Psychological Science
24(5) 751–761
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797612463080
pss.sagepub.com

The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive

A. Ross Otto[1], Samuel J. Gershman[2], Arthur B. Markman[1], and Nathaniel D. Daw[3]

[1]Department of Psychology, University of Texas at Austin; [2]Department of Psychology and Princeton Neuroscience Institute, Princeton University; and [3]Department of Psychology and Center for Neural Science, New York University

Abstract

A number of accounts of human and animal behavior posit the operation of parallel and competing valuation systems in the control of choice behavior. In these accounts, a flexible but computationally expensive model-based reinforcement-learning system has been contrasted with a less flexible but more efficient model-free reinforcement-learning system. The factors governing which system controls behavior—and under what circumstances—are still unclear. Following the hypothesis that model-based reinforcement learning requires cognitive resources, we demonstrated that having human decision makers perform a demanding secondary task engenders increased reliance on a model-free reinforcement-learning strategy. Further, we showed that, across trials, people negotiate the trade-off between the two systems dynamically as a function of concurrent executive-function demands, and people's choice latencies reflect the computational expenses of the strategy they employ. These results demonstrate that competition between multiple learning systems can be controlled on a trial-by-trial basis by modulating the availability of cognitive resources.

Keywords

cognitive neuroscience, decision making

Received 3/30/12; Revision accepted 8/24/12

Accounts of decision making across cognitive science, neuroscience, and behavioral economics posit that decisions arise from two qualitatively distinct systems that differ broadly in their reliance on controlled versus automatic processing (Daw, Niv, & Dayan, 2005; Dickinson, 1985; Kahneman & Frederick, 2002; Loewenstein & O'Donoghue, 2004). This distinction is thought to be of considerable practical importance, for instance, as a possible substrate for compulsion in drug abuse (Everitt & Robbins, 2005) and other disorders of self-control (Loewenstein & O'Donoghue, 2004).

However, one challenge for investigating such a division of labor experimentally is that, in typical formulations, it is often unclear which system produced a given behavior, and the contributions of each system can often be conclusively distinguished only by procedures that are both laborious and theory dependent (Dickinson & Balleine, 2004; Gläscher, Daw, Dayan, & O'Doherty, 2010). Moreover, although different theories share a common rhetorical theme, there is less consensus as to the fundamental, defining characteristics of the two systems, which makes it a challenge to relate data grounded in different models' predictions. One particularly large gap in this regard is between research in human and animal cognitive psychology. Human research is typically grounded in a distinction between procedural versus explicit learning and elucidated by manipulating factors such as working memory (WM) load (Foerde, Knowlton, & Poldrack, 2006; Zeithamova & Maddox, 2006). More invasive animal research has traditionally been conducted on parallel brain structures for instrumental learning (Dickinson &

Corresponding Author:
A. Ross Otto, Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003
E-mail: rotto@nyu.edu

Balleine, 2004; Yin & Knowlton, 2006) and has usually involved two-stage learning-and-transfer paradigms, such as latent learning or reward devaluation. This latter domain has been of recent interest to human cognitive neuroscientists because of the close relationship between traditional associative-learning models and the reinforcement-learning algorithms that have been used to characterize activity in dopaminergic systems in both humans and animals (temporal-difference learning; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Schultz, Dayan, & Montague, 1997).

For these reasons, reinforcement-learning theories may provide new leverage for reframing and formalizing the dual-system distinction in a manner that spans both animal and human traditions. One contemporary theoretical framework leverages the distinction between two families of reinforcement-learning algorithms: model-based and model-free reinforcement learning (Daw et al., 2005). Temporal-difference-based theories posit that the dopamine system is model free in the sense that it directly learns preferences for actions using a principle of repeating reinforced actions (akin to Thorndike's law of effect) without ever explicitly learning or reasoning about the structure of the environment. In model-based reinforcement learning, by contrast, the system learns an internal "model" of the proximal consequences of actions in the environment (such as the map of a maze) in order to prospectively evaluate candidate choices. This algorithmic distinction closely echoes theories of instrumental conditioning in animals (Dickinson, 1985), but the computational detail of Daw et al.'s (2005) framework leads to relatively specific predictions that afford clear identification of each system's contribution to choice behavior.

Consistent with prior work suggesting the parallel operation of distinct valuation systems (Dickinson & Balleine, 2004), previous research found that people appear to exhibit a mixture of both strategies in their choice patterns (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). However, it remains to be seen whether these two forms of choice behavior reflect any of the characteristics associated with controlled and automatic processing in human cognitive neuroscience and, even more fundamentally, whether they really capture distinct and separable processes. Underlining the question, recent functional MRI (fMRI) work unexpectedly revealed overlapping neural signatures of the two strategies (Daw et al., 2011).

To investigate these questions, we paired the multistep choice paradigm of Daw and colleagues (2011; Fig. 1) with a demanding concurrent task manipulation designed to tax WM resources. It has been demonstrated that concurrent WM load drives people away from explicit or rule-based systems toward reliance on putatively implicit systems in perceptual categorization (Zeithamova & Maddox, 2006), probabilistic classification (Foerde et al., 2006), and simple prediction (Otto, Taylor, & Markman, 2011). Contemporary theories differentiating model-based versus model-free reinforcement learning hypothesize that increased demands on central executive resources influence the trade-off between the two systems because model-based strategies involve planning processes that putatively draw on executive resources (Norman & Shallice, 1986), whereas model-free strategies simply apply the parsimonious principle of repeating previously rewarded actions (Daw et al., 2005; Dayan, 2009).

In Experiment 1, we utilized a within-subjects design in which some trials of the choice task were accompanied by a numerical Stroop task that has been demonstrated to displace explicit processing resources in perceptual category learning (Waldron & Ashby, 2001). We hypothesized that if learning, planning, or both in a model-based system is constrained by the availability of central executive resources, then choice behavior on these trials should, selectively, reflect reduced model-based contributions and increased model-free contributions. As a corollary, we predicted that response times (RTs)—a widely used index of cognitive cost (Payne, Bettman, & Johnson, 1993)—would be slower on trials in which model-based influence was prevalent in participants' choices than on trials in which choice appeared relatively model free. To further highlight model-based choice's dependence on central executive resources, we conceptually replicated this phenomenon in Experiment 2.

Experiment 1

Method

Our experimental procedure is described in detail in this section. Readers seeking an intuitive understanding of the task and our predictions are encouraged to advance to the Results section.

Participants. A total of 43 undergraduates at the University of Texas participated in Experiment 1 in exchange for course credit and were paid 2.5¢ per rewarded trial to incentivize choice. The data of 25 participants were used in analyses (participant inclusion criteria are detailed in the Supplemental Material available online).

Materials and procedure. Participants performed 300 trials of the two-stage reinforcement-learning task (Fig. 1a); on 150 of these trials (WM-load trials), the task was accompanied by a numerical Stroop task. These WM-load trials were positioned randomly, but with the

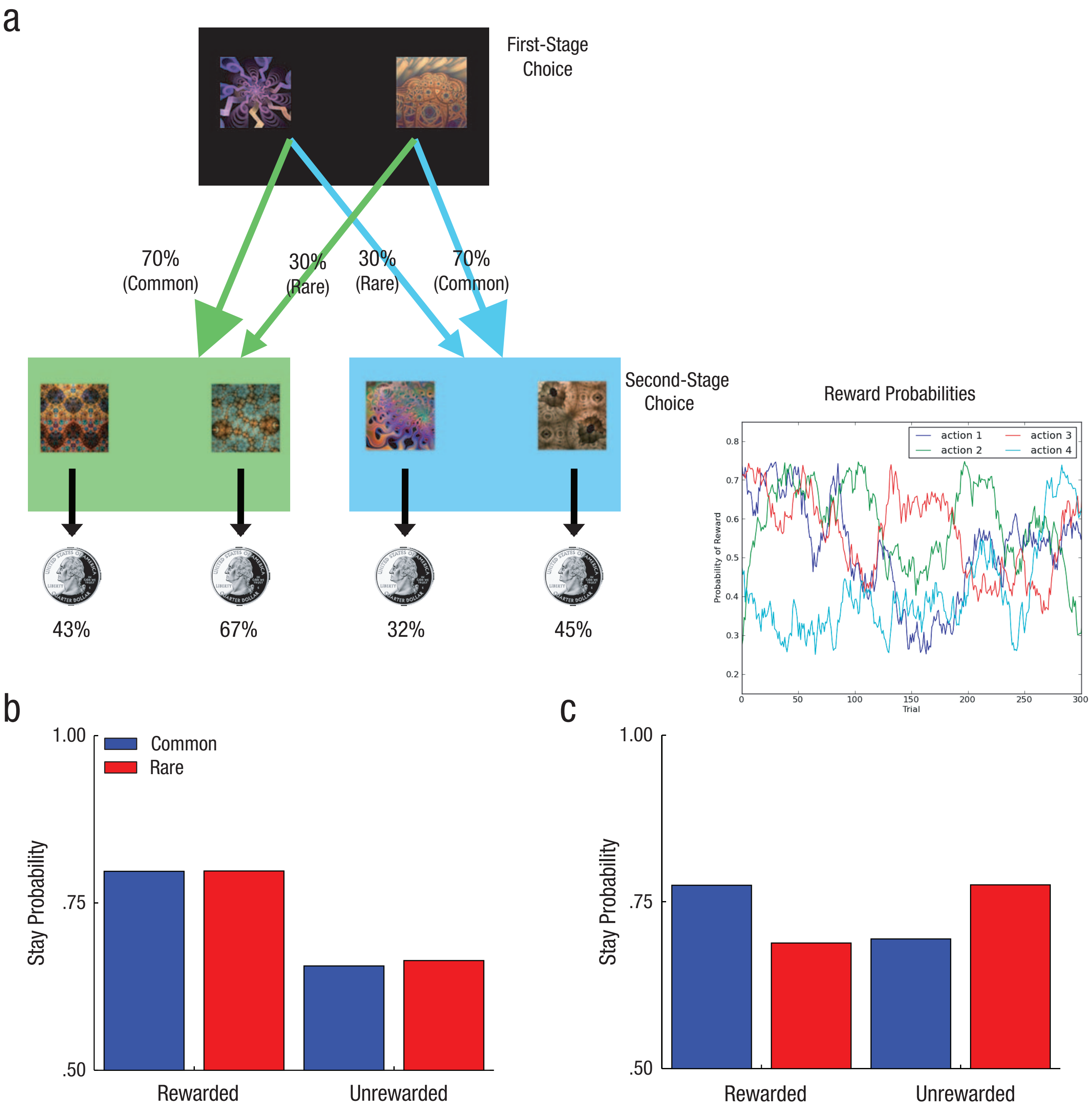

Fig. 1. Paradigm and model predictions for the two-stage choice task. In the multistep choice paradigm (a), participants are presented with two options in the first stage and asked to choose one. Each option has a 70% probability of leading to one of two second stages and a 30% probability of leading to the other. In the second stage, participants again have to choose between two options, each of which is associated with different probabilities of reward. The graphs (b and c) depict the predicted probability of repeating a first-stage action in the second stage ("stay probability") as a function of whether that choice was rewarded or unrewarded and whether the transition from the first-stage state to the second-stage state in the previous trial was common (70% probability) or rare (30% probability). Under a model-free choice strategy (b), a first-stage choice resulting in reward is more likely to be repeated on the subsequent trial regardless of whether that reward occurred after a common or a rare transition. Under a model-based choice strategy (c), rewards after rare transitions should affect the value of the unchosen first-stage option, thus leading to a predicted interaction between the factors of reward and transition probability. Panels (b) and (c) are adapted from "Model-Based Influences on Humans' Choices and Striatal Prediction Errors," by Daw, Gershman, Seymour, Dayan, and Dolan, 2011, *Neuron, 69*, p. 1206. Copyright 2011 by Elsevier.

constraint that the ordering would yield equal numbers of three trial types of interest (50 each for Lag 0, Lag 1, and Lag 2 trials, with lag defined by the number of trials since the most recent WM-load trial; see the Results section for more details). Participants were instructed to perform the WM task as well as possible and to make choices with whatever cognitive resources they had remaining (i.e., "with what was left over"). After being familiarized with the reinforcement-learning task's structure and goals, they were given 15 practice WM-load trials to familiarize them with the response procedure.

The reinforcement-learning task followed the same general procedure in both no-WM-load and WM-load trials (see Fig. 2 for a timeline). In the first step of no-WM-load trials, two fractal images appeared side by side on a black background, and participants had 2 s to choose between the left- or right-hand image using the "Z" or "?" key, respectively. After a choice was made, the selected image was highlighted for the remainder of the response period, and the background color changed according to which second-stage state the participant had been transitioned to. The second-stage state could be either common (70% probability) or rare (30% probability). After the transition, the image selected in the first stage was minimized and moved to the top of the screen. Two different fractal images were then displayed, and participants again had 2 s to choose one. The selected action was highlighted for the remainder of the response period. Then, either a picture of a quarter (indicating that they had been rewarded on that trial) or the number zero (indicating that they had not been rewarded on that trial) was shown. The reward probabilities associated with second-stage actions were governed by independently drifting Gaussian random walks (*SD* = 0.025) with reflecting boundaries at 0.25 and 0.75. Mappings of actions to stimuli and transition probabilities were randomized across participants.

WM-load trials followed the same procedure, except that participants additionally had to perform a numerical Stroop task, which required them to remember which of two numbers was physically and numerically larger (Waldron & Ashby, 2001; Fig. 2). These trials were signaled in two ways. First, during the 1-s intertrial interval preceding the first stage, participants were warned with the message "WATCH FOR NUMBERS." Second, during both stages of the choice task, the screen was outlined in red. At the beginning of the first stage, two digits were presented for 200 ms above and to the left and right, respectively, of the choice stimuli; they were then covered by a white mask for another 200 ms. After second-stage reward feedback was provided, either the word "VALUE" or "SIZE" appeared alone on a black screen, and there was a 1-s response period in which participants used the "Z" or "?" key to indicate whether the number with the larger value or larger size appeared on the left or the right side of the screen, respectively, during the first stage. Their response was followed by 1 s of feedback ("CORRECT" or "INCORRECT") and then an intertrial interval. If participants failed to choose one of the images in either response stage or in the numerical Stroop task, a red "X" appeared for 1 s to indicate that their response was too slow, and the trial was aborted. Crucially, the trial lengths were equated across WM-load and no-WM-load trials.

Results

Participants performed 300 trials of a two-stage reinforcement-learning task (Fig. 1a). In each two-stage trial, people made an initial first-stage choice between two options (depicted as fractals), which probabilistically led to one of two second-stage "states" (colored green or blue). In each of these states, participants made another choice between two options, which were associated with different probabilities of monetary reward. Each of the first-stage responses usually led to a particular second-stage state (70% of the time) but sometimes led to the other second-stage state (30% of the time). Because the second-stage reward probabilities independently changed over time, decision makers needed to make trial-by-trial adjustments to their choice behavior in order to effectively maximize payoffs.

Model-based and model-free strategies make qualitatively different predictions about how second-stage rewards influence first-stage choices on subsequent trials. For example, consider a first-stage choice that results in a rare transition to a second stage wherein the second-stage choice was rewarded. Under a pure model-free strategy—by virtue of the reinforcement principle—one would repeat the same first-stage response in the following trial because it ultimately resulted in reward. In contrast, a model-based choice strategy, utilizing a model of the transition structure and immediate rewards to prospectively evaluate the first-stage actions, would predict a decreased tendency to repeat the same first-stage option because the other first-stage action would actually be more likely to lead to that second-stage state.

These patterns of dependency of choices on the previous trial's events can be distinguished by a two-factor analysis of the effect of the previous trial's reward (rewarded vs. unrewarded) and transition type (common vs. rare) on the first-stage choice in the current trial.[1] The predicted choice pattern for a pure model-free strategy and a pure model-based strategy are depicted in Figures 1b and 1c, respectively, derived from model simulations (Daw et al., 2011; see the Reinforcement-Learning Model section in the Supplemental Material). A pure model-free strategy predicts only a main effect of reward,

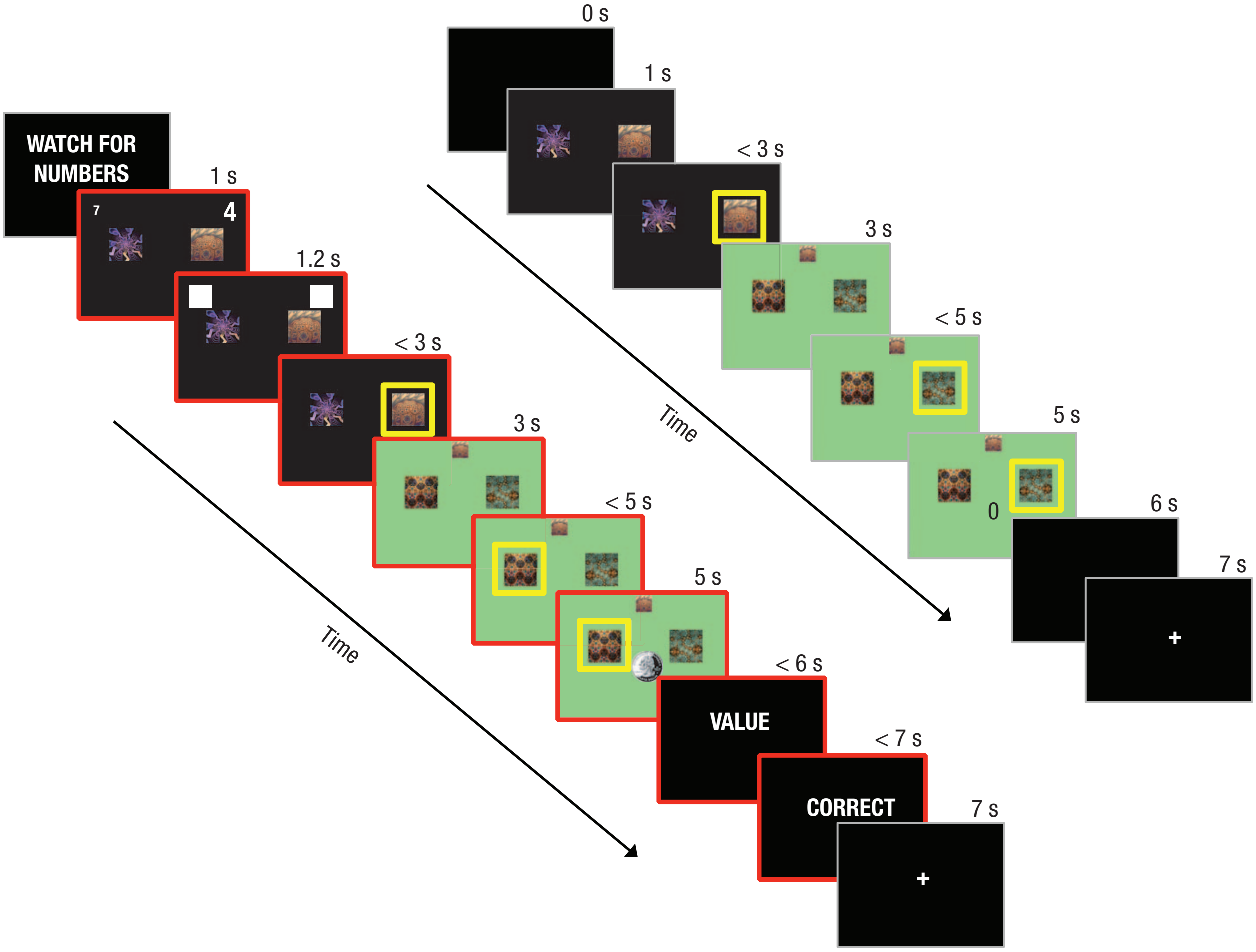

Fig. 2. Timeline of events in WM-load trials (left) and no-WM-load trials (right) in Experiment 1. The first stage of no-WM-load trials began with a blank screen, and then two fractal images appeared side by side on a black background. Participants had 2 s to choose between the two images, after which their selection was highlighted for the remainder of the response period. The first stage then transitioned to the second-stage state, which was signaled by a change in background color. The image selected in the first stage shrank and moved to the top of the screen, and two new images appeared. Participants again had 2 s to choose an image, and their response was highlighted for the remainder of the response period. Then, either a picture of a quarter (shown in the timeline on the left) or the number zero (shown in the timeline on the right) appeared to indicate that they had either been rewarded or not been rewarded, respectively, on that trial. This was followed by a blank screen and a fixation cross. WM-load trials followed the same general procedure as no-WM-load trials, with the following differences. They began with a cue that these trials would include a numerical Stroop task. Two different numbers of different physical sizes then appeared above the choice stimuli in the first stage for 200 ms and were subsequently covered by white masks. After second-stage reward feedback was provided, either the word "VALUE" (shown here) or "SIZE" appeared alone on a black screen, and participants had to indicate whether the number with the larger value or the larger size, respectively, had appeared on the left or the right during the first stage by pressing one of two keys. Feedback was given for 1 s. WM-load trials were highlighted in red throughout. Critically, event timing was equated between the two trial types.

whereas a model-based strategy predicts a full crossover interaction between reward and transition type because transition probabilities are taken into account. Following Daw et al. (2011), we factorially examined the impact that both the transition type and reward on the previous trial had on participants' tendency to repeat the same first-stage choice on the current trial. To examine the relationship between these signatures of choice strategies and the concurrent WM-load manipulation, we crossed these factors with a third factor defining the position of the most recent WM-load trial relative to the current trial. We sorted trials according to when the most recent WM-load trial had occurred relative to the current trial, which yielded three trial types of interest: Lag 0, Lag 1, and Lag 2, which refer to trials in which WM load occurred on the current trial, the previous trial, or the trial preceding the previous trial, respectively. Trials in which WM load occurred more than once across the current trial and its two predecessors did not fall into any of these categories and were excluded from analysis.

Strategy as a function of concurrent WM load. We hypothesized that if WM load interferes with model-based decision making, behavior on Lag 0 trials should be consistent with model-free decision making (Fig. 1b) because participants do not have the cognitive resources to carry out a model-based strategy on those trials. Conversely, we hypothesized that behavior on Lag 2 trials would reflect a mixture of both model-based and model-free strategies—mirroring the results of Daw and colleagues' (2011) study—because these trials involved no WM load either on the current trial or on the preceding trial, and thus participants could bring their full cognitive resources to bear on these trials. We reasoned further that if WM load disrupts participants' ability to integrate information crucial for model-based choice, then behavior on Lag 1 trials should appear model free (mirroring behavior on Lag 0 trials). In contrast, if participants are able to integrate this information while under load and apply it on the subsequent trial, then behavior on Lag 1 trials should resemble a mixture of both strategies, mirroring behavior on Lag 2 trials.

As Figure 3a shows, the pattern of results on Lag 2 trials suggests that participants' choices on these trials reflect both the main effect of reward (characteristic of model-free reinforcement learning) and its interaction with the rare or common transition (characteristic of model-based reinforcement learning); this pattern is consistent with the single-task results obtained by Daw et al. (2011). In contrast, choices on Lag 0 and Lag 1 trials (Figs. 3b and 3c) appear sensitive only to reward on the previous trial and not to the transition type. Qualitatively, these choice patterns resemble a pure model-free strategy, which suggests that WM load interferes with model-based choice.

To quantify these effects of WM load on choice behavior, we conducted a mixed-effects logistic regression

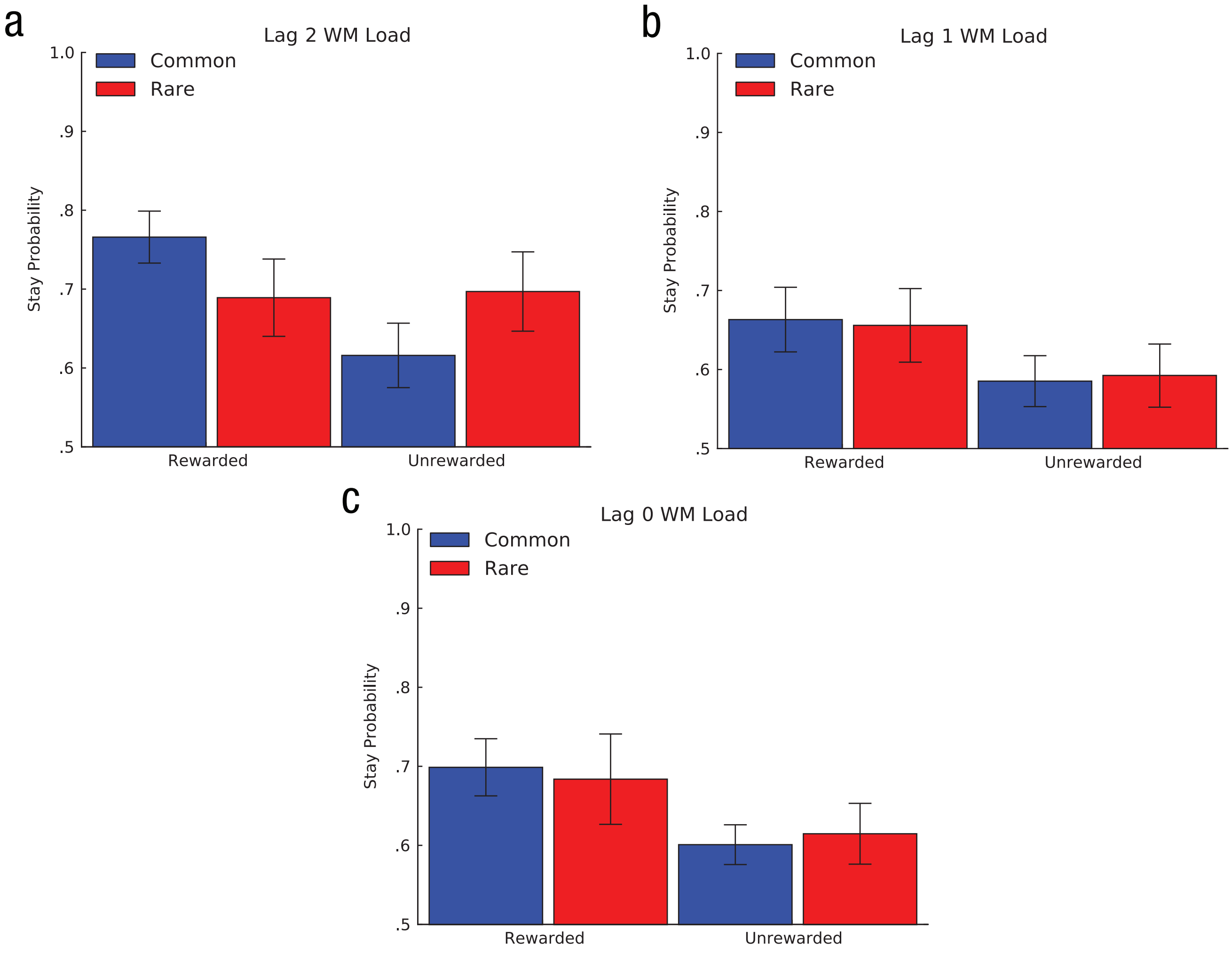

Fig. 3. Results from Experiment 1: average proportion of trials on which participants chose to stay with the response they selected in the first stage of the previous trial as a function of whether they received a reward on the previous trial and whether the second-stage state transitioned to on the previous trial was common or rare. Results are shown separately for (a) Lag 2 trials, (b) Lag 1 trials, and (c) Lag 0 trials. Lag 0, Lag 1, and Lag 2 trials were those in which working memory (WM) load was taxed on the present trial, the previous trial, and the trial preceding the previous trial, respectively. Error bars depict standard errors of the mean.

(Pinheiro & Bates, 2000) to explain the first-stage choice on each trial *t* (coded as stay vs. switch) using binary predictors indicating whether reward was received on *t* – 1 and the transition type (common or rare) that had produced it. Further, we estimated these factors under each trial type—Lag 0, Lag 1, and Lag 2, represented by binary indicators—and, to capture any individual differences, specified all coefficients as random effects over participants. The full regression specification and coefficient estimates are reported in Table 1.

We found a significant main effect of reward for each trial type (*p*s < .05), which indicates that participants had a general tendency to repeat rewarded first-stage responses, consistent with intact use of a model-free strategy. This finding also suggests that concurrent task demands did not produce trivially random or otherwise unstructured behavior. However, we found a significant three-way interaction between Lag 2, reward, and transition type ($p < .05$), which suggests that the interaction characteristic of a model-based choice strategy was evident in Lag 2 trials, as hypothesized. Neither the interactions among Lag 0, reward, and transition type nor among Lag 1, reward, and transition type were significant, which indicates that this model-based interaction was not present in these trial types (*p*s > .25).

To examine whether these differences between trial types were themselves significant, we conducted a planned contrast on the Lag 2 three-way interaction (Lag 2 × Reward × Transition Type, indicative of model-based learning). This interaction was significantly larger than the same interactions at both the Lag 1 and Lag 0 levels ($p < .05$). Further, we found no differences in model-free behavior between any of the trial types (e.g., Lag 0 × Reward, Lag 1 × Reward, and Lag 2 × Reward) that we considered (*p*s > .30). All of these results are consistent with the hypothesis that concurrent demands selectively interfere with model-based learning and planning while sparing model-free decision making. (For analyses of second-stage choice behavior and secondary task performance, refer to the Supplemental Materials.)

Table 1. Results of the Logistic Regression Investigating the Influence of Working-Memory-Load Lag, Previous Outcome, and Previous Transition Type on First-Stage Response Repetition in Experiment 1

Predictor	Estimate	*p*
Intercept	1.00 (0.18)	< .0001
Lag 0	–0.23 (0.14)	.118
Lag 1	–0.43 (0.12)	< .0001
Lag 0 × Reward	0.34 (0.13)	.010
Lag 1 × Reward	0.19 (0.09)	.031
Lag 2 × Reward	0.23 (0.12)	.044
Lag 0 × Transition Type	0.07 (0.09)	.434
Lag 1 × Transition Type	–0.07 (0.08)	.390
Lag 2 × Transition Type	0.02 (0.09)	.776
Lag 0 × Reward × Transition Type	0.06 (0.09)	.478
Lag 1 × Reward × Transition Type	–0.07 (0.08)	.383
Lag 2 × Reward × Transition Type	–0.23 (0.09)	.011

Note: Standard errors are given in parentheses. Lag 0, Lag 1, and Lag 2 refer to trials in which working memory load occurred on the current trial, the previous trial, or the trial preceding the previous trial, respectively. Previous outcome refers to whether the participant was rewarded on the previous trial. Transition type refers to whether the transition from the first stage to the second stage in the previous trial was common or rare.

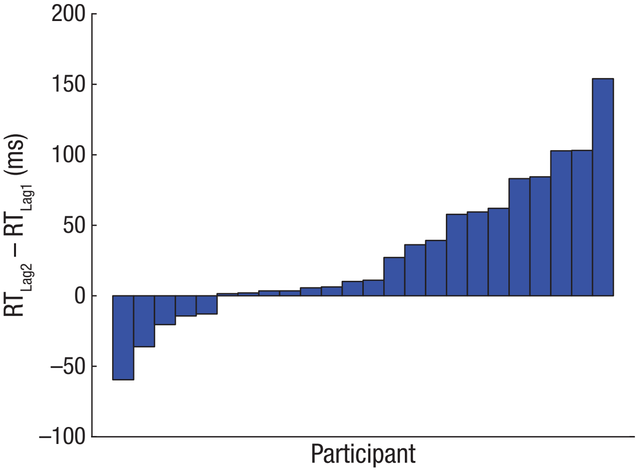

Fig. 4. Results from Experiment 1: difference in median response times (RTs) between Lag 2 and Lag 1 trials for individual participants.

Choice RTs. We also predicted that model-based choice, by virtue of its hypothesized cognitive costs, would incur larger RTs at the first-stage choice than model-free choices would (Keramati, Dezfouli, & Piray, 2011). We compared Lag 2 trials (in which behavior reflected the influence of a model-based strategy) with Lag 1 trials (in which behavior appeared to reflect only a model-free strategy). The comparison between the two single-task trial types that exhibited different degrees of model usage provided a clean test of the hypothesis: In Lag 0 trials, the RTs were confounded by the demands of the concurrent task itself. A mixed-effects linear model (see the Choice and RT Analyses section in the Supplemental Material) carried out on first-stage RTs revealed that participants exhibited significantly larger RTs on Lag 2 choices than on Lag 1 choices (Fig. 4; $\beta = 2.05$, $p < .05$), which suggests that model-based choice—evident on Lag 2 trials—indeed bore the signature of a cognitively costly process. Put another way, choice was faster on Lag 1 trials—where behavior appeared model free—which supports the notion that the process governing choice on those trials was cognitively less expensive.

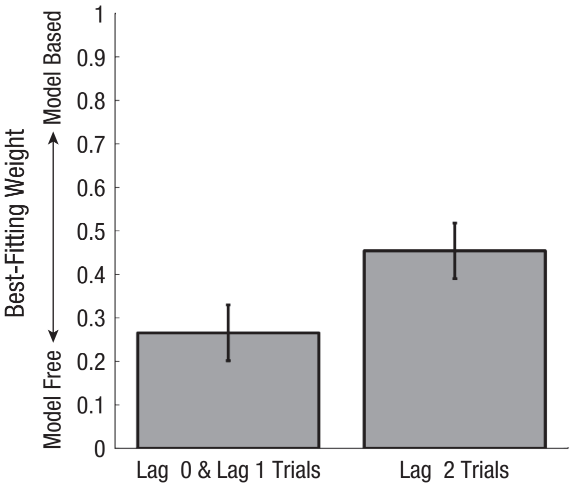

Fig. 5. Results from Experiment 1: best-fitting mixing weights across Lag 2 versus Lag 0 and Lag 1 trials (combined) resulting from fitting the reinforcement-learning algorithm to subjects' choices. Weights closer to 0 indicate more model-free control, whereas weights closer to 1 indicate more model-based control. Error bars indicate standard errors.

Reinforcement-learning model. One limitation of the foregoing regression analysis is that it only accounted for the influence of reinforcement occurring on the immediately preceding trial. Most reinforcement-learning models, in contrast, posit a decaying influence of all previous trials. We extended our regression analysis by fitting a dual-system reinforcement-learning model—a computational instantiation of the principles governing two hypothesized choice systems (Daw et al., 2011; Gläscher et al., 2010)—to behavior in this task. This model consists of a model-free system that updates estimates of choice values using temporal-difference learning and of a model-based system that learns a transition-and-reward model of the task and uses these to compute choice values on the fly (see Reinforcement-Learning Model in the Supplemental Material). The values are linearly mixed according to a weight parameter that determines the balance between model-free and model-based control—weights closer to 0 indicate model-free control, whereas weights closer to 1 indicate model-based control. The mixed value is then used to generate choices according to a softmax rule (Sutton & Barto, 1998). To accommodate the present paradigm, we fit two separate mixing weights: one for Lag 0 and Lag 1 trials (combined) and one for Lag 2 trials only. We found that Lag 2 weights were significantly larger than the Lag 0 and Lag 1 weights (Fig. 5), $t(24) = 2.94$, $p < .01$; this suggests that participants' behavior was more model-based at longer lags and corroborates the results of the regression analysis.

Experiment 2

Because the within-subjects WM-load manipulation we utilized in Experiment 1 was rather intricate and novel, we sought to provide a between-subjects replication of the study using a separate WM-load manipulation in which one group of participants counted auditory tones while performing the same choice task as in Experiment 1 (Foerde et al., 2006). In brief, we found that the behavior exhibited by single-task participants in the current experiment resembled the mixture of strategies observed in Lag 2 trials in the previous experiment, whereas the behavior of dual-task participants in the current experiment resembled the model-free pattern of choice observed in Lag 0 and Lag 1 conditions in the previous experiment (Fig. 6; Table 2; see the Supplemental Material for details of Experiment 2).

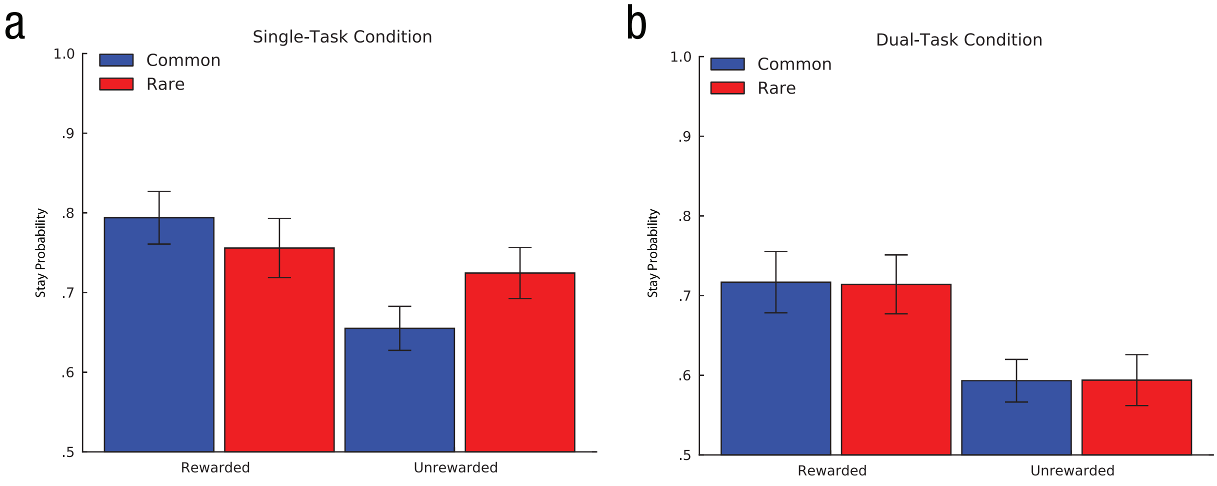

Fig. 6. Results from (a) the single-task condition and (b) the dual-task condition of Experiment 2: average proportion of trials on which participants chose to stay with the response they selected in the first stage of the previous trial as a function of whether they received a reward on the previous trial and whether the second-stage state transitioned to on the previous trial was common or rare.

Table 2. Results of the Logistic Regression Investigating the Influence of Working-Memory-Load Condition, Previous Outcome, and Previous Transition Type on First-Stage Response Repetition in Experiment 2

Predictor	Estimate	*p*
Intercept	1.15 (0.13)	< .0001
Load	–0.25 (0.13)	.058
Reward	0.42 (0.07)	.000
Transition	0.01 (0.03)	.823
Load × Reward	0.01 (0.07)	.824
Load × Transition	–0.02 (0.03)	.433
Reward × Transition Type	–0.11 (0.04)	.005
Load × Reward × Transition Type	0.08 (0.04)	.047

Note: Standard errors are given in parentheses. Working memory load was manipulated by having each participant perform either single-task or dual-task trials. Previous outcome refers to whether the participant was rewarded on the previous trial. Transition type refers to whether the transition from the first stage to the second stage in the previous trial was common or rare.

General Discussion

A number of dual-system accounts of choice behavior posit a distinction between two systems distinguished by, among other things, the extent to which central executive or prefrontal resources are employed (Dickinson & Balleine, 2004; McClure, Laibson, Loewenstein, & Cohen, 2004). Still, the contributions of the two putative systems have proven laborious to isolate behaviorally (Valentin, Dickinson, & O'Doherty, 2007) or with neuroimaging (Daw et al., 2011). Informed by a contemporary theoretical framework that makes quantitative predictions about the behavioral signatures of the two systems and the arbitration of behavioral control among the two (Daw et al., 2005), we demonstrated how human decision makers trade off the concurrent cognitive demands of the environment with their usage of computationally expensive choice strategies. In particular, when burdened with concurrent WM load, decision makers relied on a pure reinforcement-based strategy—akin to model-free reinforcement learning—and eschewed the transition structure of the environment. When unencumbered by these demands, participants' choices reflected a mixture of model-based and model-free strategies, mirroring previous results (Daw et al., 2011).

The present results are evocative of past research revealing that concurrent cognitive demands shift the onus of learning from explicit, declarative systems to procedural-learning systems (Foerde et al., 2006). It is important to note that although previous work has revealed that concurrent demands can shift people's response strategies, these studies have relied on comparing results across multiple task methodologies chosen to favor either strategy (Waldron & Ashby, 2001; Zeithamova & Maddox, 2006) or post hoc assessments of declarative knowledge (Foerde et al., 2006). The two-step reinforcement-learning task used in the experiments reported here, in contrast, afforded unambiguous identification of the simultaneous contributions of model-based and model-free choice strategies within the same task and permitted dynamic assessment of trial-by-trial arbitration of control between the two systems. Here, accordingly, we present evidence of a difference in strategy use between trial types that occurred fully interleaved, consistent with rapid strategic switching within participants and task.

These results complement previous fMRI investigations using the present task—a previous finding of convergent neural correlates for the two strategies (Daw et al., 2011) left open the question of whether they were actually psychologically or functionally distinct. Here, our behavioral results provide a compelling demonstration that model-based and model-free valuation are dissociable, and these findings further underscore the utility of within-subjects manipulations for dissociating the behavioral contributions of putatively separate neural systems. Finally, the distinction as we operationalize it is arguably of more biological relevance than previous attempts, because the model-free strategy on which participants appeared to fall back under WM load was exactly that predicted by prominent neurocomputational accounts of the dopamine system (Montague, Dayan, & Sejnowski, 1996).

It is also worth noting that model-based choice relies on at least two constituent processes: (a) learning of second-stage reward probabilities and environment-transition probabilities from feedback and (b) planning by using these reward probabilities and environment-transition probabilities prospectively to inform first-stage choice on subsequent trials (Sutton, 1990). Insofar as the learning relevant to the choice on Trial *t* occurs on earlier trials (and, specifically, for the effects quantified here on the preceding trial, *t* – 1), but the planning occurs on the trial itself, we might expect WM load occurring at Lag 1 (i.e., on trial *t* – 1) to primarily affect learning and WM load at Lag 0 (Trial t) to primarily affect planning. By this logic, our finding of a similar strategic deficit at both lags may suggest that WM load disrupted both putative subprocesses. That said, it is possible that these processes are not as temporally isolated as we ascribe (e.g., action planning on Trial *t* may begin as soon as the feedback is received on the preceding trial) or that results also reflect other executive demands not isolated to a single trial (e.g., switching between dual and single tasks from *t* – 1 to *t*), making this interpretation tentative. Future work should aim to disambiguate more precisely whether concurrent executive demands incapacitate planning, learning, or some combination thereof, perhaps by using more specifically directed distractor tasks.

Although the model-based strategy we observed in the Lag 2 trials was, by definition, not predicted by a model-free reinforcement-learning system of the sort associated with the dopamine system, it is clearly possible to produce model-free switching (win-stay-lose-shift) via a deliberative or explicit strategy. Indeed, this is the question that the present manipulation was designed to address, and the finding that the model-free, but not the model-based, behavior is robust to concurrent load is consistent with the prediction that it arises from a distinct, striatal procedural-learning system that itself is also model free. Still, it is possible in principle that load promotes a shift to increased reliance on a cheaper—but still declarative in nature—win-stay-lose-shift strategy. However, the best-fitting learning rates recovered in our computational modeling (see Table S1 in the Supplemental Material) were low,[2] which supports the idea that these influences arose from an incremental-learning process characteristic of implicit learning rather than a rule-based win-stay-lose-shift strategy.

Whereas Daw and colleagues (2011) relied in part on individual differences in model-based choice to examine the two systems' neural substrates, we explicitly manipulated reliance on these strategies within subjects and within tasks. As it is well documented that there are considerable individual differences in WM capacity and executive function (Conway, Kane, & Engle, 2003; Miyake et al., 2000), a significant portion of the individual variability reported by Daw and colleagues may be attributable to individual differences in WM capacity, and likewise, these differences could have potentially modulated the effects of WM load reported here. Exactly how individual limitations in cognitive capacity, executive control, or a combination of the two modulate model-based choice warrants additional examination. Further, characterizing more precisely how humans balance the contributions of model-based and model-free choice is of considerable practical importance because contemporary accounts of a number of serious disorders of compulsion ascribe this behavior to abnormal expression of habitual or stimulus-driven control systems (Everitt & Robbins, 2005; Loewenstein & O'Donoghue, 2004).

Acknowledgments

We gratefully acknowledge Jeanette Mumford and Bradley Doll for helpful conversations and Grant Loomis for assistance with data collection.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by National Institute of Mental Health Grant MH077708 to Arthur B. Markman. A. Ross Otto was supported by a Mike Hogg Endowment Fellowship from the University of Texas. Samuel J. Gershman was supported by a Graduate Research Fellowship from the National Science Foundation. Nathaniel D. Daw was supported in part by National Institute of Neurological Disorders and Stroke Grant R01 NS 078784, a Scholar Award from the McKnight Foundation, and an Award in Understanding Human Cognition from the McDonnell Foundation.

Supplemental Material

Additional supporting information may be found at http://pss.sagepub.com/content/by/supplemental-data

Notes

1. In general, reinforcement-learning models predict that a given trial's choice depends on learning also from even earlier trials (and in the present study, we used fits of these models to verify that our results held when these longer-term dependencies were accounted for). However, because the most recent trial exerted the largest effect on choice in these models (and this effect becomes exclusive as free-learning-rate parameters approach 1), this factorial analysis provided a clear picture of the critical qualitative features of behavior less dependent on the specific parametric and structural assumptions of the full models.

2. Further, we fitted a separate model that allowed for different learning rates across the three trial types of interest (Lag 0, Lag 1, and Lag 2) and found that learning rates did not vary significantly as a function of WM-load lag, $F = 0.83$, $p = .44$.

References

Conway, A. R. A., Kane, M. J., & Engle, R. W. (2003). Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences*, *7*, 547–552.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204–1215.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704–1711.

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks*, *22*, 213–219.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *308*, 67–78.

Dickinson, A., & Balleine, B. (2004). The role of learning in the operation of motivational systems. In R. Gallistel (Ed.), *Stevens' handbook of experimental psychology: Vol. 3. Learning, motivation, and emotion* (3rd ed.). Hoboken, NJ: Wiley.

Everitt, B. J., & Robbins, T. W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. *Nature Neuroscience*, *8*, 1481–1489.

Foerde, K., Knowlton, B. J., & Poldrack, R. A. (2006). Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences, USA*, *103*, 11778–11783.

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*, 585–595.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge, England: Cambridge University Press.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*(5), e1002055. Retrieved from http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1002055

Loewenstein, G., & O'Donoghue, T. (2004). *Animal spirits: Affective and deliberative processes in economic behavior* (Working Papers No. 04–14). Ithaca, NY: Cornell University Center for Analytic Economics. Retrieved from http://ideas.repec.org/p/ecl/corcae/04-14.html

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*, 503–507.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience*, *16*, 1936–1947.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation: Advances in research and theory* (Vol. 4, pp. 1–18). New York, NY: Plenum.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*, 329–337.

Otto, A. R., Taylor, E. G., & Markman, A. B. (2011). There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition*, *118*, 274–279.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge, England: Cambridge University Press.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY: Springer.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Sutton, R. S. (1990). Integrated architecture for learning, planning, and reacting based on approximating dynamic programming. In M. B. Morgan (Ed.), *Proceedings of the Seventh International Conference (1990) on Machine Learning* (pp. 216–224). San Francisco, CA: Morgan Kaufmann.

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

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, *27*, 4019–4026.

Waldron, E. M., & Ashby, F. G. (2001). The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review*, *8*, 168–176.

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, *7*, 464–476.

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*, 387–398.