Journal of Experimental Psychology: General
2014, Vol. 143, No. 1, 182–194

0096-3445/14/$12.00 DOI: 10.1037/a0030844

Retrospective Revaluation in Sequential Decision Making: A Tale of Two Systems

Samuel J. Gershman
Princeton University

Arthur B. Markman and A. Ross Otto
University of Texas at Austin

Recent computational theories of decision making in humans and animals have portrayed 2 systems locked in a battle for control of behavior. One system—variously termed *model-free* or *habitual*—favors actions that have previously led to reward, whereas a second—called the *model-based* or *goal-directed* system—favors actions that causally lead to reward according to the agent's internal model of the environment. Some evidence suggests that control can be shifted between these systems using neural or behavioral manipulations, but other evidence suggests that the systems are more intertwined than a competitive account would imply. In 4 behavioral experiments, using a retrospective revaluation design and a cognitive load manipulation, we show that human decisions are more consistent with a cooperative architecture in which the model-free system controls behavior, whereas the model-based system trains the model-free system by replaying and simulating experience.

Keywords: reinforcement learning, DYNA, retrospective revaluation, decision making

Humans and animals routinely face the problem of selecting a sequence of actions in order to maximize future pleasure and minimize future pain. This problem is challenging, because the number of possible sequences grows exponentially with the planning horizon (the number of steps into the future an agent is willing to consider); thus, the naïve strategy of exhaustively evaluating all possible future sequences is not generally tractable for a resource-limited computational device like the brain (or indeed modern computers!). This intractability forces consideration of alternative algorithms the brain might use to solve sequential decision problems.

Two candidate algorithms for action selection have figured prominently in contemporary theories of decision making and have firm grounding in the animal literature. One is a descendant of Thorndike's "law of effect," according to which animals habitually repeat actions that have been reinforced in the past (Thorndike, 1911). The second is a descendant of Tolman's notion of a "cognitive map," an internal model of the environment that animals use to plan goal-directed sequences of actions (Tolman, 1948). Behavioral studies have established that animals use forms of both algorithms under different training regimes (Dickinson, 1985). Moreover, the control of behavior can be shifted from one algorithm to another through focal brain lesions, suggesting the coexistence of neurally distinct decision-making systems—a habit learning system that implements the law of effect and a goal-directed system that implements the cognitive map (Balleine & O'Doherty, 2010; Dickinson & Balleine, 2002).

For example, Adams (1982) showed that rats trained to press a lever for sucrose would subsequently cease lever pressing in an extinction test after the sucrose was separately paired with illness (thereby devaluing the sucrose reinforcer), demonstrating outcome sensitivity consistent with a cognitive map or goal-directed view of instrumental behavior. It is important to note that the law of effect predicts no reduction of responding under these circumstances, because the instrumental action (lever pressing) was never directly paired with illness. However, when the rats were overtrained with the sucrose reinforcer, they continued to press the lever after the devaluation treatment, demonstrating outcome insensitivity more consistent with a habit learning system governed purely by the law of effect (Dickinson, 1985). Similar overtraining effects have recently been demonstrated in humans (Tricomi, Balleine, & O'Doherty, 2009; Valentin, Dickinson, & O'Doherty, 2007).

Our approach follows influential theoretical work (Daw, Niv, & Dayan, 2005; Keramati, Dezfouli, & Piray, 2011; Simon & Daw, 2011) that formalizes the goal-directed and habit systems in the algorithmic framework of reinforcement learning (RL; Sutton & Barto, 1998). In this framework, the decision maker is faced with a set of states (e.g., training cues) and actions (e.g., lever pressing), and the problem is to choose the action in a given state that will maximize cumulative future rewards (also known as "value"). Thorndike's (1911) law of effect is instantiated as a "model-free" RL system that maintains a look-up table containing predictions about future reward for each state–action pair ("cached" values). Learning and prediction is computationally efficient in this system,

This article was published Online First December 10, 2012.

Samuel J. Gershman, Department of Psychology and Princeton Neuroscience Institute, Princeton University; Arthur B. Markman and A. Ross Otto, Department of Psychology, University of Texas at Austin.

A. Ross Otto is now at the Center for Neural Science, New York University.

A. Ross Otto was supported by a Mike Hogg Endowment Fellowship. Samuel J. Gershman was supported by a Graduate Research Fellowship from the National Science Foundation. We are grateful to Grant Loomis for assistance with data collection.

Correspondence concerning this article should be addressed to Samuel J. Gershman, Department of Psychology, Princeton University, Princeton, NJ 08540. E-mail: sjgershm@princeton.edu

because it is only necessary to examine the contents of the look-up table and adjust these predictions incrementally. However, the model-free system is statistically wasteful in its use of experience, because it ignores the transition and reward structure of the environment (i.e., the probability distributions governing transitions between states and the rewards they generate). The practical consequence of this wastefulness is that the model-free system requires a large amount of experience to learn reliable predictions.

In contrast, the "model-based system" learns the transition and reward structure of the environment—akin to Tolman's (1948) cognitive map—and uses this model to generate predictions about future reward. Although this approach is a statistically efficient use of experience, generating predictions is computationally expensive, requiring a form of dynamic programming or tree search. In summary, the two systems have complementary strengths and weaknesses, trading off statistical and computational efficiency (Dayan, 2009). In the Appendix, we provide a more formal exposition of these learning algorithms.

Although the two systems have been viewed as competing for behavioral control (Daw et al., 2005), we argue in this article that their interaction might instead be cooperative in some circumstances. We report the results of four human behavioral experiments, all of which use a retrospective revaluation design in which decision makers are given an opportunity to change their preferences for one set of choices in light of independent experience with a different set of choices occurring later in the sequential decision problem. Our design differs from earlier retrospective revaluation experiments (e.g., Van Hamme & Wasserman, 1994) because it exploits the sequential structure of the task. As we explain in more detail below, the probability that a particular state–action pair will yield a reward never changes. Instead, we change the information available to participants about reward contingencies at states later in the sequence. This has the consequence of devaluing actions early in the sequence that were previously rewarding, while making previously unrewarding actions more valuable. Model-free and model-based RL algorithms make different predictions about how a decision maker will respond to this change.

By asking some participants to perform a demanding secondary task, we are able to diminish the use of model-based computations, which in turn reduces retrospective revaluation (as we discuss further below). To foreshadow our findings, the pattern of behavior exhibited by human decision makers' rules is inconsistent with some competitive accounts (e.g., Daw et al., 2005; Keramati et al., 2011; Simon & Daw, 2011). Although it is impossible to rule out all varieties of competitive accounts, we suggest that our experimental results have a natural explanation under a class of cooperative architectures first proposed in the artificial intelligence literature (Sutton, 1990). As an illustrative proof of concept, we explore the predictions of one popular version of this architecture.

In Sutton's DYNA architecture (shown schematically in Figure 1), behavior is controlled exclusively by the model-free system (allowing choices to be made online with minimal computational expense), but the model-based system exerts an indirect influence on behavior by training the model-free system offline (i.e., in between episodes of real experience). Specifically, the model-based system replays experienced state–action pairs and then simulates a transition and reward on the basis of its learned model of the environment. The model-free system can then learn—using an algorithm called Q-learning (Watkins, 1989)—from these simulated trajectories as though they were real experiences (see the Appendix for a more detailed description). The experiments reported in this article provide evidence for such a cooperative architecture in human choice. Using computer simulations, we demonstrate that our experimental findings can be reproduced qualitatively by a simple implementation of DYNA.

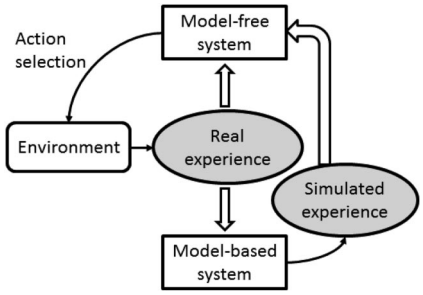

Figure 1. The DYNA architecture (from Sutton, 1990). The environment furnishes the agent with real experience (transitions and rewards), whereas the model-based system furnishes the agent with simulated experience by replaying state–action pairs from memory and then using a learned model of the environment to generate transitions and rewards. The model-free system learns in the same way from both real and simulated experience, and uses its learned values to control action selection.

Overview of the Experiments

A key distinction between model-based and model-free RL is that model-free algorithms like Q-learning can only update values along experienced state–action trajectories, whereas model-based RL propagates information across all states and actions by updating the state transition probabilities and reward functions and performing dynamic programming. This means that model-free RL will be unable to retrospectively revalue certain state–action pairs in light of new experience with other states and actions, unless they were directly experienced in the same trajectory. It is also possible to create scenarios in which model-based RL predicts no retrospective revaluation, as we describe below. Our experiments instantiate a simple multistep choice task in which neither model-based RL nor model-free RL alone predict retrospective revaluation, but a cooperative algorithm (such as DYNA) does predict retrospective revaluation.

The experimental design used in all our experiments is shown schematically in Figure 2, with numbered circles denoting states (three distinct background colors in our experiment) and letters denoting actions (button presses made to distinct visual stimuli). Because each trajectory consists of two steps, we sometimes refer to actions in State 1 as *first-step actions* and actions in States 2 or 3 as *second-step actions*. The transition structure is set up so that Action *A* in State 1 leads deterministically to State 2, whereas Action *B* leads to State 3, although (as we explain below) participants do not always experience a full two-step trajectory.

The experiment is divided into four phases. In the first phase, participants are trained on the transition structure without any rewards. In the second phase, participants make a series of decisions in the first step only and receive rewards as shown in Figure 2, establishing a baseline preference for the more valuable Action, *A*. In the third phase, participants make a series of rewarded

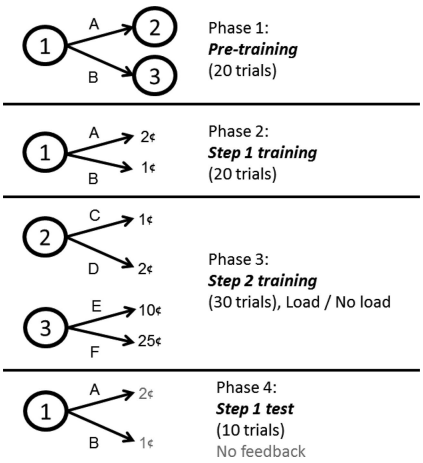

Figure 2. Experimental design. The sequential decision problem consists of three states (indicated by numbered circles) and two mutually exclusive actions in each state (indicated by letters). Deterministic transitions between states conditional upon the chosen action are indicated by arrows. Rewards for each state–action pair are indicated by amounts (in cents). In Phase 4, reward feedback is delayed until the end of the phase. In the task interface, states are signaled by background colors, and actions are signaled by fractal images.

decisions in the second step only. The reward structure is designed so that State 3 is associated with much larger rewards than States 1 and 2. As a result, the value (cumulative reward) of Action *B* in State 1 over a two-step horizon is much higher than the value of Action *A*, inducing a conflict between the value of *B* and its immediate reward in State 1. In other words, an optimal agent, planning over a two-step horizon, would change its preference from *A* to *B* in State 1 after learning the second-step reward contingencies in Phase 3 (i.e., the agent would retrospectively revalue State 1 actions). In the fourth phase, participants played the first step again without immediate feedback (i.e., their rewards were only made known after that phase), providing them with an opportunity to display a change to their preferences.

The primary dependent measure in this experiment is the "revaluation magnitude," which is the difference in preference for first-step Action *B* between the fourth phase and the second phase, or more formally, $P_4(\text{action} = B|\text{state} = 1) - P_2(\text{action} = B|\text{state} = 1)$, where $P_i(\text{action} = a|\text{state} = s)$ is the probability of choosing Action *A* in State *s* during Phase *i*. Critically, pure model-free (Q-learning) accounts and pure model-based (dynamic programming) accounts both predict a revaluation magnitude of zero, but for different reasons. Q-learning has no mechanism for propagating information from second-step rewards to first-step values without experiencing a full two-step trajectory, and hence first-step preferences should remain unchanged after second-step training in Phase 3. Dynamic programming is one mechanism by which second-step rewards can influence first-step values. However, because the decision problem in Phase 4 is restricted to the first step, the correct model-based values for Step 1 in fact do not use second-step rewards at all: The values of first-step actions over a one-step horizon are equal to their immediate rewards. Consequently, reward information from the second step in Phase 3 should have no effect on choice probabilities in Phase 4 according to a model-based account.

By contrast, a cooperative architecture such as DYNA predicts a positive revaluation magnitude. During Phase 3, the model-based system trains the model-free values; this has the effect of increasing the value of taking Action *B* in State 1, because now it is highly desirable to transition to State 3 (where large rewards await), even though the immediate reward for taking Action *B* in State 1 is less than for taking Action *A*. Because the model-free system cannot alter its cached values on the basis of the planning horizon, the updated first-step values will be used during Phase 4.

It is important to point out here that we are contrasting DYNA with only the most obvious versions of model-based and model-free algorithms that have been proposed in past research (Daw et al., 2005; Keramati et al., 2011). There are probably versions of these models that can reproduce our experimental findings. Our goal in these experiments was not to provide decisive evidence for a DYNA-like architecture per se, but rather to produce a set of constraints governing RL accounts of human sequential decision making.

The experiments reported in this article use a variety of manipulations to increase or decrease the influence of model-based computations in order to reveal the mechanisms underlying retrospective revaluation. In Experiment 1, we place a group of participants under concurrent cognitive load by imposing a demanding secondary task. As noted above, retrospective revaluation in our task crucially depends on a model-based mechanism, because model-free learning relies on unbroken trajectories through the state space that are absent in our design. Previous work (Otto, Gershman, Markman, & Daw, 2013) found that concurrent cognitive load causes participants to behave according to the law of effect, repeating previously reinforced actions even when they are unlikely to lead to reward. This suggests that load attenuates the contributions of model-based computations, as would be predicted by the cognitively demanding nature of these computations (Daw et al., 2005; Dayan, 2009; Keramati et al., 2011). Accordingly, we expected that concurrent cognitive load during Phase 3 would reduce the retrospective revaluation effect.

Furthermore, the DYNA account predicts that providing additional idle time should counteract the deleterious effects of cognitive load by providing additional opportunities for offline training of the model-free values by the model-based system. Accordingly, we explore the effects of increasing the number of Phase 3 trials (Experiment 2) or interposing a long rest interval between Phase 3 and Phase 4 (Experiment 3) on retrospective revaluation magnitude. Experiments 2 and 3 highlight a distinctive property of DYNA: transfer of model-based knowledge to the model-free system is inherently time-consuming. In Experiment 4, we manipulate cognitive load during Phase 4 to evaluate the possibility that the revaluation effects observed in the preceding experiments stem from pure model-based planning occurring during Phase 4 rather than in an offline manner prescribed by the DYNA account.

Experiment 1

The aim of our first experiment was to establish the conditions under which retrospective revaluation occurs in the sequential decision paradigm described above. Our hypothesis was that a positive revaluation magnitude would depend on resource-demanding model-based computations during Phase 3, and would therefore be attenuated when central executive resources were

depleted. To test this hypothesis, we examined the magnitude of retrospective revaluation between participants required to perform a demanding concurrent task during Phase 3 (henceforth, the load condition) and a control group of participants with no concurrent cognitive demands (henceforth, the no-load condition).

In addition to the effects of the load manipulation, we expected that individual differences in executive function would predict individual differences in retrospective revaluation magnitudes. More specifically, we reasoned that larger working memory (WM) capacities—often discussed as the hallmark of executive function (Conway, Kane, & Engle, 2003)—would lead to larger revaluation effects. Accordingly, we investigated this relationship using a common measure of WM and executive function capacity, the operation span test (OSPAN; Engle, 2002; Turner & Engle, 1989).

Method

Participants. A total of 119 University of Texas undergraduates participated in the experiment for course credit and a small cash bonus tied to earnings in the choice task. All participants gave informed consent, and the study was approved by the University of Texas Office of Human Subjects Research. Participants were randomly assigned to either the load condition or the no-load condition. To ensure that dual-task participants did not trade off performance on the concurrent task in order to perform the primary task (Zeithamova & Maddox, 2006), we excluded the data of four load participants who exhibited a root-mean-square-error (RMSE) of 4 or more on the secondary task (described below).

Materials and procedure. To assess individual participants' WM capacity, we administered an automated version of the OSPAN procedure (Unsworth, Heitz, Schrock, & Engle, 2005) prior to the choice task, which required participants to remember a series of letters while performing a concurrent task. The choice task was administered on a computer and was programmed using the PyGame library for the Python programming language (Shinners, 2011). State 1 was represented by a black background, and Stimuli A and B were represented by fractal images. States 2 and 3 were represented with green and blue backgrounds, respectively, and Stimuli C, D, E, and F were represented by unique fractal images. Assignment of fractal images to stimuli as well as the positions (left vs. right) of these stimuli were randomized across participants.

Choices in all stages of the experiment followed the same general procedure: Two fractal images appeared on a background indicating the initial state, and there was a 2-s 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 presentation of the monetary outcomes. To avoid potential verbal interference from reading numerical rewards values, images of common denominations of U.S. coins were presented. Before performing the task, participants were instructed about the general structure of the choice task—namely, that an initial choice on a black screen was made between two images that would result in a monetary reward and would transition them to either a green or a blue screen, depending on their choice, where a subsequent choice would be made between two images again for a monetary reward. Participants were instructed that because of the multistep structure of the task, their choices made on the black screen should take into account not only the immediate rewards but also the rewards available on the screen (blue or green) that their initial choice takes them to.

The initial instructions were followed by Phase 1 of the choice task, in which participants learned about the deterministic transition structure of the task (see Figure 2A) by making 20 choices starting in State 1 and observing which state (State 2 or State 3) their choice transitioned to. Critically, selections made to Stimulus A always transitioned to State 2, whereas selections made to Stimulus B always transitioned to State 3. In lieu of reward feedback, the background and stimuli of the resultant state was displayed for 2 s. Next, in Phase 2, participants made 20 rewarded choices in State 1. Choices made to Stimuli A and B were rewarded with 2¢ and 1¢, respectively, depicted as American pennies on the screen during the feedback period.

Participants subsequently began Phase 3 in which they made 30 rewarded choices starting from either State 2 or State 3 (see Figure 2A). At the beginning of each trial, one of the two states was randomly selected with equal probability, and its associated color and stimuli were shown. In State 2, selections made to Stimuli C and D resulted in rewards of 1¢ and 2¢, respectively, whereas in State 3, selections made to Stimuli E and F resulted in rewards of 10¢ and 25¢, respectively. These rewards were represented by images of American pennies, dimes, or quarters.

Finally, in Phase 4, participants made 10 unrewarded choices in State 1 to either Stimulus A or B. Before this phase, participants were informed they were making choices between the same stimuli as in Phases 1 and 2, but they were to make these choices on the basis of what they had learned over the course of the experiment. To avoid biasing choice in critical test trials, no rewards were presented during the feedback period. The precise instructions for Phase 4 were as follows:

> Lastly, we want you to make some first-stage choices based on what you have learned over the course of the experiment. The boxes you will choose between are the same ones you saw before when you learned which second-stage choices the first-stage choices take you. You will receive money today for making these choices, but you will not be able to immediately see how much each first-stage choice has earned you. We want you to use your previous knowledge to make these first stage choices.

The procedure was identical for participants in the no-load and load conditions except for a concurrent WM task imposed on load participants during Phase 3, following the general tone-counting procedure of Foerde, Knowlton, and Poldrack (2006) modified to ensure that WM demands persisted over all stages of the choice task (Otto et al., 2013; Otto, Taylor, & Markman, 2011). Two types of tones, high pitched (1000 Hz) and low pitched (500 Hz), were played during each trial, and participants were instructed to maintain a running count of the number of high tones while ignoring the low-pitched tones. Each trial was divided into 16 intervals of 250 ms, with tones occurring in intervals of 2–15 (500 ms–3,750 ms after trial onset). The number of tones presented in each trial varied uniformly between 1 and 6. The base rate of high tones was determined every 10 trials, varying uniformly between 0.3 and 0.7. At the end of each 10-trial block, load participants reported their count and were instructed to restart their count at zero, whereas no-load participants were instructed to take a break.

The timing of these trials was equated across no-load and load conditions according to the procedure described above.

Results and Discussion

Across both conditions, participants in Phase 2 overwhelmingly preferred Action *A* to Action *B* (see Table 1), confirming that they learned to choose the higher value action. In Phase 4, participants in both groups revealed a net retrospective revaluation effect, showing increased preference for Action *B* relative to Phase 2 (see Figure 3).

To test the hypothesis that decision makers in the load condition would exhibit less retrospective revaluation than participants in the no-load condition, we performed a two-sample *t* test on the revaluation magnitude across the two groups. Consistent with our hypothesis, we found that revaluation magnitude was significantly smaller in the load condition than in the no-load condition, $t(117) = 2.00$, $p < .05$.

An alternative explanation for the revaluation results is that participants in the no-load condition were using a model-based strategy (i.e., some form of dynamic programming; see the Appendix) during Phase 4, rather than choosing on the basis of model-free values. Note that for this explanation to be tenable, we would have to assume that participants are incorrectly computing values under a two-step (rather than one-step) horizon. The difference between load and no-load conditions could then be attributed to impaired model-based learning of rewards in States 2 and 3 during Phase 3. Consistent with this proposal, participants in the load condition showed a weaker preference for the higher value action in Phase 3 (collapsing across both second-step states) than participants in the no-load condition, $t(78) = 3.57$, $p < .001$. We shall pursue this alternative explanation in Experiments 4A and 4B.

To further pursue modulators of this effect, we looked for a relationship between OSPAN performance and revaluation magnitude. OSPAN scores were calculated by summing the number of letters selected for all correctly selected sets (Unsworth et al., 2005). Scores ranged from 2 to 72 ($M = 45.18$, $SE = 1.51$). In the no-load condition, we found a significant correlation between OSPAN and revaluation magnitude, $r(65) = .33$, $p < .01$. In contrast, there was no significant correlation between OSPAN and revaluation magnitude for participants in the load condition, $r(50) = -.20$, $p = .16$. We interpret this finding to mean that under no-load conditions, individual differences in WM capacity influenced the extent to which participants update their State 1 values using model-based information during Phase 3, whereas under load conditions these individual differences were possibly overwhelmed by the secondary task demands.

Table 1
Choice Probabilities in Each Condition and Phase for Experiment 1, Reported as Mean ± Standard Error

Phase	Load	No load
Phase 2	0.11 ± 0.01	0.12 ± 0.01
Phase 3	0.70 ± 0.03	0.84 ± 0.02
Phase 4	0.41 ± 0.05	0.56 ± 0.05

Note. Mean ± standard error is the probability of choosing Action B in State 1 for Phases 2 and 4. For Phase 3, mean ± standard error is the average probability of choosing Action D in State 2 or Action F in State 3 (i.e., the higher value action).

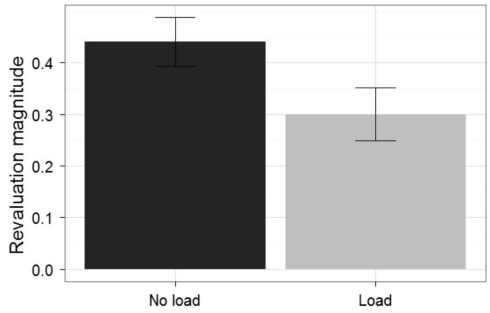

Figure 3. Experiment 1 results. Revaluation magnitude is measured as $P_4(\text{action} = B|\text{state} = 1) - P_2(\text{action} = B|\text{state} = 1)$, where $P_i(\text{action} = a|\text{state} = s)$ is the probability of choosing action *a* in State *s* during Phase *i*. Error bars indicate standard error of the mean.

To make sure that the correlation between OSPAN and revaluation magnitude is not due to impaired Phase 3 learning, we examined the relationship between OSPAN and preferences for the higher value option in the second-step states during Phase 3. Correlations were not significant, both for load ($p = .52$) and no-load ($p = .29$) conditions.

Under a model-based account, we would expect participants who show greater revaluation to also show longer response times, due to the greater processing demands involved in producing model-based revaluation (Keramati et al., 2011). Using multiple regression, the relationship between response times and revaluation magnitude was not significant when controlling for load ($p = .54$). Furthermore, second-step preferences during Phase 3 were not correlated with subsequent revaluation magnitude ($p = .17$), as one would expect if impaired model-based learning was the cause of a diminished revaluation effect under load.

Another possibility is that in Phase 4, participants initially (despite instructions to the contrary) believed they were accumulating reward over the full two-step horizon, and therefore performed model-based planning, but over the course of the 10 trials realized that they were only accumulating rewards over a one-step horizon. To evaluate this possibility, we recalculated the revaluation magnitude separately for the first half and second half of the Phase 4 trials. Confirming our previous analyses, a two-way analysis of variance (ANOVA) revealed a main effect of load, $F(238) = 7.35$, $p < .007$, but no main effect of half ($p = .73$) and no interaction between load and half ($p = .9$). Thus, it does not appear to be the case that participants are changing their behavior over the course of Phase 4. Taken together, these analyses argue against a purely model-based interpretation of our findings.

Experiment 2

In Experiment 1, we found that depleting decision makers' WM resources via concurrent cognitive demand during Phase 3 reduced their magnitude of retrospective revaluation. We also found that individual differences in retrospective revaluation were, among participants in the no-load condition, systematically related to a measure of executive function. In Experiment 2, we sought to establish some boundary conditions on this experimental effect by

manipulating the number of Phase 3 trials. In Experiment 2, we used a 2 × 2 factorial design, crossing load versus no load and 30 versus 50 Phase 3 trials.

According to DYNA, increasing the number of trials should provide further opportunities for offline training of the model-free values by the model-based system. If WM load impairs, but does not eliminate entirely, the process of updating first-step values on the basis of information about second-step rewards (e.g., through DYNA-style replay), then increasing the number of Phase 3 trials should ameliorate this impairment.

Method

Participants. A total of 172 University of Texas undergraduates participated in the experiment for course credit. All participants gave informed consent, and the study was approved by the University of Texas Office of Human Subjects Research. Participants were divided into conditions that factorially manipulated WM load (load vs. no load) and number of Phase 3 training trials (30 vs. 50), yielding four groups: load-30, load-50, no-load-30, and no-load-50. To ensure that load participants did not trade off performance on the concurrent task in order to perform the primary task, we excluded the data of four load participants who exhibited an RMSE of 4 or more on the tone-counting task.

Materials and procedure. The procedure in Experiment 2 was identical to the procedure in Experiment 1 except for one aspect: Participants in the load-50 and no-load-50 groups were given 50 trials of Phase 3 choices—in which rewarded choices were made from States 2 and 3—whereas participants in the load-30 and no-load-30 groups were given 30 trials of Phase 3, mirroring the two conditions in Experiment 1.

Results and Discussion

First-step choice probabilities are summarized in Table 2. The retrospective revaluation data are shown in Figure 4. To assess the joint effects of load and number of Phase 3 trials, we performed a two-way ANOVA (load vs. no load × 30 trials vs. 50 trials). There was a significant interaction between load and number of trials, $F(1, 168) = 6.06$, $p < .05$, indicating that difference in retrospective revaluation magnitude between load and no-load conditions was significant for 30 trials but not for 50 trials. As predicted, increasing the number of Phase 3 trials allowed participants under cognitive load to exhibit an equal degree of revaluation compared with participants not under load. Although revaluation magnitude appears to be paradoxically higher for no-load-30 compared with no-load-50, this difference is nonsignificant ($p = .34$).

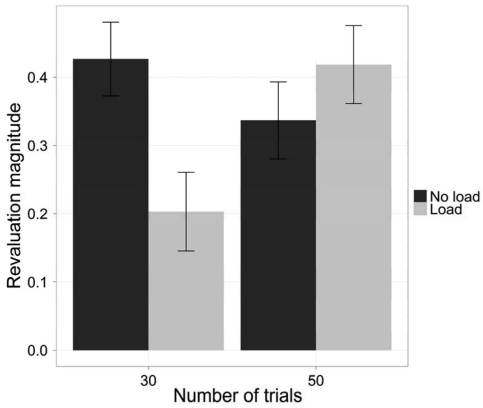

Figure 4. Experiment 2 results. Revaluation magnitude as a function of secondary task during Phase 3 (load/no load) and the number of Phase 3 trials (30/50). Error bars indicate standard error of the mean.

Consistent with the results of Experiment 1, participants in the load-30 condition showed a weaker preference for the higher value action in Phase 3 (collapsing across both second-step states) than participants in the no-load-30 condition, $t(28) = 2.43$, $p < .05$. This pattern also holds with 50 Phase 3 trials, $t(42) = 3.75$, $p < .001$. A two-way ANOVA (Load/No Load × 30/50 trials) revealed a main effect of load, $F(149) = 20.99$, $p < .0001$, and number of trials, $F(149) = 8.84$, $p < .005$. Note that revaluation is unaffected by load in the 50-trial condition, despite impaired Phase 3 learning in the same condition. This observation is consequential for our theoretical interpretation of the revaluation results: Our revaluation effects cannot be explained as simple functions of second-step reward learning in Phase 3.

The data from this study suggest that model-based information can still be propagated to first-step values under load given enough time. It is not clear, however, whether this effect is due to increased *time* or increased *number of trials*. If propagation occurs via a time-consuming memory replay and simulation process (as postulated by DYNA), then the critical variable is the amount of time during which the participants are not being asked to do anything—that is, periods of inactivity during which replay might occur. We investigated this idea further in Experiment 3.

Experiment 3

The purpose of Experiment 3 was to examine a distinctive prediction of DYNA: Simply increasing the amount of time during

Table 2
Choice Probabilities in Each Condition and Phase for Experiment 2, Reported as Mean ± Standard Error

Phase	Load/30 trials	No load/30 trials	Load/50 trials	No load/50 trials
Phase 2	0.13 ± 0.02	0.12 ± 0.01	0.10 ± 0.01	0.11 ± 0.01
Phase 3	0.77 ± 0.04	0.87 ± 0.01	0.84 ± 0.02	0.91 ± 0.01
Phase 4	0.33 ± 0.07	0.53 ± 0.05	0.54 ± 0.07	0.45 ± 0.06

Note. Mean ± standard error is the probability of choosing Action B in State 1 for Phases 2 and 4. For Phase 3, mean ± standard error is the average probability of choosing Action D in State 2 or Action F in State 3 (i.e., the higher value action).

which participants are not performing the task between Phase 3 and Phase 4 (i.e., without adding new learning trials) should increase retrospective revaluation. This prediction arises from DYNA's use of quiescent periods to replay past experiences and simulate from the learned model of the environment, providing additional training for the model-free system. Indeed, electrophysiological recordings of rodent hippocampal place cells offer one suggestive source of evidence for such offline replay and simulation (Foster & Wilson, 2006; Johnson & Redish, 2007); we return to this idea in the General Discussion.

In one group of participants, we interposed a rest period between Phase 3 and Phase 4, during which participants listened quietly to a recording of classical music. The other group proceeded immediately to Phase 4 without a rest (as in Experiments 1 and 2). Because we were mainly interested in whether a rest period would counteract the deleterious effects of load during Phase 3, both groups performed Phase 3 under load. We predicted that revaluation magnitude would increase for participants with the rest interval, because they would have more time for offline training, and thereby update their first-step values on the basis of information about second-step rewards acquired during Phase 3.

Method

Participants. A total of 106 University of Texas undergraduates participated in the experiment for course credit. All participants gave informed consent, and the study was approved by the University of Texas Office of Human Subjects Research. Participants were randomly assigned to two groups defined by the interval between Phase 3 and Phase 4. All participants experienced concurrent cognitive load.

Materials and procedure. Participants in the no-rest condition proceeded immediately to Phase 4 after completing training on second-stage rewards. Procedurally, this condition is identical to the load condition in Experiment 1. Participants in the rest condition were shown a screen instructing them to sit quietly and look at the fixation cross while listening to a piece of music (Sergei Prokofiev's Op. 12 No. 2 for piano; Prokofiev, 2008) for 3 min, after which participants proceeded to Phase 4.

Results and Discussion

Choice probabilities are summarized in Table 3. As shown in Figure 5, participants in the rest condition showed a significantly higher revaluation magnitude than participants in the no-rest condition, $t(117) = 2.13$, $p < .05$. Thus, adding a quiescent interval between Phase 3 and Phase 4 increased the extent to which participants retrospectively revalued first-step choices in light of second-step rewards. This finding is consistent with the prediction of DYNA that providing extra inactive time, without new learning trials, allows the model-based system to update model-free values.

Table 3
Choice Probabilities in Each Condition and Phase for Experiment 3, Reported as Mean ± Standard Error

Phase	Rest	No rest
Phase 2	0.11 ± 0.01	0.12 ± 0.01
Phase 3	0.88 ± 0.02	0.85 ± 0.02
Phase 4	0.54 ± 0.05	0.40 ± 0.05

Note. Mean ± standard error is the probability of choosing Action B in State 1 for Phases 2 and 4. For Phase 3, mean ± standard error is the average probability of choosing Action D in State 2 or Action F in State 3 (i.e., the higher value action). Note that both the rest and no-rest conditions were performed under load.

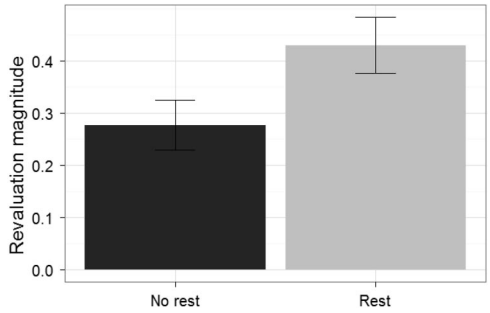

Figure 5. Experiment 3 results. Revaluation magnitude in Phase 4 as a function of rest condition. Both conditions were performed under working memory load. Error bars indicate standard error of the mean.

To examine whether there was any change in behavior over the course of Phase 4, we recalculated the revaluation magnitude for the first five trials and second five trials, and then performed a two-way ANOVA (Rest/No Rest × First Half/Second Half). Consistent with the results of Experiment 1, there was a main effect of rest, $F(212) = 8.06$, $p < .005$, but not of half ($p = .51$). The interaction between the factors was also not significant ($p = .78$).

Experiments 4A and 4B

Recent experiments have shown that cognitive load tends to reduce the influence of the model-based system in favor of model-free control of choice behavior (Otto et al., 2013). We leveraged this finding to ask: Are participants performing a pure version of model-based planning during Phase 4? As we pointed out in the introduction, planning over a two-step horizon is irrational given that participants are only being rewarded for choices in State 1. A pure model-based account in fact predicts no revaluation, given that rewards in State 1 do not change over the course of the experiment. Nonetheless, we can put this theoretical objection aside and examine the question empirically: If we assume participants are planning over a two-step horizon and this planning occurs during Phase 4 choice, then applying cognitive load during Phase 4 should impair planning and thereby reduce revaluation. In contrast, if participants are using cached values derived from a cooperative architecture-like DYNA, then cognitive load should not affect revaluation.[1]

In Experiments 4A and 4B, we tested these predictions. The two experiments are almost identical except for the instructions given to participants at the beginning of Phase 4. In Experiment 4A, we used the same instructions as in Experiments 1–3. One concern with these instructions is that they may not make sufficiently clear to participants that their rewards in Phase 4 do not depend on the second-step reward contingencies. For example, participants may

[1] We are grateful to an anonymous reviewer for suggesting this experiment.

incorrectly believe that they should take actions that would lead to the most profitable outcome on a complete two-step trajectory. To rule out the possibility that participants are simply misinterpreting the task structure, in Experiment 4B we modified the Phase 4 instructions, stating clearly that rewards do not depend on second-step choices.

Method

Participants. A total of 132 University of Texas undergraduates (40 in Experiment 4A and 92 in Experiment 4B) participated in the experiments for course credit. All participants gave informed consent, and the study was approved by the University of Texas Office of Human Subjects Research. Participants were randomly assigned to two groups (20 in load and 20 in no load).

Materials and procedure. The procedure is the same as in Experiment 1, with the exception that participants in the load condition were asked to perform the concurrent task during Phase 4 trials. No concurrent task was used during Phase 3. In Experiment 4B, we modified the Phase 4 instructions to read as follows:

> Lastly, we want you to make some first-stage choices. The boxes you will choose between are the same ones you saw before when you learned which second-stage choices the first-stage choices take you. You will receive money today for making these choices, but you will not be able to immediately see how much each first-stage choice has earned you. Note that the rewards you earn in this phase do not depend on second-stage choices.

Results and Discussion

As shown in Figure 6 and Table 4, we found no evidence in Experiment 4A for a difference between revaluation in the load and no-load conditions, $t(38) = 0.31$, $p = .76$. There was no change in revaluation between the first and second halves of the Phase 4 trials ($p = .98$). We also compared the average preference for B in the load and no-load conditions, again finding no significant difference, $t(38) = 0.75$, $p = .46$.

Of course, the t test merely indicates that we have failed to reject the null hypothesis that revaluation magnitudes should be equal across the two conditions; we would like to assert the null effect. For this purpose, we turn to the Bayesian t test proposed by Rouder, Speckman, Dongchu, and Morey (2009) using the scaled

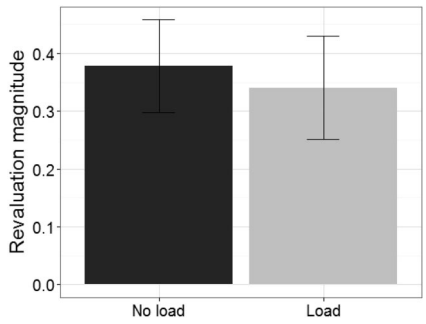

Figure 6. Experiment 4A results. Revaluation magnitude in Phase 4 as a function of working memory load during Phase 4. Phase 3 was performed without load in both conditions. Error bars indicate standard error of the mean.

Table 4
Choice Probabilities in Each Condition and Phase for Experiment 4A, Reported as Mean ± Standard Error

Phase	Load (Phase 4)	No load (Phase 4)
Phase 2	0.12 ± 0.02	0.17 ± 0.03
Phase 3	0.90 ± 0.04	0.91 ± 0.04
Phase 4	0.46 ± 0.09	0.55 ± 0.08

Note. Mean ± standard error is the probability of choosing Action B in State 1 for Phases 2 and 4. For Phase 3, mean ± standard error is the average probability of choosing Action D in State 2 or Action F in State 3 (i.e., the higher value action).

Jeffrey-Zellner-Siow prior, which is based on a Cauchy distribution on the effect size. Using the default scale parameter of 1 for the effect size, the Bayes factor in favor of the null was 4.12. According to Jeffreys' (1961) scale, this constitutes moderate evidence for the null hypothesis.

Essentially, we found the same results in Experiment 4B, in which the instructions were slightly modified to avoid misinterpretations of the Phase 4 reward structure. We failed to find a difference between load and no-load conditions, $t(90) = 0.48$, $p = .64$. A Bayesian t test reinforces this conclusion with a Bayes factor of 5.12 in favor of the null hypothesis. In summary, the results of Experiments 4A and 4B indicate no difference in revaluation behavior as a function of Phase 4 load, disfavoring a pure model-based account of retrospective revaluation behavior.

Model Simulations

As proof of concept that a cooperative reinforcement learning architecture can qualitatively capture our findings, in this section we present simulations of the DYNA algorithm (Sutton, 1990). According to DYNA (see Figure 1), model-based and model-free systems learn value functions in parallel, whereas the model-free system exclusively controls behavior. The role of the model-based system is to train the model-free system in an offline manner by replaying experienced state–action pairs and then simulating transitions and reward from the learned model. This provides sufficient information for the model-free system to compute a prediction error and update its value estimate.

In light of evidence that concurrent cognitive load can debilitate rehearsal of WM contents (Baddeley, 1992), we assumed that our load manipulation affects the learning algorithm by reducing replay. To emulate the effect of the load manipulation, we simulated

Table 5
Choice Probabilities in Each Condition and Phase for Experiment 4B, Reported as Mean ± Standard Error

Phase	Load (Phase 4)	No load (Phase 4)
Phase 2	0.08 ± 0.01	0.10 ± 0.01
Phase 3	0.94 ± 0.01	0.94 ± 0.01
Phase 4	0.48 ± 0.05	0.46 ± 0.04

Note. Mean ± standard error is the probability of choosing Action B in State 1 for Phases 2 and 4. For Phase 3, mean ± standard error is the average probability of choosing Action D in State 2 or Action F in State 3 (i.e., the higher value action).

the DYNA algorithm in our experimental design with different amounts of replay after each trial in Phase 3 (one replay per trial in the load condition, two in the no-load condition). Similarly, we emulated the effect of the rest period by examining revaluation magnitude after varying amounts of replay between Phase 3 and Phase 4. Simulation details are described in the Appendix.[2]

Figure 8 shows the simulated revaluation magnitude as a function of rest period replays for the four different conditions in Experiment 2. Consistent with the results from that experiment, the revaluation magnitude grows with the number of Phase 3 trials and shrinks under load. This pattern reflects that having more trials translates into more opportunities for replay, whereas load is modeled as reducing opportunities for replay. As we found in Experiment 2, the model simulations show no difference between no-load-30 and no-load-50 conditions. Figure 8 also reproduces the main result of Experiment 3: The revaluation magnitude grows with increased numbers of replays during the rest period between Phase 3 and Phase 4. In summary, the DYNA algorithm—Sutton's (1990) instantiation of a cooperative model-free/model-based architecture—is one implemented framework for interactions between model-free RL and model-based RL systems that is consistent with the pattern of data from these studies.

It is worth emphasizing here that, although we did not perform formal model comparison, pure model-based and model-free algorithms cannot in principle explain our data, regardless of their parameter settings, under the assumption we made above. Of course, it is entirely possible that these assumptions are wrong, but they highlight the basic empirical commitments of the different theories. More elaborate versions of these theories, making different assumptions (e.g., participants erroneously perform model-based planning over a two-step horizon), could explain some of our results.

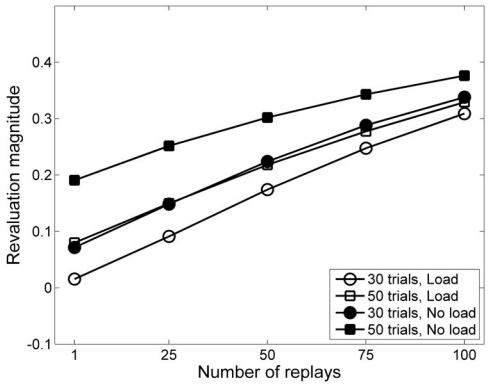

Figure 8. Predictions of the DYNA algorithm for the two-step task. Simulated revaluation magnitude increases as a function of the number of replays between Phase 3 and Phase 4. The simulation was run for different numbers of Phase 3 trials (30/50) and different numbers of replays per Phase 3 trial (one in the load condition, two in the no-load condition).

General Discussion

In four experiments, we explored the conditions under which people exhibit retrospective revaluation in a sequential decision problem. We measured revaluation to the degree to which participants changed their choice preference in the first step of the task after experiencing reward contingencies in the second step of the task. In Experiment 1, participants asked to perform a demanding secondary task during the second-step trials (in Phase 3) exhibited less revaluation than participants who did not have to perform the secondary task. In Experiment 2, the effect of cognitive load was modulated by the number of second-step trials: When given more second-step trials, participants under load exhibited the same amount of revaluation as participants not under load. In Experiment 3, interposing a rest period between Phase 3 and Phase 4 increased the revaluation effect for participants under load. Finally, in Experiments 4A and 4B, WM load during Phase 4 had no effect on revaluation.

Our favored theoretical interpretation of these results is based on the idea that model-free and model-based reinforcement learning systems interact cooperatively (Sutton, 1990). In particular, we propose, along the lines of Sutton's (1990) DYNA algorithm, that the model-based system trains the model-free system offline (e.g., during periods of quiescence). This proposal is motivated by several observations about our experimental design and results. First, Phase 4, in which participants made first-step choices after experiencing the second-step rewards, was restricted to a one-step planning horizon; that is, participants were never given the opportunity to make a full sequence of Step 1 and Step 2 choices. This aspect of the design is important because if participants were using a model-based planning algorithm (e.g., dynamic programming or tree search) over a one-step horizon, they should show no revaluation: The values for this horizon are identical to the immediate rewards, which did not change between Phase 1 and Phase 4. Second, the fact that the revaluation magnitude was sensitive to the rest interval in Experiment 3 is consistent with the idea that offline training occurs during periods of quiescence. This interpretation is bolstered by neural evidence that rats simulate future trajectories while they are asleep or during waking periods of inactivity (see below). Third, load during Phase 4 had no effect on revaluation (Experiment 4), which seems to contradict a model-based account, which would predict sensitivity to load (Otto et al., 2013). We do

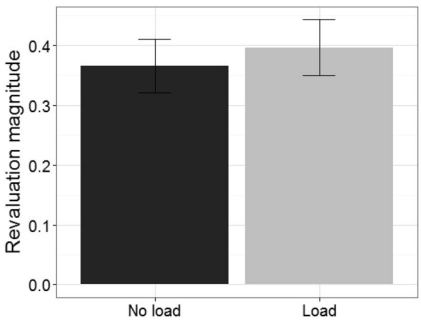

Figure 7. Experiment 4B results. Revaluation magnitude in Phase 4 as a function of working memory load during Phase 4. Phase 3 was performed without load in both conditions. Error bars indicate standard error of the mean.

[2] Our intention in simulating rather than fitting the model to data was to remain as agnostic as possible with respect to the particular algorithmic details of the implementation. Our goal here was to provide a proof-of-concept for a particular *family* of models.

not claim that this is the only cooperative architecture that can capture our findings, or even that our findings rule out all possible competitive architectures. At present, the available data do not sufficiently constrain the space of viable models.

One alternative interpretation of our results is that participants are actually performing model-based planning using an incorrect model of the task. For example, if they erroneously believe that the planning horizon during Phase 4 is two steps instead of one step, then the basic effect of load versus no load in Experiment 1 can be explained as a consequence of impaired model-based learning during Phase 3. Although we cannot rule out this interpretation, it does make an additional prediction for which we found no evidence: Response times should correlate with revaluation magnitude, under the assumption that revaluation will occur to the extent that participants plan over two steps and that planning over two steps is more computationally expensive than planning over one step. We found no correlation between response times and revaluation magnitude. Although this is a null result, and should therefore be interpreted cautiously, it fails to support the alternative theoretical interpretation. A more compelling argument against the model-based account is provided by the results of Experiment 4: Load during Phase 4 does not appear to affect revaluation, which, as we have argued above, is inconsistent with cognitively demanding model-based planning (Otto et al., 2013).

Previous Work

Our experiments were inspired by the seminal devaluation studies of Dickinson and his colleagues (Dickinson, 1985). These studies showed that rats were able to reason about the causal effects of their actions. Rats were first trained to press a lever for food reinforcement, and then the reinforcer was separately paired with illness (i.e., the reinforcer was devalued). After moderate amounts of training, rats ceased pressing the lever following devaluation. After extensive training, however, rats continued to press the lever in order to receive the devalued reinforcer. Dickinson interpreted this as a shift from goal-directed to habitual control with extended training, or, in more modern parlance, a shift from model-based to model-free control (Daw et al., 2005).

Another way to shift control from model based to model free is to tax executive resources by asking participants to perform a demanding secondary task. Fu and Anderson (2008) used a two-stage sequential decision task and showed that under dual-task conditions, participants learned actions proximal to feedback faster than distal actions, whereas this pattern was reversed under single-task conditions. They interpreted these results to indicate the predominance of model-based control under single-task conditions and model-free control under dual-task conditions. The model-free system gradually propagates reward information to earlier states in a trajectory, explaining why second-step values are learned about faster than first-step values under dual-task conditions. In contrast, reward information is propagated instantly by the model-based system.

Using a similar two-step task, but with continuously drifting rewards (which enabled a more fine-grained analysis of the learning process), Otto et al. (2013) found that choice behavior is more consistent with model-free control when people are placed under load—that is, choices were primarily influenced by whether they were previously rewarded (cf. Thorndike's, 1911, law of effect). In contrast, when unfettered by concurrent demands, people's behavior registered sensitivity to the transition structure of the environment, indicating the influence of model-based knowledge (see also Daw, Gershman, Seymour, Dayan, & Dolan, 2011).

Also related to our work is the literature on retrospective revaluation (Larkin, Aitken, & Dickinson, 1998; Shanks, 1985; Van Hamme & Wasserman, 1994). In a typical retrospective revaluation experiment, two cues (A and B) are first trained in compound, and then one of the cues (A) is trained individually. *Retrospective revaluation* refers to changes in the response to B following A-alone training. This phenomenon is important from a theoretical perspective because it contradicts important models of associative learning (e.g., Rescorla & Wagner, 1972) that assert that associations between cues and outcomes are only modified when a cue is present. These findings have prompted the development of models that allow modification to occur in the absence of cues (e.g., Dickinson & Burke, 1996; Markman, 1989; Van Hamme & Wasserman, 1994). Of particular relevance are theories that postulate "rehearsal" of previous experiences, conceptually similar to the replay mechanism of DYNA (Chapman, 1991; Melchers, Lachnit, & Shanks, 2004). However, these theories are "trial-level" models, and hence do not accommodate sequential decision tasks such as the one presented here. In particular, the models do not incorporate a notion of value or cumulative reward, the crucial ingredient of the retrospective revaluation effects described in this article.

Neural Mechanisms

Although there is no direct neural evidence for the cooperative architecture outlined in this article, the brain possesses plausible neural machinery for implementing such an architecture. One candidate is a network centered on the hippocampus and basal ganglia (Johnson & Redish, 2005; Johnson, van der Meer, & Redish, 2007). The role of the basal ganglia in model-free RL is well established (see Niv, 2009, for a review), with different subregions thought to compute various components of the temporal difference learning equations described in the introduction, although the precise nature of these computations is still in dispute. In accordance with the DYNA architecture, we suggest that the hippocampus contributes in two ways: (2) by learning a model of the environment and (b) by replaying and simulating state–action-reward trajectories to the basal ganglia (see Johnson & Redish, 2005, for an implementation of this idea).

The idea that the hippocampus is involved in learning a statistical model of the environment has received considerable theoretical attention (Fuhs & Touretzky, 2007; Gershman, Blei, & Niv, 2010; Gluck & Myers, 1993; Levy, Hocking, & Wu, 2005). Experiments demonstrate that the hippocampus tracks the statistics of transitions (Bornstein & Daw, 2012; Harrison, Duggins, & Friston, 2006; Shohamy & Wagner, 2008; Strange et al., 2008) and rewards (Hölscher, Jacob, & Mallot, 2003), the necessary ingredients for building a Markov decision process model. The hippocampus uses this model to simulate forward trajectories, as evidenced by the sequential firing of place-coding cells along anticipated spatial paths, a phenomenon known as "preplay" (Diba & Buzsáki, 2007; Dragoi & Tonegawa, 2011; Johnson & Redish, 2007). Forward trajectories can also originate from previously experienced locations rather than the animal's current location (Gupta, van der Meer, Touretzky, & Redish, 2010), consistent with

the memory retrieval mechanism posited by DYNA. Furthermore, hippocampal replay appears to be temporally coordinated with activity in the ventral striatum (a subregion of the basal ganglia) during sleep (Lansink, Golstein, Lankelma, McNaughton, & Pennartz, 2009). These physiological findings are complemented by functional brain imaging evidence of hippocampal activation while humans imagine future events (see Buckner, 2010, for a review).

Although we have focused on the hippocampus as the locus of model-based computations, most research on goal-directed learning has focused on an extended network of regions including the dorsomedial striatum and prefrontal cortex (Balleine & O'Doherty, 2010). Lesions to the hippocampus have deleterious effects on goal-directed instrumental behavior in spatial tasks (McDonald & White, 1993). Nonetheless, the role of the hippocampus within the model-based RL system remains poorly understood.

Conclusions

Psychology and neuroscience are rich with variations of dual-systems theories, some of which are closely linked to the distinction between model-based and model-free RL (Dayan, 2009). At the heart of these theories is the idea that the brain can solve RL problems in two different ways: either by performing costly deliberative computations to arrive at the optimal solution or by using cheap habitual solutions that may be suboptimal. The brain, according to dual-systems theories, has evolved separate systems dedicated to these different computational strategies. This inevitably invites the question of how behavioral control is arbitrated between the systems. Recent theoretical work has framed this arbitration as competitive (Daw et al., 2005; Keramati et al., 2011; Simon & Daw, 2011). In contrast, we suggest, on the basis of behavioral evidence, that the interactions between systems might be more cooperative in nature. The cooperative scheme, which we have simulated using Sutton's (1990) DYNA algorithm, is able to capture our behavioral data and has some neural plausibility (Johson & Redish, 2005), though our data do not decisively rule out all competitive schemes. Evidence for cooperation between the two systems has been suggested by a recent brain imaging study (Daw et al., 2011), which found that model-based and model-free values overlap in their neural substrates. An important open challenge is now to find direct evidence of cooperation between reinforcement learning systems in the brain.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology: Section B: Comparative and Physiological Psychology, 34,* 77–98.

Baddeley, A. (1992). Working memory. *Science, 255,* 556–559. doi:10.1126/science.1736359

Balleine, B. W., & O'Dohety, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35,* 48–69. doi:10.1038/npp.2009.131

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Bornstein, A. M., & Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience, 35,* 1011–1023. doi:10.1111/j.1460-9568.2011.07920.x

Buckner, R. L. (2010). The role of the hippocampus in prediction and imagination. *Annual Review of Psychology, 61,* 27–48. doi:10.1146/annurev.psych.60.110707.163508

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 837–854. doi:10.1037/0278-7393.17.5.837

Conway, A. R., 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. doi:10.1016/j.neuron.2011.02.027

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. doi:10.1038/nn1560

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks, 22,* 213–219. doi:10.1016/j.neunet.2009.03.004

Diba, K., & Buzsáki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience, 10,* 1241–1242. doi:10.1038/nn1961

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society: B, 308,* 67–78. doi:10.1098/rstb.1985.0010

Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems. In H. Pashler & R. Gallistel (Eds.), *Stevens' handbook of experimental psychology, third edition, Vol. 3: Learning, motivation, and emotion* (pp. 497–534). New York, NY: Wiley.

Dickinson, A., & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *Quarterly Journal of Experimental Psychology: Section B: Comparative and Physiological Psychology, 37,* 60–80.

Dragoi, G., & Tonegawa, S. (2011). Preplay of place cell sequences by hippocampal cellular assemblies. *Nature, 469,* 397–401. doi:10.1038/nature09633

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11,* 19–23.

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

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature, 440,* 680–683.

Fu, W.-T., & Anderson, J. R. (2008). Dual learning processes in interactive skill acquisition. *Journal of Experimental Psychology: Applied, 14,* 179–191. doi:10.1037/1076-898X.14.2.179

Fuhs, M. C., & Touretzky, D. S. (2007). Context learning in the rodent hippocampus. *Neural Computation, 19,* 3173–3215. doi:10.1162/neco.2007.19.12.3173

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning and extinction. *Psychological Review, 117,* 197–209. doi:10.1037/a0017808

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus, 3,* 491–516. doi:10.1002/hipo.450030410

Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron, 65,* 695–705. doi:10.1016/j.neuron.2010.01.034

Harrison, L. M., Duggins, A., & Friston, K. J. (2006). Encoding uncertainty in the hippocampus. *Neural Networks, 19,* 535–546. doi:10.1016/j.neunet.2005.11.002

Hölscher, C., Jacob, W., & Mallot, H. A. (2003). Reward modulates neuronal activity in the hippocampus of the rat. *Behavioural Brain Research, 142,* 181–191. doi:10.1016/S0166-4328(02)00422-9

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, England: Oxford University Press.

Johnson, A., & Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks, 18,* 1163–1171. doi:10.1016/j.neunet.2005.08.009

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience, 27,* 12176–12189. doi:10.1523/JNEUROSCI.3761-07.2007

Johnson, A., van der Meer, M. A. A., & Redish, A. D. (2007). Integrating hippocampus and striatum in decision making. *Current Opinion in Neurobiology, 17,* 692–697. doi:10.1016/j.conb.2008.01.003

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology, 7,* e1002055. doi:10.1371/journal.pcbi.1002055

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., & Pennartz, C. M. A. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biology, 7,* e1000173. doi:10.1371/journal.pbio.1000173

Larkin, M., Aitken, M., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1331–1352. doi:10.1037/0278-7393.24.6.1331

Levy, W. B., Hocking, A. B., & Wu, X. (2005). Interpreting hippocampal function as recoding and forecasting. *Neural Networks, 18,* 1242–1264. doi:10.1016/j.neunet.2005.08.005

Markman, A. B. (1989). LMS rules and the inverse base-rate effect: Comment on Gluck and Bower (1988). *Journal of Experimental Psychology: General, 118,* 417–421. doi:10.1037/0096-3445.118.4.417

McDonald, R. J., & White, N. M. (1993). A triple dissociation of memory systems: Hippocampus, amygdala, and dorsal striatum. *Behavioral Neuroscience, 107,* 3–22. doi:10.1037/0735-7044.107.1.3

Melchers, K. G., Lachnit, H., & Shanks, D. R. (2004). Within-compound associations in retrospective revaluation and in direct learning: A challenge for comparator theory. *The Quarterly Journal of Experimental Psychology: Section B: Comparative and Physiological Psychololgy, 57,* 25–53. doi:10.1080/02724990344000042

Niv, Y. (2009). Reinforcement learning in the brain. *The Journal of Mathematical Psychology, 53,* 139–154. doi:10.1016/j.jmp.2008.12.005

Otto, A. R., Gershman, S. G., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement learning systems by taxing the central executive. *Psychological Science, 24,* 751–761.

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. doi:10.1016/j.cognition.2010.11.009

Prokofiev, S. (2008). Gavotte Op. 12 No. 2 [Recorded by Sergei Prokofiev]. *On Prokofiev: The Pianist* [CD]. London, England: Future Noise Music.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Rouder, J. N., Speckman, P. L., Dongchu, S., & Morey, R. D. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16,* 225–237. doi:10.3758/PBR.16.2.225

Shanks, D. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology: Section B: Comparative and Physiological Psychology, 37,* 1–21.

Shinners, P. (2011). *PyGame–Python game development*. Available from http://www.pygame.org

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron, 60,* 378–389. doi:10.1016/j.neuron.2008.09.023

Simon, D. A., & Daw, N. D. (2011). Environmental statistics and the trade-off between model-based and TD learning in humans. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 24, pp. 127–135). Retrieved from http://books.nips.cc/nips24.html

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., & Friston, K. J. (2008). Information theory, novelty and hippocampal responses: Unpredicted or unpredictable? *Neural Networks, 18,* 225–230. doi:10.1016/j.neunet.2004.12.004

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

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

Thorndike, E. L. (1911). *Animal Intelligence*. New York, NY: The Macmillian Company.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55,* 189–208. doi:10.1037/h0061626

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience, 29,* 2225–2232. doi:10.1111/j.1460-9568.2009.06796.x

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28,* 127–154.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37,* 498–505.

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. doi:10.1523/JNEUROSCI.0564-07.2007

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25,* 127–151. doi:10.1006/lmot.1994.1008

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Unpublished doctoral dissertation). Cambridge University, Cambridge, England.

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

(*Appendix follows*)

Appendix

Algorithms for Sequential Decision Making

In this Appendix, we briefly summarize the mathematical basis of model-based (dynamic programming) and model-free (Q-learning) algorithms (see Sutton & Barto, 1998, for a comprehensive overview). The environment is formalized as a *Markov decision process* (MDP), which consists of a set of states (e.g., the location of the agent), actions (what the agent can do in each state), and rewards. State transitions and rewards depend only on the current state and action (i.e., transitions and rewards are conditionally independent of an agent's history given its current state and action; this is the Markov property). A common example of an MDP is the game of chess, where moves only depend on the current board configuration. The *value* of a state–action pair (s,a) is defined as the expected future return[3] over a horizon of length H conditional on taking action a in State s:

$$Q(s,a) = E\left[r_t + r_{t+1} + \ldots + r_H \,\middle|\, s_t = s,\ a_t = a\right], \tag{A1}$$

where r_t denotes the reward received at time t, and E denotes an average over future state and reward trajectories. The optimal policy (a mapping from states to actions) is to choose the action that maximizes the value function.

An important consequence of the Markov property is that the value function can be written in a recursive form known as *Bellman's equation* (Bellman, 1957):

$$Q(s,a) = R(s,a) + \max_{a'} \Sigma_{s'} T(s',a,s)\, Q(s',a'), \tag{A2}$$

where $R(s,a)$ is the expected reward for taking action a in State s, and $T(s',a,s)$ is the probability of transitioning to State s' after taking action a in State s. The simplest form of dynamic programming, known as *value iteration*, harnesses this recursion by iteratively updating its value estimate $Q_k(s,a)$ according to:

$$Q_{k+1}(s,a) = R(s,a) + \max_{a'} \Sigma_{s'} T(s',a,s)\, Q_k(s',a'). \tag{A3}$$

It can be shown that $Q_k(s,a)$ will eventually converge to $Q(s,a)$. We refer to T and R as the agent's *model* of the world; in practice, this model is typically learned from experience, in which case dynamic programming operates over estimates of T and R.

The classic example of model-free RL is Q-learning (Watkins, 1989), which updates an estimate of the value function according to:

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha\delta_t, \tag{A4}$$

where α is a learning rate parameter, and

$$\delta_t = r_t + \max_{a'} Q_t(s_{t+1},a') - Q_t(s_t,a_t) \tag{A5}$$

is the temporal difference error at time t.

For the DYNA simulations, we used a learning rate $\alpha = .01$ and a softmax policy:

$$P(a_t = B \,|\, s_t = s) = \frac{\exp\{\beta Q(s,B)\}}{\left[\exp\{\beta Q(s,A)\} + \exp\{\beta Q(s,B)\}\right]}, \tag{A6}$$

with the inverse temperature parameter β set to 2. Transition and reward functions were estimated by maximum likelihood; this simply corresponds to setting the transition and reward functions to their sample averages.[4] Because choices are stochastic, we averaged the simulation results over 1,000 repetitions.

Offline training in DYNA proceeds as follows: (a) A previous state–action pair (s, a) is retrieved with uniform probability; (b) a new state (s') is stochastically selected using the learned transition function; (c) a reward (r') for the new state is stochastically generated using the learned reward function; (d) the value function $Q(s,a)$ is updated using Equation A4. Under load, this cycle of computations is performed once between each experienced transition; under no load, it is repeated twice.

[3] For simplicity, we have ignored temporal discounting here.

[4] The particular method for estimating the transition and reward functions is not important here, as the ground truth in this case is deterministic, and thus any reasonable algorithm (e.g., Bayesian, maximum likelihood) will converge very quickly on the correct solution.

Received March 7, 2012
Revision received October 15, 2012
Accepted October 15, 2012 ■