Published in final edited form as:
Psychol Sci. 2013 May ; 24(5): . doi:10.1177/0956797612463080.

The Curse of Planning: Dissecting multiple reinforcement learning systems by taxing the central executive

A. Ross Otto,
University of Texas at Austin

Samuel J. Gershman,
Princeton University

Arthur B. Markman, and
University of Texas at Austin

Nathaniel D. Daw
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. Along these lines, 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. Based on the hypothesis that model-based reinforcement learning requires cognitive resources, we demonstrate that having human decision-makers perform a demanding secondary task engenders increased reliance on a model-free reinforcement learning strategy. Further, we show that across trials, people negotiate this tradeoff dynamically as a function of concurrent executive function demands and their choice latencies reflect the computational expenses of the strategy employed. 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.

Accounts of decision-making across cognitive science, neuroscience, and behavioral economics posit that decisions arise from two qualitatively distinct systems, which 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, on typical formulations, most behaviors are ambiguous as to which system produced them, and their contributions can often only be conclusively distinguished by procedures that are both laborious and theory-dependent (Dickinson & Balleine, 2002; Gläscher et al., 2010). Moreover, although different theories share a common rhetorical theme, there is less consensus as to what are the fundamental, defining characteristics of the two systems, making it a challenge to relate data grounded in different models' predictions. One

Please address all correspondence to: A. Ross Otto, Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003, rotto@nyu.edu.
ARO is now at Center for Neural Science, New York University. SJG is now at Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology.

particularly large gap in this regard is between research in human cognitive psychology, which is typically grounded in a distinction between procedural versus explicit learning and elucidated using manipulations such as working memory (WM) load (Foerde, Knowlton, & Poldrack, 2006; Zeithamova & Maddox, 2006) and another tradition of more invasive animal research on parallel brain structures for instrumental learning (Dickinson & Balleine, 2002; Yin & Knowlton, 2006), usually investigated with two-stage learning/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 (RL) algorithms that have been used to characterize activity in dopaminergic systems in both humans and animals (temporal-difference learning, TD; O'Doherty et al., 2003; Schultz, Dayan, & Montague, 1997).

For these reasons, RL 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 RL algorithms: model-based and model-free RL (Daw et al., 2005). TD-based theories of the dopamine system are model-free in the sense that they directly learn 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. Model-based RL, by contrast, 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. (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, 2002), people appear to exhibit a mixture of the signatures of both strategies in their choice patterns (Daw et al., 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 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; Figure 1) with a demanding concurrent task manipulation designed to tax WM resources. Concurrent WM load has been demonstrated to drive people away from explicit or rule-based systems towards 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 RL hypothesize that increased demands on central executive resources influence the tradeoff between the two systems because model-based strategies involve planning processes that putatively draw upon 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). We hypothesized that if learning and/or planning in a model-based system were 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.

Experiment 1 utilizes a within-subject design in which some trials of the choice task were accompanied by a concurrent Numerical Stroop task that has been demonstrated to displace explicit processing resources in perceptual category learning (Waldron & Ashby, 2001). We hypothesized that if learning and/or planning 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—a widely used index of cognitive cost (Payne et al., 1993)—should be slower on trials in which model-based influence was prevalent in participants' choices compared to trials in which choice appears relatively model-free. To further highlight model-based choice's dependence on central executive resources, Experiment 2 provides a conceptual replication of this phenomenon.

Experiment 1

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

Participants

A total of 43 undergraduates at the University of Texas participated in exchange for course credit and were paid 2.5 cents per rewarded trial to incentivize choice. The data of 25 participants were used in analyses (participant exclusion criteria are detailed in the Supplemental Materials).

Materials and Procedure

Participants performed 300 trials of the two-step RL task (Figure 1A) accompanied by a concurrent Numerical Stroop task on 150 trials selected as WM-load trials. These WM-load trials were positioned randomly, but with the constraint that the ordering would yield equal numbers of the three trial types of interest (50 each). Participants were instructed to perform the WM task as well as possible and make choices with "with what was left over." After being familiarized with the RL task structure and goals, they were given 15 practice trials under WM-load to familiarize themselves with the response procedure.

The RL task followed the same general procedure in both trial types (see Figure 2 for a timeline). In the first step, two fractal images appeared on a black background (indicating the initial state), and there was a two-second response window in which participants could choose the left- or right-hand response using the "Z" or "?" keys respectively. After a choice was made, the selected action was highlighted for the remainder of the response window followed by the background color changing according to the second-stage state the participant had transitioned to. After the transition, the background color changed to reflect the second-stage state and the selected first-stage action moved to the top of the screen. Two fractal images, corresponding to the actions available in the second stage, were displayed and participants again had two seconds to make a response. The selected action was highlighted for the remainder of the response window. Then, either a picture of a quarter was shown (indicating that they had been rewarded that trial) or the number zero (indicating that they had not been rewarded 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.

On WM-load trials, participants additionally had to perform a numerical Stroop task, which required the participant to remember which of two numbers were physically and numerically larger (Waldron & Ashby, 2001; Figure 2). These trials were signaled in two ways. First, during the one-second inter-trial interval preceding the first stage, participants were warned

with the message “WATCH FOR NUMBERS.” Second, during both stages of the choice task on WM-load trials, the screen was outlined in red. At the beginning of the first-stage response window, two digits were presented for 200 ms above the response stimuli, followed by a white mask for another 200 ms. After second-stage reward feedback was provided, either the word “VALUE” or “SIZE” appeared on screen, and there was a one-second response window in which participants were to indicate the side of the screen on which the number with the larger value or larger size was presented. Participants used the “Z” or “?” keys to indicate the left and right side respectively. This was followed by one second of feedback (“CORRECT” or “INCORRECT”) followed by the inter-trial interval preceding the next trial. If the participant failed to make a choice in the response window of either response stage or the numerical Stroop judgment, a red X appeared for one second indicating 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-step RL task (Figure 1A). In each two-stage trial, people made an initial first-stage choice between two options (depicted as fractals), which probabilistically leads to one of two second-stage “states” (colored green or blue). In each of these states participants make another choice between two options, which were associated with different probabilities of monetary reward. One 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 change over time, decision-makers need 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 subsequent first-stage choices. For example, consider a first-stage choice that results in a rare transition to a second stage wherein that 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 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 was actually 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 versus unrewarded) and transition type (common versus rare) on the current trial’s first-stage choice[1]. The predicted choice pattern for a pure model-free strategy and a pure model based-strategy are depicted in Figures 1A and 1B, respectively, derived from model simulations (Daw et al., 2011, see Supplemental Materials). A pure model-free strategy predicts only a main effect of reward, while a full crossover interaction is predicted under a model-based strategy because transition probabilities are taken into account. Following Daw et al. (2011), we factorially examined the impact of both the transition type (common versus rare) and reward (rewarded versus not rewarded) on the previous trial upon participants’ tendency to repeat the same first-stage choice on the current choice. To examine the relationship between these signatures of choice strategies and the concurrent WM load manipulation

[1]In general, RL models predict that a trial’s choice depends on learning also from even earlier trials (and below we use fits of these models to verify that our results hold when these longer-term dependencies are accounted for). However, since in these models, the most recent trial exerts the largest effect on choice (and this effect becomes exclusive as free learning rate parameters approach 1), this factorial analysis provides a clear picture of the critical qualitative features of behavior less dependent on the specific parametric and structural assumptions of the full models.

(Figure 2), we crossed these factors with a third defining the position of the most recent WM-load trial relative to the current trial. We sorted trials according to where the most recent WM-load trial had occurred relative to the current trial, yielding three trial types of interest. Thus Lag-0, Lag-1, and Lag-2 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 had 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 appear model-free (Figure 1B), as 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 (Figures 1B and C)—mirroring the results of Daw and colleagues' (2011) study—as these trials involved no WM load either on the choice 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 Lag-0 trials). On the other hand, 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 Lag-2 trials.

Figure 3 plots participants' choices as a function of previous reward and transition type, broken down by WM condition. 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 RL) and its interaction with the rare or common transition (characteristic of model-based RL), consistent with the previous single-task result (Daw et al., 2011). In contrast, choices on Lag-0 and Lag-1 trials (Figures 3B and C) 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 (Figure 2A), suggesting 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 (Pinheiro & Bates, 2000) to explain the first-stage choice on each trial t (coded as stay versus switch) using binary predictors indicating if 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 subjects. The full regression specification and coefficient estimates are reported in Table 1.

We found a significant main effect of reward for each trial type (ps<.05), indicating that participants had a general tendency to repeat rewarded first-stage responses, consistent with spared use of a model-free strategy, and suggesting that the 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 (lag-2×reward×transition, p<.05) suggesting that the interaction characteristic of a model-based choice strategy was evident in Lag-2 trials as hypothesized. Neither interaction between Lag-0, reward, and transition type nor Lag-1, reward, and transition type were significant indicating that this model-based interaction was not present in these trial types (ps>.25).

To examine whether these differences between trial types were themselves significant, a planned contrast revealed that the size of the Lag-2 three-way interaction (lag-2×reward×transition, indicative of model-based learning) 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) considered ($ps>.30$). All of these results are consistent with the hypothesis that concurrent demand selectively interferes with model-based learning and/or planning while sparing model-free decision-making.

Choice Response Times

We also predicted that model-based choice, by virtue of its hypothesized cognitive costs, should incur larger RTs at the first-stage choice than model-free choices (Keramati et al., 2011). We compared Lag-2 trials (in which behavior reflected the influence of a model-based strategy; Figure 3A) with Lag-1 trials (in which behavior appeared only to reflect a model-free strategy; Figure 3B). The comparison between the two single-task trial types that exhibit different degrees of model usage provides a clean test of the hypothesis: in Lag-0 trials, the RTs are confounded by the demands of the concurrent task itself. A mixed-effects linear model (see Supplemental Materials) carried out on first-stage RTs revealed that participants exhibited significantly larger RTs on Lag-2 choices than on Lag-1 choices (Figure 4; $\beta=2.05, p<.05$), suggesting 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—where behavior appeared model-free—supporting the notion that the process governing choice on those trials was cognitively less expensive.

Reinforcement Learning Model

One limitation of the foregoing regression analysis is that it only considers the influence of reinforcement occurring on the immediately preceding trial. Most RL models, in contrast, posit a decaying influence of all previous trials. We extended our regression analysis by fitting a dual-system RL 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 TD learning, and 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 Supplemental Materials). 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-0/1 trials and one for Lag-2 trials. We found that Lag-2 weights were significantly larger than the Lag-0/Lag-1 weights (Figure 5; $t=2.94, p<.01$) suggesting that participants' behavior was more model-based at longer lags, corroborating the results of the regression analysis.

Experiment 2

Because this within-subjects WM load manipulation is 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 concurrently counted auditory tones (Foerde et al., 2006). In brief, we found that the behavior exhibited by Single-task participants resembled the mixture of strategies observed in Lag-2 trials while Dual-task participants would resemble the model-free pattern of choice observed in Lag-0 and Lag-1 conditions (Figure 6, Table 2, see Supplemental Materials for study details).

Discussion

A number of dual-systems 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, 2002; McClure et al., 2004). Still, the contributions of the two putative systems have proven laborious to isolate behaviorally (Valentin et al., 2007) or with neuroimaging (Daw et al., 2011). Informed by a contemporary theoretical framework which makes quantitative predictions about the behavioral signatures of the two systems and the arbitration of behavioral control amongst the two (Daw et al., 2005), we demonstrate 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 RL—eschewing 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 demand shifts the onus of learning from explicit/declarative systems to procedural learning systems (Foerde et al., 2006). It is important to note while previous work has revealed that concurrent demands can shift response strategies people employ, these studies rely 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 RL task, in contrast, affords unambiguous identification of model-based and model-free choice strategies' simultaneous contributions within the same task, and permits dynamic assessment of trial-by-trial arbitration of control between the two systems. Here, accordingly, we present evidence of for a difference in strategy use between trial types that occurred fully interleaved, consistent with rapid strategic switching within participant and task.

These results complement previous fMRI investigations using the present task, since a 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 result provides a compelling demonstration that model-based and model-free valuation are dissociable and further underscores 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, since the model-free strategy upon which participants appear to fall back under load is 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: 1) learning of second-stage reward probabilities and environment transition probabilities from feedback, and 2) planning, by using these reward probabilities and environment transition probabilities prospectively to inform subsequent first-stage choice (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 sub-processes. 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.

Finally, although the model-based strategy we observe in the lag-2 trials is, by definition, not predicted by a model-free RL 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 which 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 (Supplemental Table 1) were low[2], supporting 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.

While Daw and colleauges (2011) relied in part upon individual differences in model-based choice to examine the two systems' neural substrates, we were able to explicitly manipulate reliance upon these strategies within-subject and within-task. As it is well documented that there are considerable individual differences in WM capacity and/or 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 potentially modulate the effects of WM load reported here. Exactly how individual limitations in cognitive capacity and/or executive control 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).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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

References

Conway ARA, Kane MJ, Engle RW. Working memory capacity and its relation to general intelligence. Trends in Cognitive Sciences. 2003; 7(12):547–552. doi:10.1016/j.tics.2003.10.005. [PubMed: 14643371]

[2]Further, we fit a separate model that allowed for different learning rates across the 3 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=.83, p=.44.

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-Based Influences on Humans' Choices and Striatal Prediction Errors. Neuron. 2011; 69(6):1204–1215. doi:10.1016/j.neuron.2011.02.027. [PubMed: 21435563]

Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8(12):1704–1711. doi:10.1038/nn1560. [PubMed: 16286932]

Dayan P. Goal-directed control and its antipodes. Neural Networks. 2009; 22(3):213–219. doi:10.1016/j.neunet.2009.03.004. [PubMed: 19362448]

Dickinson A. Actions and Habits: The Development of Behavioural Autonomy. Philosophical Transactions of the Royal Society of London. B, Biological Sciences. 1985; 308(1135):67–78. doi:10.1098/rstb.1985.0010.

Dickinson, Anthony; Balleine, B. Stevens' Handbook of Experimental Psychology. 2002. The Role of Learning in the Operation of Motivational Systems.

Everitt BJ, Robbins TW. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nature Neuroscience. 2005; 8:1481–1489. doi:10.1038/nn1579.

Foerde K, Knowlton BJ, Poldrack RA. Modulation of competing memory systems by distraction. Proceedings of the National Academy of Sciences. 2006; 103(31):11778–11783. doi:10.1073/pnas.0602659103.

Foerde K, Poldrack R, Knowlton B. Secondary-task effects on classification learning. Memory & Cognition. 2007; 35(5):864–874. doi:10.3758/BF03193461. [PubMed: 17910172]

Gläscher J, Daw N, Dayan P, O'Doherty JP. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron. 2010; 66(4):585–595. doi:10.1016/j.neuron.2010.04.016. [PubMed: 20510862]

Payne, John W.; Bettman, James R.; Johnson, Eric J. The adaptive decision maker. Cambridge University Press; Cambridge ;;New York NY USA: 1993.

Kahneman, D.; Frederick, S. Representativeness revisited: Attribute substitution in intuitive judgment. In: Gilovich, T.; Griffin, D.; Kahneman, D., editors. Heuristics and Biases: The Psychology of Intuitive Judgment. Cambridge University Press; Cambridge: 2002. p. 49-81.

Keramati M, Dezfouli A, Piray P. Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. PLoS Comput Biol. 2011; 7(5):e1002055. doi:10.1371/journal.pcbi.1002055. [PubMed: 21637741]

Loewenstein, G.; O'Donoghue, T.; Cornell University. Center for Analytic Economics. Animal Spirits: Affective and Deliberative Processes in Economic Behavior (Working Papers No. 04-14). 2004. Retrieved from http://ideas.repec.org/p/ecl/corcae/04-14.html

McClure SM, Laibson DI, Loewenstein G, Cohen JD. Separate Neural Systems Value Immediate and Delayed Monetary Rewards. Science. 2004; 306(5695):503–507. doi:10.1126/science.1100907. [PubMed: 15486304]

Metcalfe J, Mischel W. A hot/cool-system analysis of delay of gratification: dynamics of willpower. Psychological Review. 1999; 106(1):3–19. [PubMed: 10197361]

Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. Cognitive Psychology. 2000; 41(1):49–100. doi:10.1006/cogp.1999.0734. [PubMed: 10945922]

Montague PR, Dayan P, Sejnowski TJ. A Framework for Mesencephalic Dopamine Systems Based on Predictive Hebbian Learning. The Journal of Neuroscience. 1996; 16(5):1936–1947. [PubMed: 8774460]

Norman, DA.; Shallice, T. Attention to action: Willed and automatic control of behavior. In: Davidson, RJ.; Schwartx, GE.; Shapiro, D., editors. Consciousness and self-regulation: Advances in research and theory. Vol. 4. Plenum; New York: 1986. p. 1-18.

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal Difference Models and Reward-Related Learning in the Human Brain. Neuron. 2003; 38(2):329–337. doi:10.1016/S0896-6273(03)00169-7. [PubMed: 12718865]

Otto AR, Taylor EG, Markman AB. There are at least two kinds of probability matching: Evidence from a secondary task. Cognition. 2011; 118(2):274–279. doi:doi: DOI: 10.1016/j.cognition.2010.11.009. [PubMed: 21145046]

Pinheiro, JC.; Bates, DM. Mixed-Effects Models in S and S-PLUS. Springer; New York: 2000.

Schultz W, Dayan P, Montague PR. A Neural Substrate of Prediction and Reward. Science. 1997; 275(5306):1593–1599. doi:10.1126/science.275.5306.1593. [PubMed: 9054347]

Sutton, RS.; Barto, AG. Reinforcement Learning. MIT Press; Cambridge, MA: 1998.

Sutton, Richard S. Integrated architecture for learning, planning, and reacting based on approximating dynamic programming; Proceedings of the seventh international conference (1990) on Machine learning; San Francisco, CA, USA. 1990. p. 216-224.

Valentin VV, Dickinson A, O'Doherty JP. Determining the neural substrates of goal-directed learning in the human brain. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience. 2007; 27(15):4019–4026. doi:10.1523/JNEUROSCI.0564-07.2007. [PubMed: 17428979]

Waldron EM, Ashby FG. The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. Psychonomic Bulletin & Review. 2001; 8:168–176. doi:10.3758/BF03196154. [PubMed: 11340863]

Yin HH, Knowlton BJ. The role of the basal ganglia in habit formation. Nat Rev Neurosci. 2006; 7(6):464–476. doi:10.1038/nrn1919. [PubMed: 16715055]

Zeithamova D, Maddox WT. Dual-task interference in perceptual category learning. Memory & Cognition. 2006; 34:387–398. doi:10.3758/BF03193416. [PubMed: 16752602]

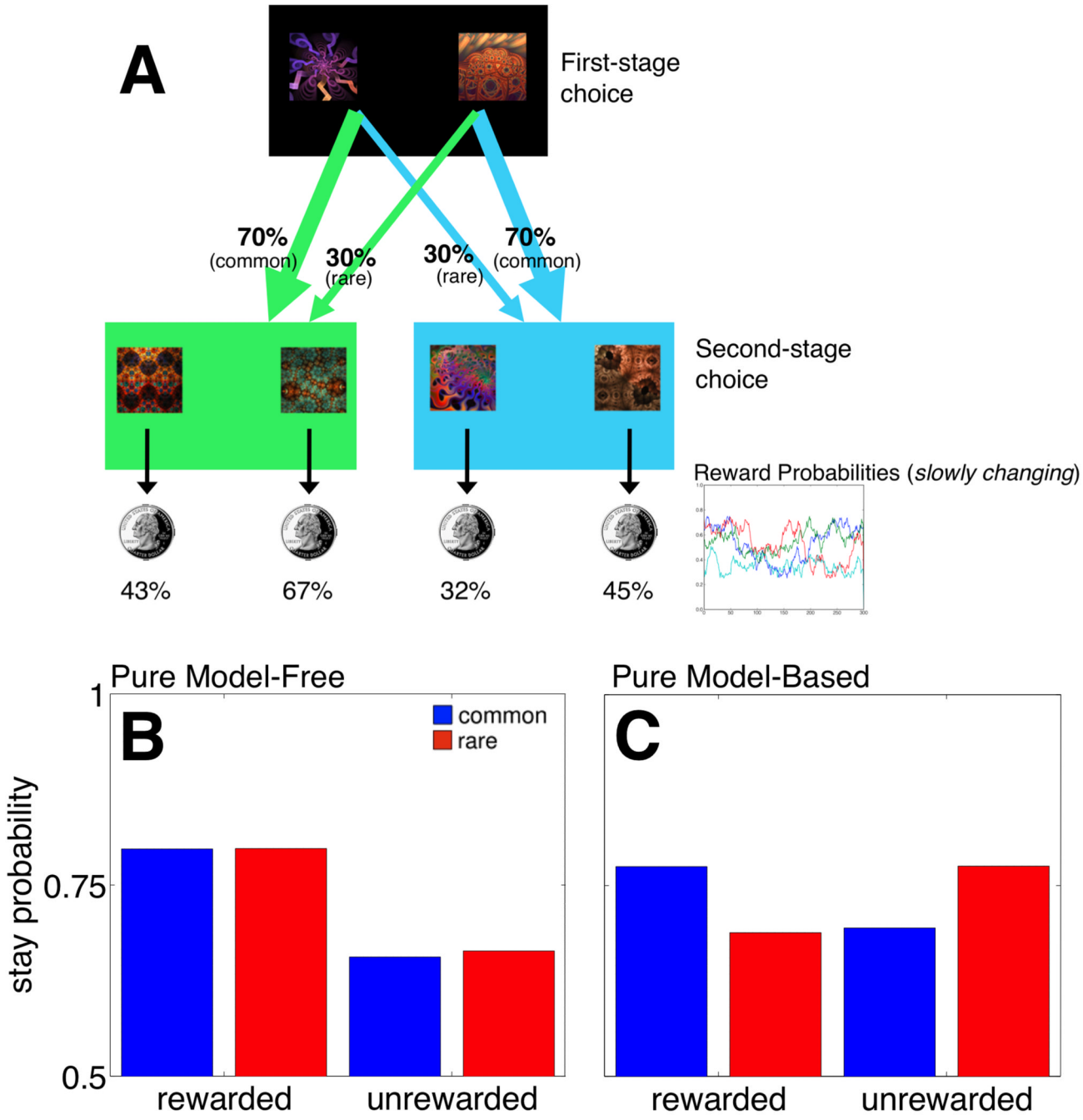

Figure 1.

A) State transition and reward structure in the Two-step task. Each first-stage choice (black background) is predominantly associated with one or the other of the second-stage states (green and blue backgrounds), and leads there 70% of the time. These second-stage choices are probabilistically reinforced with money (see main text for a detailed explanation). (B) Choice behavior predicted by a model-free strategy, which predicts that 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 rare transition. (C) A Model-based based choice strategy predicts that rewards after rare transitions should affect the value of the

unchosen first-stage option, leading to a predicted interaction between the factors of reward and transition probability (reprinted from Daw et al., 2011).

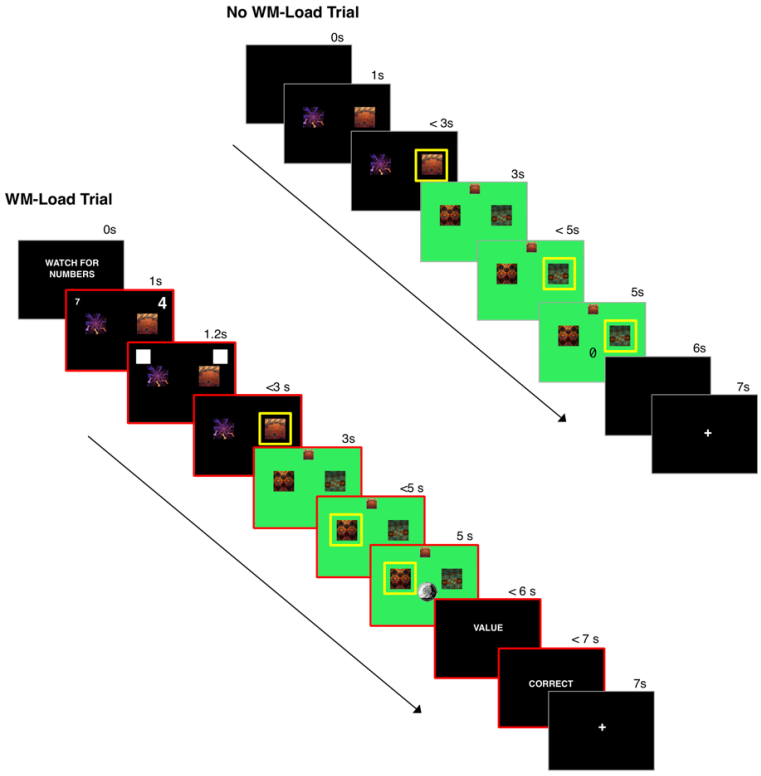

Figure 2.
Timeline of events in No WM-load trials (top; second-stage response was not rewarded) and WM-Load trials (bottom; second stage response was rewarded). Critically, event timing was equated across the two trial types.

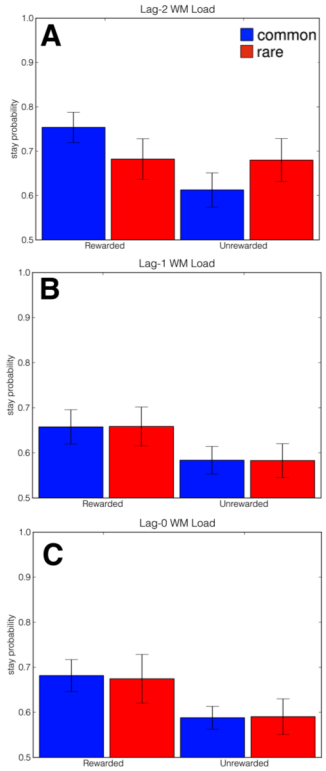

Figure 3.

Average proportion of "stay" trials as a function of reward on previous trial (rewarded versus unrewarded) and transition type on previous trial (common versus rare) across the three trial types of interest. Lag-0, Lag-1, and Lag-2 WM-load refers to trials in which concurrent WM load occurred with the present trial, the previous trial, or the trial preceding the previous trial, respectively. Behavior on Lag-2 trials, in which no WM load was imposed on the current or previous trial, reflects the contribution of a model-based strategy. Conversely, behavior on Lag-0 and Lag-1 trials, in which concurrent WM load was respectively imposed on the current and previous trial, reflects only the contribution of a model-free strategy, suggesting that cognitive demand reverted subjects to model-free choice. Error bars depict standard error of the mean.

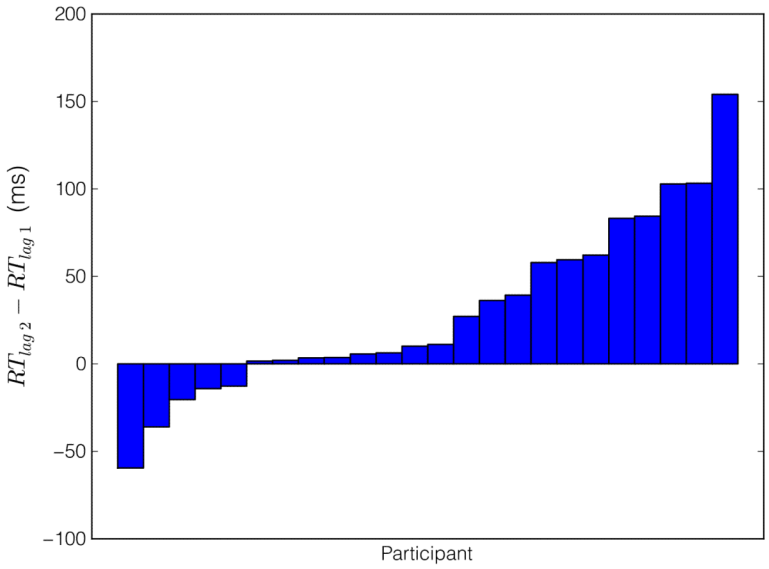

Figure 4.

Individual subjects' choice response times (RT) difference between Lag-2 WM-load trials and Lag-1 WM-load trials. Subjects with positive differences exhibited slower RTs on Lag-2 trials (where behavior appeared more model-based) than on Lag-1 trials (where behavior appeared more model-free). Differences in median adjusted RTs are reported in milliseconds.

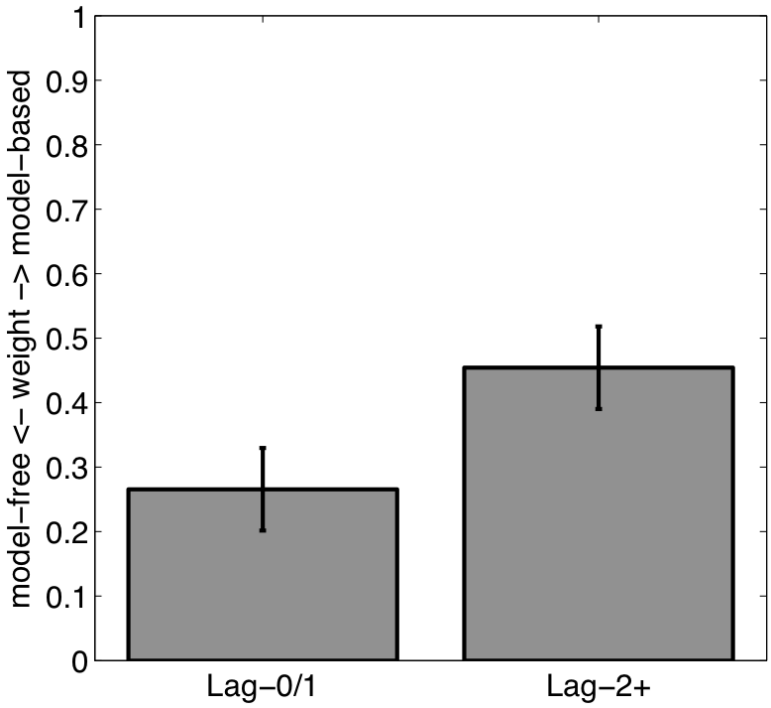

Figure 5.
Best-fitting mixing weights across Lag-2 versus Lag-0/1 WM-load trials resulting from fitting the RL algorithm to subjects' choices (see text). Error bars indicate standard errors.

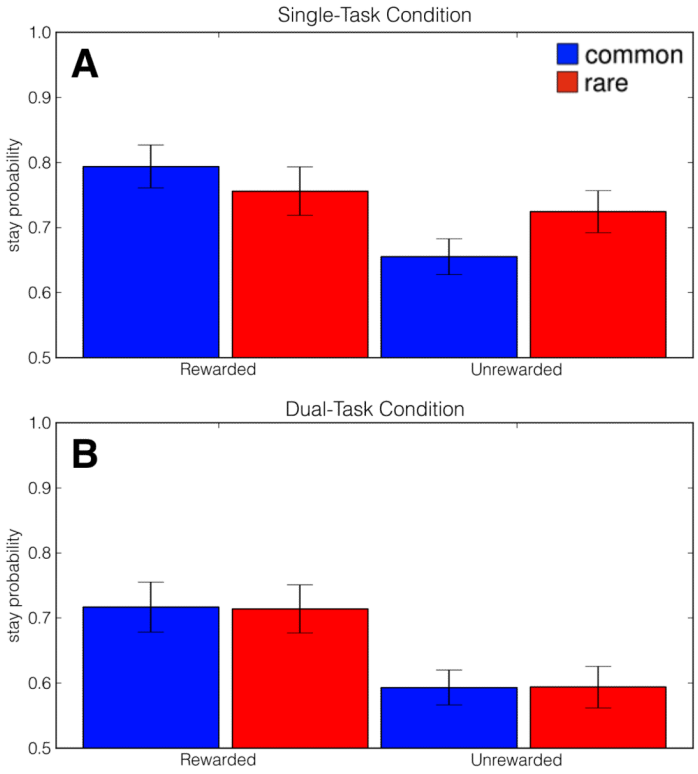

Figure 6.

Average proportion of "stay" trials as a function of reward on previous trial (rewarded versus unrewarded) and transition type on previous trial (common versus rare) across Single-Task (A) and Dual-Task (B) conditions in Experiment 2. Corroborating the results of Experiment 1, concurrent WM load reverted participants to a pure model-free strategy, while participants unfettered by concurrent WM demands exhibited a mixture of model-based and model-free influences.

Table 1

Logistic regression coefficients indicating the influence of WM-load lag, outcome of previous trial, and transition type of previous trial upon first-stage response repetition in Experiment 1. Asterisks denote significance at the .05 level.

Coefficient	*Estimate (SE)*	*p-value*
(Intercept)	1.00 (0.18)	< .0001*
lag-0	−0.23 (0.14)	0.118
lag-1	−0.43 (0.12)	< .0001*
lag-0 × reward	0.34 (0.13)	0.010*
lag-1 × reward	0.19 (0.09)	0.031*
lag-2 × reward	0.23 (0.12)	0.044*
lag-0 × transition	0.07 (0.09)	0.434
lag-1 × transition	−0.07 (0.08)	0.390
lag-2 × transition	0.02 (0.09)	0.776
lag-0 × reward × transition	0.06 (0.09)	0.478
lag-1 × reward × transition	−0.07 (0.08)	0.383
lag-2 × reward × transition	−0.23 (0.09)	0.011*

Table 2

Logistic regression coefficients indicating the influence of WM load condition, outcome of previous trial, and transition type of previous trial upon first-stage response repetition in Experiment 2. Asterisks denote significance at the .05 level.

Coefficient	Estimate (SE)	*p*-value
(Intercept)	1.15 (0.13)	< .0001*
load	−0.25 (0.13)	0.058
reward	0.42 (0.07)	0.000*
transition	0.01 (0.03)	0.823
load × reward	0.01 (0.07)	0.824
load × transition	−0.02 (0.03)	0.433
reward × transition	−0.11 (0.04)	0.005*
load × reward × transition	0.08 (0.04)	0.047*