Running Head: MODEL-BASED CHOICE

Dissecting multiple reinforcement learning systems by taxing the central executive

A. Ross Otto

University of Texas at Austin

Samuel J. Gershman

Princeton University

Nathaniel D. Daw

New York University

Arthur B. Markman

University of Texas at Austin

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 are still unclear. Based on the hypothesis that model-based reinforcement learning requires executive resources, we demonstrate that requiring human decision-makers to perform a demanding secondary task engenders increased reliance on a model-free reinforcement learning strategy. Further, we show that people negotiate this tradeoff dynamically as a function of concurrent executive function demands. 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 (1–6). This distinction is thought to be of considerable practical importance, for instance, as a possible substrate for compulsion in drug abuse (7) and other disorders of self-control (5).

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 (e.g., 8–10). 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 particularly large gap in this regard is between research in human cognitive psychology, which is typically grounded in a distinction between and procedural versus explicit learning and elucidated using manipulations such as working memory (WM) load (11, 12), and another tradition of more invasive animal research on parallel brain structures for instrumental learning (8, 13), 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); 14–17).

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 (2). 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" (18)) 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, or the rules of chess) in order to prospectively evaluate candidate choices. This algorithmic distinction closely echoes theories of instrumental conditioning in animals ((3, 8), but the computational detail of Daw and colleagues' (2) framework leads to relatively specific predictions that afford clear identification of each system's contribution to choice behavior. Recent experiments have demonstrated that human choice behavior in multi-step choice tasks is consistent with a mixture of both types of choice (9, 19). 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. Indeed, functional magnetic resonance imaging (fMRI) unexpectedly revealed overlapping neural signatures of the two strategies (19).

To investigate these questions, we pair the multistep choice paradigm of (19) 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 categorization (12, but see 20), probabilistic classification (11), and simple prediction (21). 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 (22) whereas model-free strategies simply apply the parsimonious principle of repeating previously rewarded actions (2, 23, 24). Thus we hypothesized that concurrent WM load during choice should drive decision-makers towards reliance upon the computationally inexpensive model-free strategy.

To address these questions, we use 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 (12, 25). 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.

Results

Participants performed 300 trials of a two-step RL task (Figure 1; (19)). 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.

Strategy as a function of concurrent WM load

In the two-step task, 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 law of effect—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 rewarding 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 (19). The predicted choice pattern for a pure model-free strategy and a pure model based-strategy are depicted in Figures 2A and B, respectively, derived from simulations of RL algorithms described in (19). 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.

Accordingly, to examine the relationship between these signatures of choice strategies and the concurrent WM load manipulation, 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. If

WM load interferes with model-based learning and/or planning, we would expect behavior on Lag-0 and/or Lag-1 trials to be more model-free (Figure 2A). Conversely, we hypothesized that behavior on Lag-2 trials would resemble a mixture of both strategies (Figure 2B)—mirroring the results of Daw and colleagues' (19) single-task study—as these trials involved no WM load either on the choice trial or on the preceding trial. Figure 3 plots choices as a function of previous reward and transition type, broken down by WM condition. The pattern of results on Lag-2 trials (Figure 3A) 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 study. In contrast, choices on Lag-1 and Lag-2 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 (26) 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 (lag-0 trials: $p<.05$, lag-1 trials: $p<.01$), indicating that participants had a general tendency to repeat rewarded first-stage responses, consistent with both model-based and model-free strategies. However, we found a

significant three-way interaction between Lag-2, reward, and transition type (lag-2 × reward × transition, $p<.05$) suggesting that the characteristic interaction 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 (all p's$>.25$, see Table 1).

To examine whether these differences between trial types were themselves significant, we used a planned contrast, revealing that the size of the Lag-2 three-way interaction (lag-2 × reward × transition, indicative of model-based learning) was significantly larger than both the Lag-1 and the Lag-0 interactions (both $p<.05$) suggesting that WM load on choice trials and trials immediately preceding choice modulated participants' use of model-based choice strategies.

Discussion

We examined how concurrent cognitive demand influences the tradeoff between model-based and model-free choice in a multistep choice task that allowed us to disambiguate model-based and model-free strategies by virtue of their differing predictions about the effect of second-stage rewards on subsequent first-stage choices. Model-based RL, by its nature, incurs greater computational costs than model-free strategies (24), and thus, concurrent cognitive demands should in turn influence the relative contributions of the two strategies (23). Consistent with these expectations, we found that concurrent cognitive demands markedly shifted strategy use. In particular, when burdened with concurrent WM load, decision-makers relied on a pure reinforcement-based strategy—akin to model-free RL in the framework of (2)—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 work (19).

The present result is evocative of past research revealing that concurrent cognitive demand shifts the onus of learning from explicit/declarative systems to procedural learning systems (11, 12). It is important to note while previous work has revealed that concurrent demands can shift response strategies employed by participants, these studies rely on comparing results across multiple task methodologies chosen to favor either strategy (12, 25) or post-hoc assessments of declarative knowledge (11). 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 a difference in strategy use between trial types that occurred fully interleaved, consistent with rapid strategic switching within subject and task. Our results are also complementary to previous fMRI investigations using the present task, since a finding of convergent neural correlates for the two strategies (19) left open the question 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 underscores the utility of within-subjects manipulations for dissociating the behavioral contributions of separate neural systems.

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 (27). 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 putat1ive 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.

The present study represents an initial demonstration of how humans dynamically trade off the usage of model-based versus model-free strategies in decision-making. While (19) 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 (28, 29), a significant portion of the individual variability reported by (19) 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 (5–7).

Materials and Methods

A total of 43 undergraduates at the University of Texas participated in this experiment in exchange for course credit and were paid 2.5 cents per rewarded trial to incentivize choice. We excluded the data of 11 participants whose concurrent task accuracy was less than 75% and the data of one participant who failed to meet a response deadline greater than 20 times. The exclusion of these participants does not change the significance of the reported results. The final group of participants missed an average of 3.16 trials and was rewarded, on average, on 53% of trials, which translated to an average payment of $3.94. The overall proportion of correct Numerical Stroop responses was 86%.

Participants performed 300 trials of a two-step RL task (Figure 1; 19) accompanied by a concurrent Numerical Stroop task on the 100 trials we pseudo-randomly selected as WM-load trials. 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. 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. Choosing the left action moved the participant to the blue state 70% of the time and the green state 30% of the

time, while choosing the right action transitioned to the green state 70% of the time and the blue state 30% of the time (Figure 1).

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 (25). 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.

The mixed-effects logistic regression was performed using the lme4 package (30) in the R programming language (31). All the coefficients in the model, as listed in Table 1, were taken as random effects across subjects, and the estimates and statistics reported are at the population level. The planned comparisons were conducted using the esticon function (package doBy; 33) on the estimated model.

References

1. Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM (1998) A neuropsychological theory of multiple systems in category learning. *Psychological Review* 105:442-481.

2. Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8:1704-1711.

3. Dickinson A (1985) Actions and Habits: The Development of Behavioural Autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* 308:67 - 78.

4. Kahneman D, Frederick S (2002) in *Heuristics and Biases: The Psychology of Intuitive Judgment*, eds Gilovich T, Griffin D, Kahneman D (Cambridge University Press, Cambridge), pp 49-81.

5. Loewenstein G, O'Donoghue T (2004) *Animal Spirits: Affective and Deliberative Processes in Economic Behavior* (Cornell University, Center for Analytic Economics) Available at: http://ideas.repec.org/p/ecl/corcae/04-14.html.

6. Metcalfe J, Mischel W (1999) A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychol Rev* 106:3-19.

7. Everitt BJ, Robbins TW (2005) Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nature Neuroscience* 8:1481-1489.

8. Dickinson A, Balleine B (2002) in *Stevens' Handbook of Experimental Psychology,*.

9. Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron* 66:585-595.

10. Tolman EC (1948) Cognitive maps in rats and men. *Psychological Review* 55:189-208.

11. Foerde K, Knowlton BJ, Poldrack RA (2006) Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences* 103:11778 -11783.

12. Zeithamova D, Maddox WT (2006) Dual-task interference in perceptual category learning. *Memory & Cognition* 34:387-398.

13. Yin HH, Knowlton BJ (2006) The role of the basal ganglia in habit formation. *Nat Rev Neurosci* 7:464-476.

14. Barto AG (1995) Adaptive Critics and the Basal Ganglia. *IN*:215--232.

15. Schultz W, Dayan P, Montague PR (1997) A Neural Substrate of Prediction and Reward. *Science* 275:1593-1599.

16. McClure SM, Berns GS, Montague PR (2003) Temporal Prediction Errors in a Passive Learning Task Activate Human Striatum. *Neuron* 38:339-346.

17. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron* 38:329-337.

18. Thorndike EL (1911) *Animal intelligence: experimental studies* (The Macmillan company).

19. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-Based Influences on Humans' Choices and Striatal Prediction Errors. *Neuron* 69:1204-1215.

20. Newell BR, Dunn JC, Kalish M (2010) The dimensionality of perceptual category learning: A state-trace analysis. *Memory & Cognition* 38:563-581.

21. Otto AR, Taylor EG, Markman AB (2011) There are at least two kinds of probability matching: Evidence from a secondary task. *Cognition* 118:274-279.

22. Norman DA, Shallice T (1986) in *Consciousness and self-regulation: Advances in research and theory*, eds Davidson RJ, Schwartx GE, Shapiro D (Plenum, New York), pp 1-18.

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

24. Keramati M, Dezfouli A, Piray P (2011) Speed/Accuracy Trade-Off between the Habitual and the Goal-Directed Processes. *PLoS Comput Biol* 7:e1002055.

25. Waldron EM, Ashby FG (2001) The effects of concurrent task interference on category learning: Evidence for multiple category learning systems. *Psychonomic Bulletin & Review* 8:168-176.

26. Pinheiro JC, Bates DM (2000) *Mixed-Effects Models in S and S-PLUS* (Springer, New York).

27. Sutton RS (1990) in *Proceedings of the seventh international conference (1990) on Machine learning* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA), pp 216–224. Available at: http://portal.acm.org/citation.cfm?id=101883.102055 [Accessed June 2, 2011].

28. Conway ARA, Kane MJ, Engle RW (2003) Working memory capacity and its relation to general intelligence. *Trends in Cognitive Sciences* 7:547-552.

29. Miyake A et al. (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.

30. Bates D, Maechler M (2009) *lme4: Linear mixed-effects models using S4 classes* Available at: http://CRAN.R-project.org/package=lme4.

31. R Core Development Team (2009) *R: A Language and Environment for Statistical Computing* (Vienna, Austria) Available at: http://www.R-project.org.

32. Højsgaard S, Halekoh U (2009) *doBy: Groupwise computations of summary statistics, general linear contrasts and other utilities* Available at: http://CRAN.R-project.org/package=doBy.

Acknowledgements

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

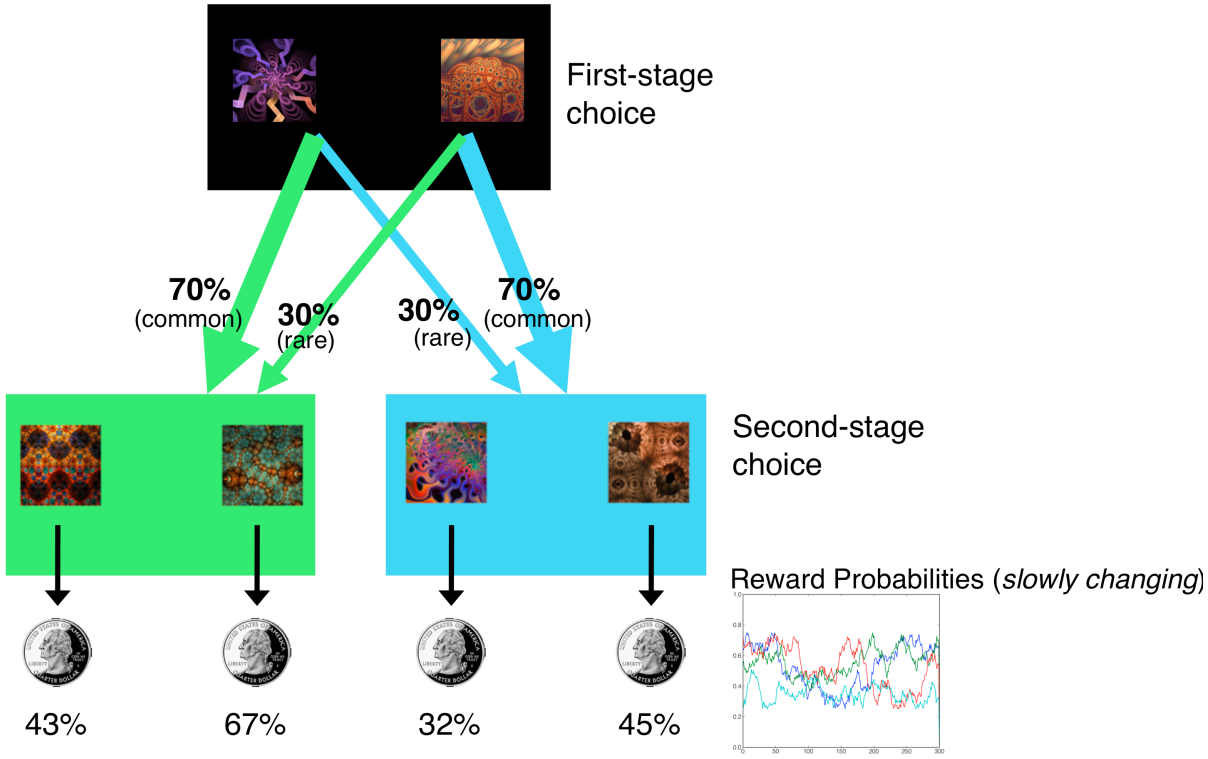

Figure 1. State transition and reward structure in the Two-step task. Each first-stage choice (black background) is predominantly associated with one or the other of the second-stage states (green and blue backgrounds), and leads there 70% of the time. These second-stage choices are probabilistically reinforced with money (see main text for a detailed explanation).

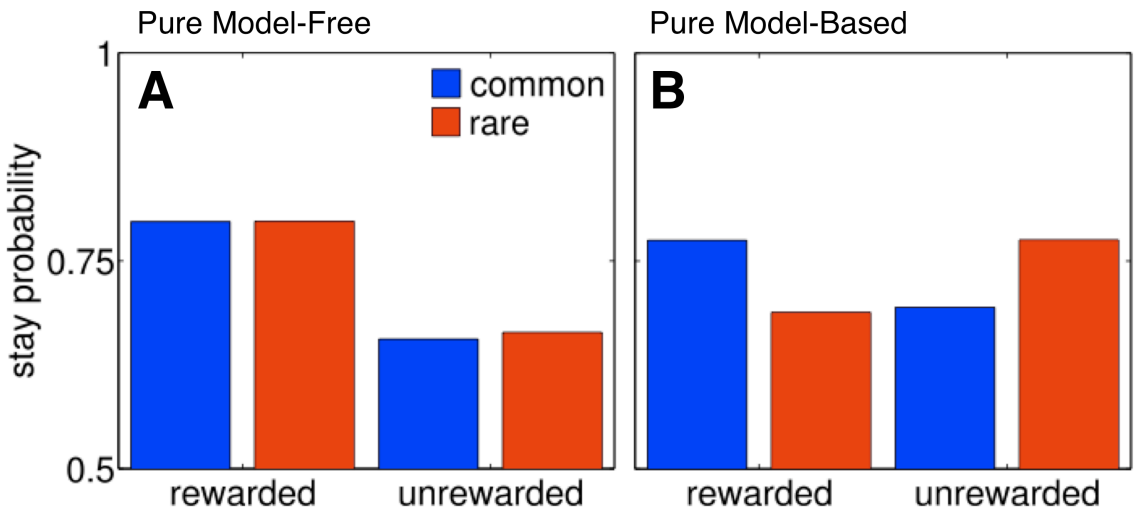

Figure 2. Predicted behavior under the two strategies (reprinted from Daw et al., 2011). Panel A: Choice behavior predicted by a model-free strategy, which predicts that a first-stage choice resulting in reward is more likely to be repeated on the subsequent trial regardless of whether that reward occurred after a common or rare transition. Panel B: 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.

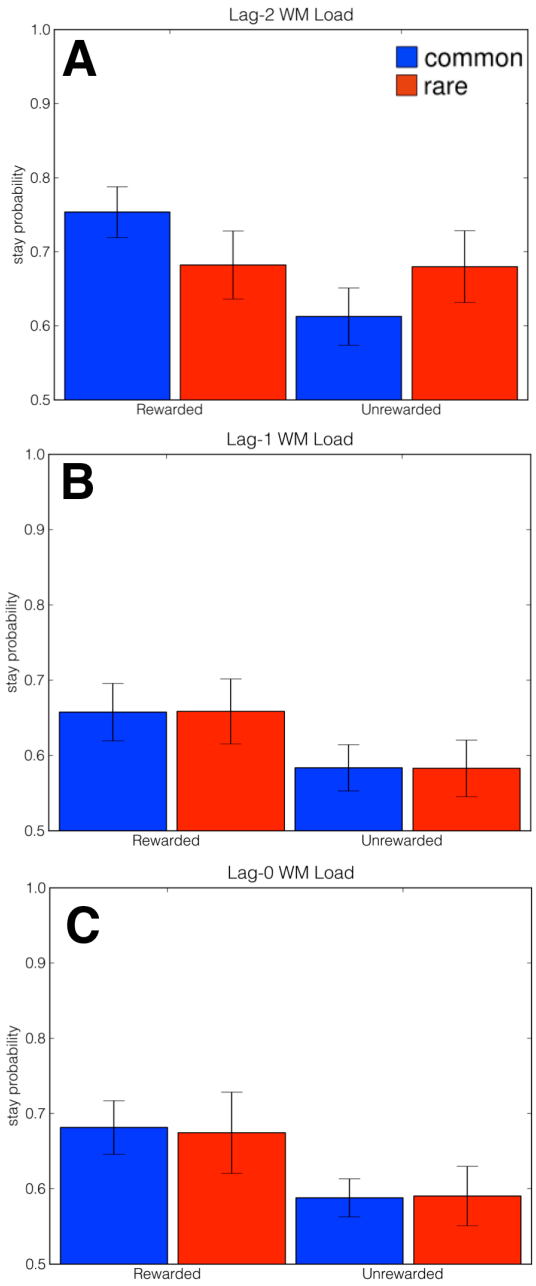

Figure 3. Average proportion of "stay" trials as a function of reward (rewarded versus unrewarded) and transition type (common versus rare) across the three trial types of interest. Lag-0, Lag-1, and Lag-2 WM-load refers to trials in which concurrent WM load occurred with the present trial, the previous trial, or the trial preceding the previous trial, respectively. Error bars depict standard error of the mean.

Table 1. Logistic regression coefficients indicating the influence of WM-load lag, outcome of previous trial, and transition type of previous trial, upon response repetition (see main text). Asterisks denote significance at the .05 level.

Coefficient	*Estimate (SE)*	*p-value*
(Intercept)	1.03 (0.19)	<0.0001
lag-0	-0.25 (0.13)	0.045
lag-1	-0.39 (0.11)	0.001*
lag-0 × reward	0.36 (0.11)	0.002*
lag-1 × reward	0.16 (0.08)	0.041*
lag-2 × reward	0.21 (0.10)	0.028*
lag-0 × transition	0.05 (0.08)	0.502
lag-1 × transition	-0.06 (0.07)	0.388
lag-2 × transition	0.06 (0.08)	0.455
lag-0 × reward × transition	0.08 (0.08)	0.296
lag-1 × reward × transition	-0.07 (0.07)	0.35
lag-2 × reward × transition	-0.21 (0.09)	0.016*