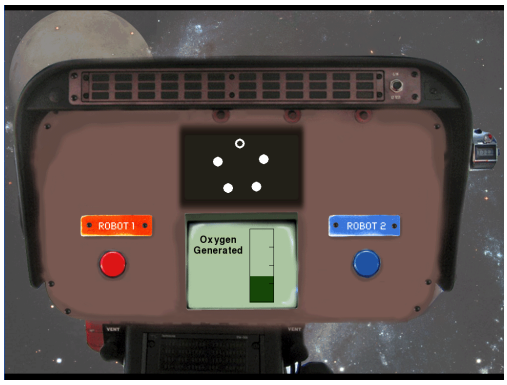

Figure 3: Example of the task interface used in the experiment. The display shows the indicator lights used in the five-cue condition. Additionally, the screen illustrates how rewards were conveyed to participants.

Materials and design The experiment was administered on standard Macintosh computers using an in-house data collection system written in Python[1]. Participants were tested individually over a single one-hour session. Extraneous display variables, such as which robot corresponded to the left or right choice option, the position of the lights, and which direction the active light moved (clockwise or counter-clockwise), were counterbalanced across participants. On each trial, the payoff for selecting the Long-term robot was $40+70*h/9$, where h is the number of times the Long-term robot was selected in the last 9 trials. In contrast, the payoff on each trial for the Short-term robot was $30 + 70 * h/9$. The final values were scaled by 110 and displayed as a percentage on the sliding oxygen meter.

Procedure Participants were tested in the basic Farming on Mars task as described above. At the beginning of the experiment, subjects were presented instructions on the screen which conveyed the basic cover story for the task. The instructions were identical for all conditions, and there was no explicit reference to the function or purpose of the indicator lights/cues. On each trial, participants were shown a display with two large response buttons. Between these buttons was a video display which presented trial-relevant feedback. After a robot selection was made, the quantity of oxygen produced for that trial was presented on the video display. The amount of oxygen points earned was presented visually with a vertical, sliding bar which filled green to

[1] http://www.pypsyexp.org

varying levels. The oxygen level display was shown for 800 ms, after which the the screen was reset to indicate the start of a new trial. No information regarding cumulative oxygen generation was presented, but instructions did emphasize that participants should try to "maximize the number of oxygen points generated over the entire experiment." In the two-light, five-light, and ten-light conditions (but not in the no-cue condition), the screen was augmented with an array of indicator lights as described above and shown in Figure 3. The experiment consisted of 500 separate trials divided into five blocks of 100 trials. In order to maintain motivation, participants were informed that they would receive a small cash bonus of $2-5 dollars based on total oxygen generated by the end of the task.

Results

The primary dependent measure in our experiment was the proportion of Long-term robot selections (i.e., reward-maximizing responses) made by the participant. Total mean proportions by condition are presented in Figure 4. Overall, the proportion of Long-term choices were significantly higher than chance in all conditions, except for the five-cue condition (all $p < .05$). Given the binary outcome choice data, we conducted a series of binomial regressions using the χ^2 distributed deviance-based test as our measure of model selection[2]. There was an overall significant effect of condition $\chi^2(3) = 15.6$, $p = .001$. In addition, the pattern of results across conditions was best predicted as a quadratic function of the number of perceptually distinct task states compared to a linear relationship ($\chi^2(1) = 11.32$, $p < .001$, the quadratic term was reliably above zero, $\beta_{cond^2} = .02$, $p < .001$). Pairwise contrasts (using an Bonferroni-adjusted $\alpha = .05/4 = .012$) between the individual conditions revealed a significantly higher proportion of maximizing responses in the ten-cue condition compared to both the five-cue condition, $\chi^2(1) = 13.46$, $p < .001$, and the two-cue condition, $\chi^2(1) = 11.62$, $p < .001$. Surprisingly, there was a relatively small difference between the ten-cue and no-cue conditions which did not reach significance, $\chi^2(1) = 3.59$, $p = .06$. Note, however, that in a similar task, Gureckis & Love (2009b) and Otto, et al. (2009) found an advantage for one-to-one percept-state representations. Also, note that when given only two state cues, performance was not significantly better than when participants are given five state cues, $\chi^2(1) = 1.04$, $p = .3$.

In order to better understand the genesis of the aliasing effect, we examined the *dynamics* of exploration in the task. In particular, even if the marginal proportion of maximizing choices is constant, it is possible that the distribution of those choices in time could vary. For example, participants in the different conditions might adopt alternative strategies for exploring the task. One way to quantify these differences is to plot the percentage of total trials participants spent in each true (latent) state in the task. Remember that "states" in this dynamic task are defined by the percent allocation of choices to the Long-term option over the last nine trials. Figure 1 plots this distribution for each of the four conditions. Interestingly, the structure of the cues in the task has a strong impact on the way participants explored the task dynamic. In particular, participants in the two-cue condition spent a much larger percentage of time in intermediate states (indicated roughly equal allocation to both choices for extended periods of time). For example, a one-way ANOVA on proportion of time spent in states 3-7 revealed an effect of condition, $F(3, 132) = 4.57$, $p < .005$. Specifically, participants in the two-cue condition spent more total time in these intermediate states than in the no-cue, $t(64) = 2.95, p < .005$, five-cue, $t(66) = 2.31, p < .02$, and ten-cue, $t(66) = 3.43, p = .001$, conditions (since these are post-hoc analyses significance should be interpreted using a conservative $\alpha = .05/3 = .016$). On the other hand, there was also a significant effect of condition on how long participants spend in the end point states (i.e., state 1 & 2 and 9 & 10), $F(3, 132) = 3.25$, $p < .025$. Post-hoc test revealed this was driven primarily by the lower percentage of total time spent in these states in the two-cue condition condition compared to the 10-cue condition, $t(66) = 3.17$, $p < .003$.

Discussion

The results of Experiment 1 show that participant's conceptualization of the state structure of the task can influence both their exploration strategies as well as their ability to identify a reward maximizing strategy. In particular, when cues about the underlying state of the states were more highly aliased (the two-cue and five-cue conditions) participant's overall task performance suffered. Closer examination of the way in which participants explored the task revealed that the alignment of the cues in the task had a dramatic effect on behavior, even when overall performance differences appeared smaller. In particular, relative to the other conditions, participants in the two-cue condition spent a considerably longer time in intermediate states, consistent with a choice strategy involving alternations between the short-term and long-term options.

Experiment 2

In Experiment 1 we found that reward-maximizing performance was worst when a circle of five indicator lights was presented on the interface, such that two different task states mapped to the same perceptual display. However, it is as yet unclear if the performance difference for highly aliased environments results from the num-

[2]We also analyzed these data through a one-way ANOVA and a series of t-tests which revealed an identical pattern of significant results.

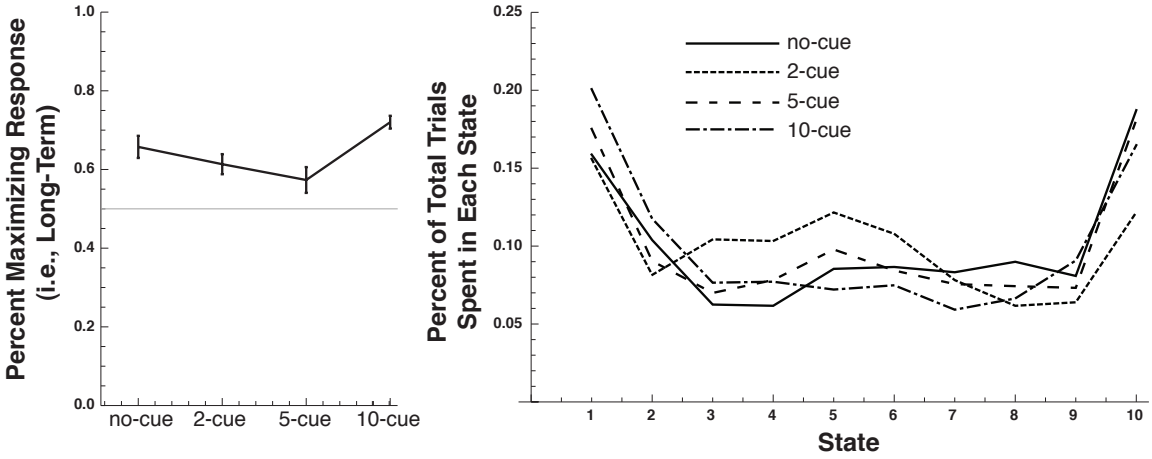

Figure 4: *Panel A*: Average proportion of Long-Term (maximizing) responses made throughout the experiment as a function of condition. The horizontal line at 0.5 shows chance performance. Error bars are standard errors of the mean. *Panel B*: Average percentage of total experiment spent in each state. State 1 corresponds to 0 of the last nine choices being to the Long-term option. State 10 corresponds to 9 of the last 9 choices being to the Long-term option.

ber of implied states (5) or how those states "blend together" by the dynamics of the focal cue (i.e., the active light). For example, in the *five-cue* condition of Experiment 1, the active cue moved one position either to the left or right as the state of the underlying system was updated. Thus, a participant who steadily progressed from states 1-10 would experience the active light looping twice around the circle of indicator lights. An alternative display which maintains the same level of perceptual aliasing (two true states for every one distinct percept) would be to have the active light remain in the same position across two consecutive state updates. In this design, a participant who steadily progressed from states 1-10 would observe the active light making a single loop around the five indicator lights, 'doubling-up' at each individual light position. In other words, if the letter A-E represent the five locations for the state cue, then the mapping from the 10 latent task states to the display would be 1,2,3,4,5,6,7,8,9,10→A,A,B,B,C,C,D,D,E,E. In Experiment 2, we compare task performance in this *single-looped* condition with performance in the *twice-looped* condition (which is identical to the 'five-cue' condition of Experiment 1).

Our prediction was that performance in the twice-looped condition would be lower than in the single-looped condition. The rationale was that participants in the single-looped condition would be better able to recognize that the "gradient" of reward was rising as the light moved in a particular direction. In contrast, the twice-looped condition would be more likely to be confused as a state that they had previously experienced to have low reward (e.g., state cue position A) might later also be associated with high reward. The prediction that the perception of a correlation between the movement of the light and the magnitude of the reward is supported by previous studies showing that participants use such information even when it is against their best interest in the task (Otto et al., 2009).

Methods

Participants Forty New York University undergraduates participated for course credit and a small cash bonus based on task performance. Participants were randomly assigned to either the *twice-looped* condition (N=21) or the *single-looped* condition (N = 19).

Materials and design All aspect of the materials and design were identical to Experiment 1, except for the changes to the five-cue display described above.

Procedure The general procedure was the same as in Experiment 1.

Results

As before, the primary dependent measure in our experiment was the proportion of Long-term robot selections (i.e., reward-maximizing responses) made by the participant. However, there was no overall effect of condition $\chi^2(1) = 0.26$, $p = .61$, M=0.52 in the twice-looped condition and M=0.54 in the single-looped condition. Closer examination of the distribution of overall performance scores indicated that the distribution was strongly bimodal in the twice-looped condition, while it was uni-model in the single-looped condition. As shown in Figure 5, this bi-modality arose from the way that participants explored the latent task states. In particular, a 2-way repeated measures ANOVA on condition

and time spent in each state found a significant effect of state, $F(9, 342) = 4.12$, $p < .001$, and a significant state by condition interaction, $F(9, 342) = 3.17$, $p < .001$. At least a subset of participants in the twice-looped condition appeared to have spent a disproportion amount of time in state 6 which is the point where the display looped back on itself suggesting that they were attempting to keep the state cue from crossing back around to the state associated with the lowest reward. In contrast, participants in the single-looped condition spent more time in the lower states (1-4) indicating that they had an overall bias towards the short-term option that a subset of participants eventually overcame.

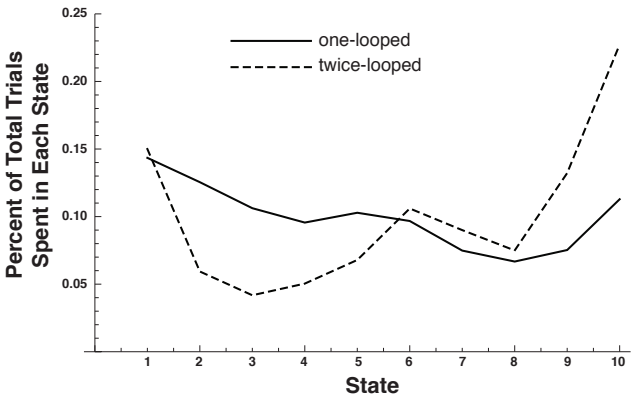

Figure 5: Average percentage of total experiment spent in each state in Experiment 2.

General Discussion

Across a set of two experiments we explored how perceptual cues concerning the underlying state structure of a dynamic decision making task influenced learning. Consistent with previous work (Gureckis & Love, 2009b, 2009a), we find that when task states are aliased, participants' ability to identify an optimal task strategy is impaired. It is important to point out that the effects we see here are unlikely to be a simple consequence of participants ignoring the primary task (to earn oxygen points) and instead exploring aspects of the display. First, participants were clearly instructed that the primary goal was to control the system to earn as many points as possible. In addition, participants were paid a small cash bonus tied to their performance in the task which increased the relevance of the primary task. Finally, our analysis of the dynamics of exploration (i.e., the percent of time spent in each state) reveal systematic differences related to the structure of the cues we provided.

One possibility is that the structure of the perceptual cues provide a kind of strategy "affordance" in the task, limiting the space of exploration/response policies that participants considered. Note that in a separate study, we recently found that motivational manipulations can also impact participant's exploration behavior in a similar task (Otto, Markman, Gureckis, & Love, in review). A theoretical analysis of these results and evaluation of their implication for the Gureckis & Love (2009) model are currently underway. However, preliminary simulations show a close correspondence between the results reported here and the behavior of the model. Future work will continue to evaluate how RL models can be used to understand the motivational and cognitive influences underlying dynamic decision-making.

Acknowledgements We thank Louis Tur and Nathaniel Blanco for programming assistance and discussion in the development of this project.

References

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.

Dayan, P., & Daw, N. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, and Behavioral Neuroscience*, *8*, 429-453.

Gureckis, T., & Love, B. C. (2009a). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, *53*, 180-193.

Gureckis, T., & Love, B. C. (2009b). Short term gains, long term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*(3), 293-313.

McCallum, R. (1993). Overcoming incomplete perception with utile distinction memory. In *The proceedings of the tenth international machine learning conference (ml'93)*. Amherst, MA.

Otto, A., Gureckis, T., Love, B., & Markman, A. (2009). Navigating through abstract decision spaces: Evaluating the role of state knowledge in a dynamic decision making task. *Psychonomic Bulletin and Review*, *16*(5), 957-963.

Otto, A., Markman, A., Gureckis, T., & Love, B. (in review). Regulatory fit in a dyanmic decision-making environment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

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

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.