Actions that Modify Schedules of Reinforcement

Mac Strelioff (mstrelio@uci.edu)
Mimi Liljeholm (m.liljeholm@uci.edu)
Department of Cognitive Sciences, University of California Irvine
Irvine, CA 92697-5100 USA

Abstract

Many everyday activities involve the use of one action to modify the effects of another: When driving, shifting gears modifies the influence of pressing the gas pedal on acceleration; when cooking, the rate of adding a particular ingredient modifies the influence of stirring on viscosity. Here, we investigate a general ability to learn how to use actions to control schedules of reinforcement. In Experiment 1, participants quickly discovered the optimal rate of responding on an action that controlled the rate of reward contingent on performing a different action. In Experiment 2, when the modifying action was itself rewarded, participants failed to discover the optimal rate. Implications for formal theories of instrumental behavior are discussed.

Keywords: Schedules of reinforcement; reward learning; instrumental contingencies.

Introduction

Since the early 20th century, researchers have investigated the influence of various reward schedules on the rate and selection of instrumental responses. For example, ratio schedules, in which reward delivery depends on the number of responses since the last reward, produce higher rates of responding than do interval schedules, in which reward delivery depends on the time elapsed since the last reward (Fester & Skinner, 1957). When two or more action alternatives are available, that which yields the greatest, most immediate, or most certain reward is, all other things being equal, generally that most frequently selected (e.g., Rachlin et al., 1991). However, in the real world, many responses serve only to modulate the effects of other actions: The rate and pattern of pressing strings on a guitar does not itself yield music, but profoundly impacts the sounds produced by strumming. Here, we assess a domain-general capacity for learning about actions that control schedules of reinforcement on other actions.

Formally, the relationship between a particular action and its outcome has been modeled as a complex associative structure (Dickinson & Balleine, 1993), as the difference between probabilities of reward given the presence versus absence of the action (Hammond, 1980), as the probability and subjective utility of the outcome given the action (Savage, 1954), or as a cached value assigned to the action based on its reinforcement history (Watkins, 1989). What these diverse approaches have in common is that they address the identity and/or latency of a single action at a time, ignoring situations in which multiple actions are performed in concert and potentially interact. In our paradigm, an intermediate rate of responding on one action maximizes the reward contingent on performing a different, concurrently available, action.

Experiment 1

Methods

Participants Thirty undergraduates at the University of California, Irvine (22 females; mean age=20±2.17) participated in the study for course credit. All participants gave informed consent and the study was approved by the Institutional Review Board of the University of California, Irvine.

Task & Procedure The task is illustrated in Figure 1. We used a free operant paradigm in which participants were allowed to respond at will on either or both of two concurrently available actions, graphically represented on the computer screen, by pressing the corresponding keys on the computer keyboard. Whenever a response was made a selection square appeared around the chosen action for 300ms. If the response was rewarded, an image of a quarter appeared center screen for 500ms and a count of the cumulative monetary earning, continuously displayed above the quarter image location, would increment by +$0.25. The task was comprised of ten 2-minute blocks separated by brief rest periods. All monetary earnings were fictitious.

In the "Modify" group (n=15), the rate of responding on a "modifying" action influenced the probability that the concurrently available "modified" action would produce a reward. When the modifying action was performed at an "optimal" rate of 1.25 to 2.75 presses per second, the probability of reward given a response on the modified action was 0.9. When response rates on the modifying action were outside of the 1.25 to 2.75 range, the probability of reward given the modified action was 0. The modifying action did not itself produce any reward. Response rates on the modifying action were tracked using a differential equation that increased by an impulse of 1 at the time of a response and decayed each impulse at a linear rate of 0.2 per second, so that each impulse from a response decayed to zero after 5 seconds.

Specifically, for an impulse (a_i), which was 1 if an action were taken during the current iteration of the program and 0 otherwise, a decay rate of 0.2, a counter for the number of responses that occurred within the last 5 seconds (N_5) and the difference in time between the current iteration of the program and the previous iteration (dt), the response rate variable (R) was updated on each iteration i by:

$$R_i \leftarrow R_{i-1} + a_i - 0.2N_5 dt$$

This method adjusts more quickly to changes in response rate than the commonly used approach of dividing the number of responses in a time window by the length of the window (e.g., Soto et al., 2006). The probability of reward on the modified action was set to 0.9 whenever the response rate variable, R, was in the optimal, 1.25 to 2.75, range and 0.0 otherwise.

Note that the optimal rate of responding on the modifying action was intermediate; this was done to rule out the contribution of systematic biases of either very high or very low responding. On the other hand, an intermediate rate might represent an average towards which most responders converge in free operant tasks. To address this possibility, a second, "Yoked", group was included (n=15), in which the rate of responding on the modifying action had no influence, while the probability of reward on the modified action was yoked to that of a participant in the Modify group. We predicted that, by the end of the session, participants in the Modify group would respond on the modifying action at a rate falling within the optimal range, while those in yoked group would not.

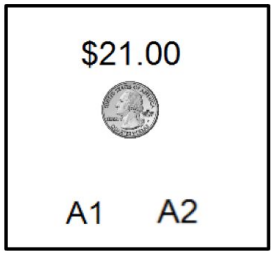

Figure 1: Task Illustration, see text for details.

Results

We divided the number of responses on the modifying action in each 5-second bin of task performance with 5 (i.e., responses per second), and computed the distance of this response rate from the bounds of the optimal range, for the first and last 5 seconds of the task. We then used a mixed analysis of variance (ANOVA) with "group" as the between-subject factor and "bin" as the within-subject factor to assess a change in optimal responding between the first and last bins. There was no main effect of bin, F(1,28)=3.19, *p=0.09*, but a main effect of group, F(1,28)=6.53, *p<0.05*, and, critically, a bin-by-group interaction, F(1,28)=5.78, *p<0.05*. Planned comparisons revealed that while the two groups did not differ with respect to optimal responding on the modifying action in the first bin, t(28)=0.13, *p=0.89,* by the last bin, participants in the Modify group were significantly closer to the optimal response rate than were participants in the Yoked group, t(28)=3.61, *p<0.01*. As can be seen in Figure 2, while the mean deviation from the optimal rate significantly decreased from the first to the last bin in the Modify group, t(14)=2.69, *p<0.05,* they remained unchanged across bins in the Yoked group, t(14)=0.05, *p=0.63*. The apparent absence of a change in optimal responding by Yoked participants reflects a tendency to either increase or decrease responding on the modifying action across blocks, resulting in no net change for the group; in contrast participants in the Modify group coherently converged towards the optimal rate.

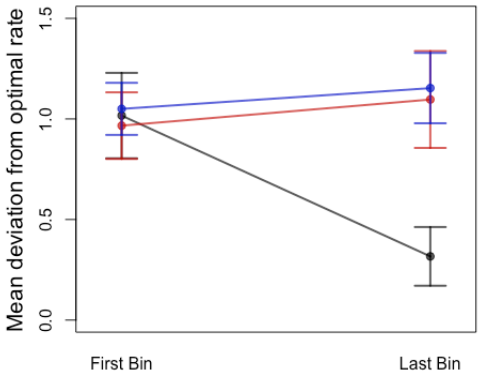

Figure 2: Mean deviation of response rates on the modifying action from the optimal range in the first and last 5 seconds of task performance, for subjects in the Modify (black) and Yoked (blue) groups, and for the single group of Experiment 2, in which the modifying action was rewarded (red). Error bars=SEM.

We also assessed performance in terms of the proportion of bins with optimal response rates, early and late in the task. Bins were scored as optimal if the windowed (5 seconds) response rate was in the optimal range of 1.25 to 2.75 responses per second. For each subject, we assessed the number of optimal 5-second bins in the first and last 30 seconds of the task. (We used 30 seconds, rather than the full 2-minute blocks, to ensure that the index of early learning did not include already asymptotic performance.) The results using this metric were consistent with those described above: The groups did not differ in the first 30-second block, t(28)=1.12, p=0.24, but by the last 30-second block, the mean proportion of optimal bins was significantly greater for the Modify group than for the Yoked group, t(28)=6.93, p<0.01. Indeed, while the proportion of optimal bins increased significantly from the first to the last block in

the Modify group, t(14)=3.60, $p<0.01$, it *decreased*, albeit with marginal significance, t(14) = 2.09, $p=0.06$, in the Yoked group. The mean proportion of optimal bins in each 30-second block throughout the task is shown in Figure 3.

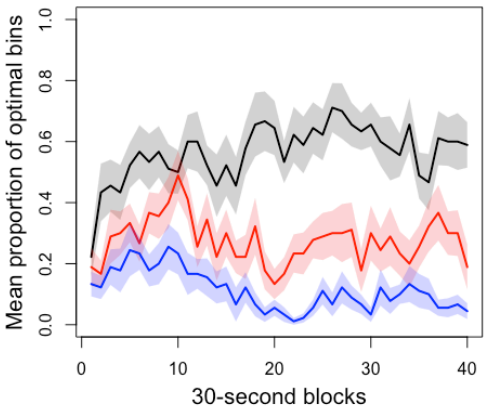

Figure 3: Mean proportion of bins with an optimal response rate on the modifying action in each 30-second block of the task for the Modify (black) and Yoked (blue) groups of Experiment 1, and for the Reward group of Experiment 2 (red). Shading=SEM.

With respect to the modified action, response rates were, overall, higher than those on the modifying action, for both the Modify, t(14)=3.10, $p<0.05$, and Yoked, t(14)=6.76, $p<0.05$, groups. This likely reflects the fact that, while the probability of reward given the modified action was either a function of (Modify) or independent of (Yoked) responding on the modifying action, the actual *delivery* of reward was contingent only on performing the modified action.

Experiment 2

A well-studied phenomenon closely related to our query is that of "melioration" – a tendency to select an action alternative that produces a greater immediate pay-off, but that, when selected repeatedly, lowers the overall rate of reward (Herrnstein, 1991). Such tendencies are commonly attributed to impulsivity (Herrnstein, 1991; Otto, Markman, & Love, 2012), but have also been described as rational choices under uncertainty (Gureckis & Love, 2009a, 2009b; Sims et al., 2013). Other related paradigms, such as delay discounting (Ainslie, 1975; Johnson & Bickel, 2002) and differential reinforcement of low response rates (Wilson & Keller, 1953; Carter & MacGrady, 1966), have convincingly demonstrated the interfering influence of salient reward on rational decision-making (Ainslie, 1975; Van den Broek, Bradshaw, & Szabadi, 1987).

In Experiment 2, we assess whether the lure of an immediate reward results in a failure to suppress responding on the modifying action, thus interfering with the ability to control the schedule of reinforcement on the modified action.

Methods

Participants Fifteen undergraduates at the University of California, Irvine (10 females; mean age=19.7±1.1) participated in the study for course credit. All participants gave informed consent and the study was approved by the Institutional Review Board of the University of California, Irvine.

Task & Procedure Participants performed a task that was identical to that of the Modify group in Experiment 1, with one exception: In addition to modulating the schedule of reinforcement on the modified action, the modifying action was itself rewarded by $0.25, with a probability of 0.2. Note that, since this reward probability is much lower than the conditional, 0.9, probability of reward on the modified action, maintaining an optimal, intermediate, response rate on the modifying action dramatically increases the average reward rate.

Results

We computed the same measures of optimal responding as those used in Experiment 1. Comparing the first and last 5 seconds of performance, there was no change in the deviation of response rates on the modifying action from the bounds of the optimal rate, t(14)=0.48, $p=0.64$ (see Figure 2). Likewise, the proportion of optimal bins did not differ between the first and last 30-second blocks of the task, t(14)=0.00, $p=1.00$. In the absence of random assignment, we refrain from making any statistical comparisons between the results of this experiment and those obtained in Experiment 1. Nonetheless, it is worth noting that, as illustrated in Figures 2 and 3, when the modifying action was itself rewarded, the rate of responding on the modifying action was apparently closer to that in the Yoked group than in the Modify group. Finally, although, overall, response rates were again higher on the modified than the modifying action, unlike for the groups in Experiment 1, this difference was only marginally significant, t(14)=2.09, $p=0.06$, presumably reflecting the fact that, in Experiment 2, reward delivery was potentially contingent on performing either action.

General Discussion

In two experiments, we assessed the discovery and performance of an action that controlled the schedule of reinforcement on another, concurrently available, action. In Experiment 1, participants quickly discovered and implemented an optimal, intermediate, response rate on a modifying action that, while not producing any rewards itself, modulated the reward contingent on a distinct, concurrently available, action. Response rates

in a yoked control group confirmed that convergence to the optimal rate was due to the influence of the modifying action on the reward schedule of the modified action. In Experiment 2, consistent with a large literature on the failure to suppress inappropriate responding in the face of immediate reward (Ainslie, 1975; Carter & MacGrady, 1966; Van den Broek *et al*., 1987; Wilson & Keller, 1953), reinforcement of the modifying action apparently prevented discovery of the optimal response rate. The focus in the existing literature on the disruptive effects of immediate reward has largely overshadowed the question raised here of whether, and how, agents learn about actions that modify schedules of reinforcement. Our results suggest that, in the absence of interfering or competing reward contingencies, increasing levels of instrumental control can be achieved by incorporating information about dependencies between actions.

In a model-free reinforcement learning account of free operant responding, Niv et al. (2007) proposed that, for each decision, the agent selects both the latency and the identity of the to-be-executed action, based on the relative degree to which that action increases the average reward rate. Although it is possible that participants in the Modify group of Experiment 1 similarly learned about the modifying action based on its reinforcement history, several aspects of our task depart from the specification of Niv et al. (2007). Most notably, participants in our task would have to include a representation of the modified action in their state space when updating the value of the modifying action – that is, assess the value of a particular latency of the modifying action *given* that the modified action is simultaneously[1] or proximally performed – since the modifying action is never itself rewarded. Likewise, the value of the modified action has to be specified conditional on the performance of the modifying action, since the probability of reward on the former is zero whenever responding on the latter falls outside the optimal range. It is of course possible to specify a model-free learner that has enough conditionals built into its state-representation to identify the combination of responding on modifying and modified actions that maximizes reward[2].

An alternative, model-based, approach is for the agent to create a graphical probabilistic model representing the dependencies between actions, states and rewards (e.g., Acuna & Schrater, 2010). Although initially ignorant of the nature of these dependencies, a Bayesian reinforcement learner generates beliefs over a set of possible dependency structures and updates those beliefs, after each observation, using Bayesian inference. For example, a learner in our task might postulate two possible worlds: one in which the latency to respond on an action can modulate the probability of reward given that same, or some other, action, and one in which response latencies have no influence on schedules of reinforcement. The former possibility must of course be further partitioned into several putative structures, each with a particular set of links (e.g., an action modifying its own probability of reward vs. that of a different action) and associated parameters. The learner then updates the belief distribution over structures based on sequences of actions, latencies and rewards.

Critically, the approach sketched in the previous paragraph, to address model-based inferences regarding action dependencies, can also be used to explain some of the most basic aspects of instrumental behavior, such as the distinction between interval and ratio schedules – Recall that, whereas on interval schedules a response is rewarded based on the amount of time elapsed since the last reward, on ratio schedules a response is rewarded based on the number of responses since the last reward. These qualitatively different schedules produce distinct response profiles (Fester & Skinner, 1957), suggesting some, implicit or explicit, discrimination by the agent. Notably, the interval schedule can be conceptualized as a case in which the rate of performing an action modifies the *schedule of reinforcement*, rather than just the rate of reward: Specifically, on a given interval schedule, any response rate greater than "one per the required interval" will decrease the probability of reward conditional on that action. Other well-established schedules, such as differential reinforcement of high or low responding (Van den Broek, *et al*., 1987) can also be characterized as actions modifying schedules of reinforcement, as can the "seeking" component of seeking-taking schedules (Balleine, Garner, Gonzalez, & Dickinson, 1995). Thus, the framework proposed here potentially applies to a wide range of instrumental phenomena.

At the neural level, model-free and model-based RL approaches have been mapped to dissociable neural substrates, with the ventral striatum, posterior putamen and premotor cortex being implicated in model-free responding (Glascher et al., 2010; Lee et al., 2014; Tricomi et al., 2009; de Wit et al., 2012), and the caudate, ventromedial prefrontal cortex and inferior parietal lobule in model-based computations (de Wit et al., 2012; Liljeholm et al., 2011, 2013, 2015; Lee et al., 2014). It should be noted, however, that, with some exceptions (e.g., Liljeholm et al., 2013), the work

[1] Note that even the possibility of simultaneously performing multiple responses falls outside the scope of Niv et al.'s (2007) model, according to which all action-latency pairs are serially implemented (i.e., no alternative actions may be executed while the time indicated by the chosen latency passes).

[2] Indeed, Niv et al.'s (2007) model hard-codes into the definition of each state several variables that are needed to discover an optimal policy in the environments addressed by the model (e.g., the time elapsed since the last response when modeling interval schedules and the number of presses since the last reward when modeling ratio schedules).

identifying such dissociations has focused on relatively simple model-based processes, such as the encoding of individual action-outcome contingencies or sensitivity to changes in an outcomes utility. In contrast, the model-based learner postulated here engages in complex reasoning regarding how actions may be used to control action-outcome relationships. Such processes may warrant the involvement of brain regions known to support relational and inductive reasoning, including the rostolateral and dorsolateral prefrontal cortex (e.g., Krawczyk et al., 2011).

Finally, an important point regarding action dependencies such as those addressed here is how they relate to the actual representations of actions. In our task, the instructions and materials clearly defined and distinguished between action alternatives (see Figure 1), so that there could be little doubt about how many, and exactly what, actions were available. It is interesting to consider how inferences and performance might have differed had the grouping of elements into discrete action alternatives been more ambiguous. One possibility is that increasing ambiguity would afford a more rapid acquisition of relevant dependencies (Pezzulo, Rigoli, & Friston, 2015) and, further, that those inferred dependencies might serve to configure action elements into more clearly delineated action representations based on reinforcement learning principles (e.g. Reynolds, & O'Reilly, 2009).

In conclusion, we have demonstrated a domain-general ability to learn about, and take advantage of, an action that modifies the schedule of reinforcement on a different action. We have also sketched a model that, by making inferences about dependencies between response latencies and conditional reward probabilities, might account for behavior across a wide range of instrumental schedules. Future work will focus on extensions of our experimental paradigm, further development of formal accounts, and investigations of mediating neural substrates.

References

Acuna, D., & Schrater, P. (2010). Structure learning in human sequential decision-making. *PLoS Computational Biology*, 6, 1–8.

Ainslie, G. (1975). Specious reward: a behavioral theory of impulsiveness and impulse control. *Psychological bulletin*, *82*(4), 463.

Balleine, B. W., Garner, C., Gonzalez, F., & Dickinson, A. (1995). Motivational control of heterogeneous instrumental chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*(3), 203.

Carter, D. E. & MacGrady, G. J. (1966). Acquisition of a temporal discrimination by human subjects. *Psychonomic Science*, 5, 309-310.

de Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., van de Vijver, I., & Ridderinkhof, K. R. (2012). Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. *Journal of Neuroscience*, *32*(35), 12066-12075.

Dickinson, A., & Balleine, B. (1993). Actions and responses: The dual psychology of behaviour. In N. Eilan, R. A. McCarthy, & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology* (pp. 277-293). Malden: Blackwell Publishing

Fester, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. New York, NY: Appleton-Century-Croft.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585-595.

Gureckis, T. M., & Love, B. C. (2009a). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, 53(3), 180–193.

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

Hammond, L. J. (1980). The effect of contingency upon the appetitive conditioning of free‐operant behavior. *Journal of the experimental analysis of behavior*, *34*(3), 297-304.

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

Hogarth, L., Dickinson, A., Wright, A., Kouvaraki, M. & Duka, T. (2007). The role of drug expectancy in the control of human drug seeking. *Journal of Experimental Psychology-Animal Behavior Processes*, *33*(4), 484-496.

Johnson, M. W., & Bickel, W. K. (2002). Within-subject comparison of real and hypothetical money rewards in delay discounting. *Journal of the experimental analysis of behavior*, *77*(2), 129-146.

Krawczyk, Daniel C., M. Michelle McClelland, and Colin M. Donovan. "A hierarchy for relational reasoning in the prefrontal cortex." *Cortex* 47.5 (2011): 588-597.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning.*Neuron*, *81*(3), 687-699.

Liljeholm, M., Tricomi, E., O'Doherty, J. P., & Balleine, B. W. (2011). Neural correlates of instrumental contingency learning: differential effects of action–reward conjunction and disjunction. *Journal of Neuroscience*, *31*(7), 2474-2480.

Liljeholm, M., Wang, S., Zhang, J., & O'Doherty, J. P. (2013). Neural correlates of the divergence of

instrumental probability distributions. *Journal of Neuroscience*, *33*(30), 12519-12527.
Liljeholm, M., Dunne, S., & O'doherty, J. P. (2015). Differentiating neural systems mediating the acquisition vs. expression of goal‐directed and habitual behavioral control. *European Journal of Neuroscience*, *41*(10), 1358-1371.
Niv, Y., Daw, N. D., Daphna, J., & Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology, 191(3), 507–520.
Otto, A. R., Markman, A. B., & Love, B. C. (2012). Taking more, now: The optimality of impulsive choice hinges on environment structure. Social Psychological and Personal-ity Science, 3, 131–138.
Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology*, *134*, 17-35.
Rachlin, H., Raineri, A., & Cross, D. (1991). Subjective probability and delay. *Journal of the experimental analysis of behavior*, *55*(2), 233-244.
Reynolds, J. R., & O'Reilly, R. C. (2009). Developing PFC representations using reinforcement learning. *Cognition*, *113*(3), 281-292.
Savage, Leonard J. (1954). *The Foundations of Statistics*. New York, Wiley.
Sims, C. R., Neth, H., Jacobs, R. A., & Gray, W. D. (2013). Melioration as rational choice: Sequential decision making in uncertain environments. Psychological Review, 120(1), 139–154.
Soto, P. L., McDowell, J. J., & Dallery, J. (2006). Feedback functions, optimization, and the relation of response rate to reinforcement rate. *Journal of the Experimental Analysis of Behavior*, *85*, 57-71.
Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, *29*(11), 2225-2232.
Van den Broek, M. D., Bradshaw, C. M., & Szabadi, E. (1987). Behaviour of 'impulsive'and 'non-impulsive'humans in a temporal differentiation schedule of reinforcement. *Personality and Individual Differences*, *8*(2), 233-239.
Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Doctoral dissertation, University of Cambridge).
Wilson, M. P. & Keller, F. S. (1953). On the selective reinforcement of spaced responding. *Journal of Comparative and Physiological Psychology*, 46, 190-193.