Neurobiology of Learning and Memory 117 (2015) 4–13

Contents lists available at ScienceDirect

Neurobiology of Learning and Memory

journal homepage: www.elsevier.com/locate/ynlme

Multiple memory systems as substrates for multiple decision systems

Bradley B. Doll [a,b,*], Daphna Shohamy [b], Nathaniel D. Daw [a,c]

[a] Center for Neural Science, New York University, New York, NY, United States
[b] Department of Psychology, Columbia University, New York, NY, United States
[c] Department of Psychology, New York University, New York, NY, United States

ARTICLE INFO

Article history:
Received 19 November 2013
Revised 22 April 2014
Accepted 29 April 2014
Available online 15 May 2014

Keywords:
Reinforcement learning
Relational memory

ABSTRACT

It has recently become widely appreciated that value-based decision making is supported by multiple computational strategies. In particular, animal and human behavior in learning tasks appears to include habitual responses described by prominent model-free reinforcement learning (RL) theories, but also more deliberative or goal-directed actions that can be characterized by a different class of theories, model-based RL. The latter theories evaluate actions by using a representation of the contingencies of the task (as with a learned map of a spatial maze), called an "internal model." Given the evidence of behavioral and neural dissociations between these approaches, they are often characterized as dissociable learning systems, though they likely interact and share common mechanisms.

In many respects, this division parallels a longstanding dissociation in cognitive neuroscience between multiple memory systems, describing, at the broadest level, separate systems for declarative and procedural learning. Procedural learning has notable parallels with model-free RL: both involve learning of habits and both are known to depend on parts of the striatum. Declarative memory, by contrast, supports memory for single events or episodes and depends on the hippocampus. The hippocampus is thought to support declarative memory by encoding temporal and spatial relations among stimuli and thus is often referred to as a relational memory system. Such relational encoding is likely to play an important role in learning an internal model, the representation that is central to model-based RL. Thus, insofar as the memory systems represent more general-purpose cognitive mechanisms that might subserve performance on many sorts of tasks including decision making, these parallels raise the question whether the multiple decision systems are served by multiple memory systems, such that one dissociation is grounded in the other.

Here we investigated the relationship between model-based RL and relational memory by comparing individual differences across behavioral tasks designed to measure either capacity. Human subjects performed two tasks, a learning and generalization task (acquired equivalence) which involves relational encoding and depends on the hippocampus; and a sequential RL task that could be solved by either a model-based or model-free strategy. We assessed the correlation between subjects' use of flexible, relational memory, as measured by generalization in the acquired equivalence task, and their differential reliance on either RL strategy in the decision task. We observed a significant positive relationship between generalization and model-based, but not model-free, choice strategies. These results are consistent with the hypothesis that model-based RL, like acquired equivalence, relies on a more general-purpose relational memory system.

1. Introduction

There can be multiple paths to a decision. For example, as we learn by trial and error about the value of different options, we can select among them based simply on how much reward has previously followed each of them, or instead flexibly reevaluate them in the moment by taking into account their particular expected consequences and the current value of those consequences. The latter strategy allows us to choose flexibly if our current needs or tastes have changed: for instance, if we progress from thirst to hunger. Two distinct classes of control systems developed in the engineering literature, called model-free and model-based reinforcement learning (RL), describe these computationally and representationally different approaches to value-based decision making (Sutton & Barto, 1998).

* Corresponding author. Address: 4 Washington Pl, Room 873B, New York, NY 10003, United States.
E-mail address: bradley.doll@nyu.edu (B.B. Doll).

http://dx.doi.org/10.1016/j.nlm.2014.04.014
1074-7427/

Much evidence in both humans and animals supports the idea that the brain implements both of these approaches (Doll, Simon, & Daw, 2012). In particular, model-free RL theories (Montague, Dayan, & Sejnowski, 1996) maintain an estimate of the net reward previously received following each action, updating it in light of each experience using a reward prediction error signal. These theories explain the behavioral phenomena of habits (Daw, Niv, & Dayan, 2005)—inflexible response tendencies that often arise after overtraining (Adams, 1982)—and, neurally, offer the predominant computational account of the reward prediction error-like phasic responses of dopamine neurons and of similar signals in human fMRI at striatal dopamine targets (Glimcher, 2011). The reward prediction error signal, the difference in reward received and reward expected, is the computational core of model-free RL that drives increases in action values following rewards and decreases following punishments. The phasic bursts and dips of midbrain dopamine neurons following rewards and punishments mirror this teaching signal (Montague et al., 1996). Dopaminergic projections to striatum modulate activation and plasticity (Reynolds & Wickens, 2002) of corticostriatal synapses, driving reward and punishment learning (Hikida, Kimura, Wada, Funabiki, & Nakanishi, 2010). Optogenetic work indicates a causal role for this signaling pathway in reward learning, consistent with the predictions of model-free RL (Steinberg et al., 2013; Tsai et al., 2009). Despite the success of this theory, model-free RL cannot explain more flexible, goal-directed actions that have been demonstrated experimentally using tasks such as latent learning or reward devaluation (Adams, 1982; Tolman & Honzik, 1930), nor the neural correlates of these behavioral effects (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Glascher, Daw, Dayan, & O'Doherty, 2010; Tricomi, Balleine, & O'Doherty, 2009), nor the correlates of in-the-moment evaluation of candidate actions (Pfeiffer & Foster, 2013; van der Meer, Johnson, Schmitzer-Torbert, & Redish, 2010). These latter phenomena, instead, are well explained by model-based RL. Instead of net action values, model-based algorithms learn an "internal model" of the task—how different actions lead to resulting situations or outcomes, and how those outcomes map onto value—which then can be used to evaluate candidate actions' values through a sort of mental simulation or search at decision time.

Many experiments suggest that these two learning approaches trade off differentially, depending on task features such as the amount of training (Adams, 1982; Tricomi et al., 2009), dual-task interference (Otto, Gershman, Markman, & Daw, 2013, perhaps by promoting an advantage to striatal over hippocampal function, Foerde, Knowlton, & Poldrack, 2006), pharmacological interventions (Dickinson, Smith, & Mirenowicz, 2000; Wunderlich, Smittenaar, & Dolan, 2012) or brain lesions affecting areas associated with either system (Yin, Knowlton, & Balleine, 2004; Yin, Ostlund, Knowlton, & Balleine, 2005). Further, these learning approaches also differ spontaneously across individuals (Skatova, Chan, & Daw, 2013). All these results suggest that these two sorts of learning rely on neurally and cognitively dissociable systems (though interacting and sharing some common mechanisms, e.g. Daw et al. (2011)). However, although the cognitive and neural mechanisms supporting model-free RL are reasonably well understood, those supporting model-based RL are currently much less clear.

This profile of relatively better and worse understanding about the two systems is complementary to that for a similar dichotomy in another area of research, memory. Decades of work in cognitive neuroscience concerns the brain's multiple systems for memory. Traditionally, the study of memory systems has focused on a distinction between procedural and declarative memory, which are thought to differ in the kinds of representations they form, the contexts in which they are elicited, and the neural systems that support them (e.g. Squire, 1992; Gabrieli, 1998; Knowlton, Mangels, & Squire, 1996). The distinction between memory systems was initially dramatically supported by observations that declarative memory was impaired in patients with amnesia due to hippocampal damage, while procedural memory was relatively spared in the same patients (Corkin, 1968). This effect has subsequently been demonstrated in amnestics with varying etiologies in numerous incrementally acquired, feedback-driven procedural learning tasks, such as mirror-reading (Cohen & Squire, 1980; Gabrieli, Corkin, Mickel, & Growdon, 1993), the pursuit-rotor (Heindel, Salmon, Shults, Walicke, & Butters, 1989), and weather prediction tasks (Knowlton et al., 1996). In contrast, degeneration of the nigrostriatal dopamine system in Parkinson's disease and loss of striatal integrity in Huntington's disease impairs procedural memory, but leaves declarative memory relatively intact (Heindel et al., 1989; Knowlton et al., 1996; Martone, Butters, Payne, Becker, & Sax, 1984). The dissociable behavioral and neural characteristics of these memory systems are now well-described—procedural memory is thought to be inflexible, implicit, incremental, and reliant on striatum, whereas declarative memories are more flexible, relational, possibly subject to conscious access, and reliant on hippocampus.

Thus, in many ways, this distinction appears to be closely related to that between model-free and model-based RL, with model-free RL corresponding to procedural memory (Knowlton et al., 1996), and model-based corresponding to declarative memory (Daw & Shohamy, 2008; Dickinson, 1980). The computational and neural mechanisms formalized by model-free theories and tasks also explain classic procedural learning tasks, which feature incremental learning from trial-to-trial feedback, and implicate the striatum and its dopaminergic inputs (Knowlton et al., 1996; Shohamy et al., 2004). Evidence that distraction by cognitive load disrupts model-based, but not model-free RL (Otto, Gershman, et al., 2013) is mirrored by evidence that such interference disrupts declarative, but not procedural task learning (Foerde et al., 2006). Better understanding of the relationship between decision making and memory systems has the potential to shed light on both areas, in particular because in memory, much is known about the brain's systems for declarative memory, which might provide a crucial relational encoding mechanism underlying model-based RL. (Conversely, our relatively strong knowledge of the brain's mechanisms for model-free RL may fill in many gaps in our understanding of procedural memory.) Thus, in the present study we aimed to examine evidence for these correspondences by studying the relationship between tasks from the memory and decision literature.

A key feature of the hippocampal memory system, which is particularly relevant to model-based RL is the encoding of relations—associations between multiple stimuli or events (Cohen & Eichenbaum, 1993). A hallmark of relational memories that suggests a parallel to model-based RL is that multiple, previously learned associations can be combined in novel ways. A family of classic memory tasks including acquired equivalence (Honey & Hall, 1989), associative inference (Bunsey & Eichenbaum, 1996), and transitive inference (Davis, 1992) assess this feature of relational memory. In these tasks, subjects first learn sets of overlapping stimulus relationships and then generalize to or infer the relationships between never-before-seen stimulus combinations in a test phase. Neurally, successful performance of these relational memory tasks is associated with the hippocampus (Greene, Gross, Elsinger, & Rao, 2006; Myers et al., 2003; Preston, Shrager, Dudukovic, & Gabrieli, 2004; Shohamy & Wagner, 2008).

These features of relational memory suggest a correspondence to model-based RL. Unlike the response- or reward-related associations underlying model-free learning—which are clearly in the domain of procedural memory (Knowlton, Squire, & Gluck, 1994; Nissen & Bullemer, 1987)—model-based RL relies on learning a world model, that is, an arbitrary set of associations between stimuli or situations (as with a map of a spatial task), which can

be flexibly used to plan actions. Not only is such learning essentially relational, but such a decision making strategy and the laboratory tasks used to probe it bear a striking similarity to generalization or inference in the relational memory tasks described above. This parallel is further supported by suggestive, though mixed, evidence for hippocampal involvement in model-based RL (Corbit & Balleine, 2000; Corbit, Ostlund, & Balleine, 2002; Johnson & Redish, 2007; Simon & Daw, 2011).

Accordingly, in this study, we consider the extent to which multiple decision systems make differential use of multiple memory systems. In particular, we assess the possibility that model-based RL relies on relational memory. If the one distinction arises from the other, this would help to situate model-based learning on a more general foundation of mnemonic mechanisms that have been well-characterized both cognitively and neurally.

To examine whether relational memory might serve as a substrate for model-based RL we compare individual differences across tasks that separately measure these cognitive faculties. In particular, we use an acquired equivalence task, a well-characterized measure of memory generalization that is known to recruit the hippocampus (Myers et al., 2003; Shohamy & Wagner, 2008), to train equivalencies between stimuli, then test whether subjects rely on these equivalences to generalize flexibly to novel probes. We then use the same stimuli as the basis of the world model in a subsequent RL task. Leveraging the substantial between subject variability in both generalization and in model-based RL, we ask whether performance on each of the tasks is related by a common computation or capacity. Our results indeed demonstrate a positive relationship between generalization performance in acquired equivalence and the use of model-based RL, consistent with a role for the relational memory system (and, more speculatively, the hippocampus) in model-based RL.

2. Materials and methods

2.1. Experimental tasks

29 Columbia University students completed two behavioral tasks, and were paid $10 for participation. The study was approved by the local review board. Subjects first completed an acquired equivalence task (a variant of that presented in Myers et al. (2003) and Meeter, Shohamy, & Myers (2009), Fig. 1A, Table 1). This task measures learning of familiar trained associations and generalization to novel ones composed of trained components. Subjects next completed a sequential reinforcement learning task (a variant of the task introduced in Daw et al. (2011), Fig. 1B), which measures the degree to which subjects choices are consistent with model-based or model-free strategy. We investigated the relationship of generalization in the relational memory task to use of model-based and model-free strategies in the RL task across subjects.

In the training phase of the acquired equivalence task, subjects saw three pictures on each trial: a person's face and two outdoor scenes, and were given the goal of learning which of the two scenes was preferred by each person. Subjects were told that each trial type had a single consistent answer. Correct choices were followed by positive feedback (a picture of a dollar coin), while incorrect ones were followed by negative feedback (the dollar coin with an *X* through it).

Equivalence training proceeded in three parts (Table 1). First (shaping phase), subjects completed trials featuring faces *A* and *X*, and scenes S1 and S3 (*A* preferred S1, *X* preferred S3; 8 randomly ordered trials total). Equivalence training followed, in which *A′* and *X′* were presented with scenes S1 and S3 (*A′* preferred S1, *X′* preferred S3; 8 trials, intermingled with a repetition of the first 8 trials). Finally (new consequents), new scenes were introduced. *A* and *X* were presented with scenes S2 and S4 (*A* preferred S2, *X* preferred S4; 8 trials interleaved with a repetition of the first 16 trials).

Following the training phase, subjects completed a test phase in which familiar training stimulus pairs as well as novel combinations of stimuli were presented. Participants were instructed again to select the scene preferred by each person, and were told that they would not receive feedback during this portion of the experiment. The novel trials (*A′* and *X′* with scenes S2 and S4, 16 trials, 8 each) which measure acquired equivalence (*A′* preferring S2, *X′* preferring S4), were intermingled with familiar training trials (8 repetitions of each of the 6 trial types). In training and test, subjects had 5 s to respond.

Next, subjects completed 200 trials of a two-step sequential reinforcement learning task, designed to distinguish between reliance on model-based and model-free strategies. In this task

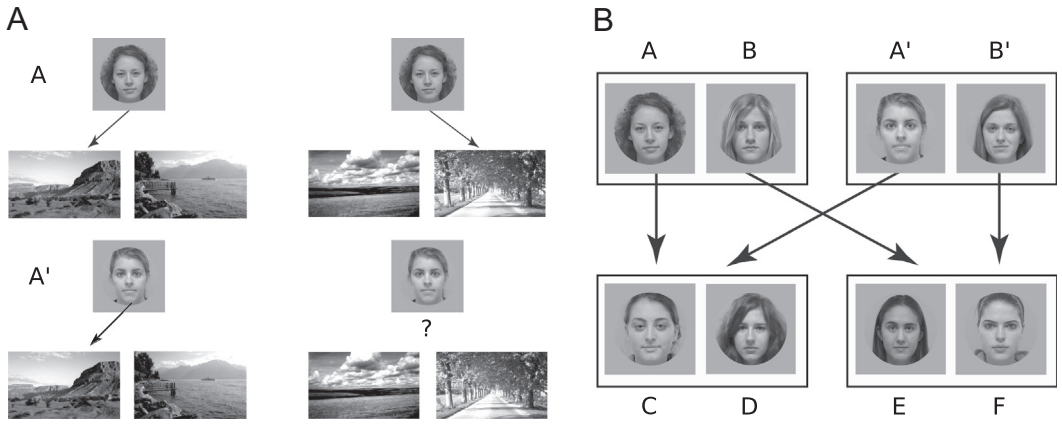

Fig. 1. Task structure. (A) Acquired equivalence. Subjects learned face-scene associations by trial and error and received corrective feedback. Example training trials shown for the *A* face stimulus (upper row), and the *A′* face stimulus (lower left). Note that faces *A* and *A′* are both associated with the same scene (left column; in the task, left/right scene presentation was counterbalanced across trials). In the test phase, participants reported these associations again without feedback. Interleaved with familiar training trials were novel test trials (lower right). Subjects who learned the equivalency of *A* and *A′* generalize the association for *A* (upper right) to this novel test trial. Subjects were also trained on the equivalence relation between *X* and *X′* faces (stimuli not shown; see discussion). (B) Reinforcement learning task. Subjects started each trial on either the stage one states with actions *A*, *B* (left) or the state with actions *A′*, *B′* (right) at random (left/right face presentation was randomized on each trial). Selection of *A* or *A′* transitioned to the stage two state with actions *C*, *D*, while selection of *B* or *B′* transitioned to the stage two state with actions *E*, *F*. Each action in the stage two states produced reward with a continually drifting probability. Note that acquired equivalence of *A* and *A′* is advantageous in the reinforcement learning task, as these actions are functionally equivalent.

Table 1
Acquired equivalence trial structure. Corrective feedback followed choice in acquisition phase. No feedback was given in transfer phase. The transfer phase featured novel probes interleaved with familiar probes from previous acquisition phases. A, A', X, X' denote face stimuli used in subsequent sequential reinforcement learning task. S1, S2, S3, S4 denote scene stimuli.

Acquisition Stage 1: Shaping	Acquisition Stage 2: Equivalence training	Acquisition Stage 3: New consequents	Transfer phase: Equivalence testing
$A \rightarrow$ S1	$A \rightarrow$ S1 $A' \rightarrow$ S1	$A \rightarrow$ S1 $A' \rightarrow$ S1 $A \rightarrow$ S2	$A' \rightarrow$ S2
$X \rightarrow$ S3	$X \rightarrow$ S3 $X' \rightarrow$ S3	$X \rightarrow$ S3 $X' \rightarrow$ S3 $X \rightarrow$ S4	$X' \rightarrow$ S4

subjects were given the task of winning as many virtual coins as possible. Each trial consisted of two stages. The first stage began with the presentation of one of two initial start "states" (choice sets). In each state, subjects chose between one of two face stimuli (state 1: actions A, B; state 2: actions A', B', where A and A' but not B and B' were previously used in the acquired equivalence task, see discussion). Whether a trial started with state 1 or 2 was alternated randomly over trials. Selection of one of the two stimuli in the start state transitioned subjects to a second stage consisting of one or the other of two possible states, 3 and 4. The relationship between the first-stage choices and the second-stage states was deterministic, with the selection of stimulus A or A' always producing state 3, and selection of B or B' always producing state 4.

At the second stage, subjects chose between another two face stimuli (state 3: C, D; state 4: E, F; 2.5 s response window at all stages) in an attempt to receive reward (either one coin or nothing, as in acquired equivalence, displayed for 1.5 s). The probability of winning a coin for the choice of each second-stage stimulus diffused randomly and independently over the course of the experiment between 0.25 and 0.75 according to a Gaussian random walk with reflecting boundary conditions. Thus participants were incentivized to learn constantly in order to select the first- and second-stage actions most likely to produce reward.

The main differences between this task and the one used by Daw et al. (2011) is the addition of the second start state, state 2, and the elimination of any stochasticity in the transition to second-stage state conditional on the first-stage choice. In this design, as detailed below, the pattern by which learning about choices generalizes across the two first-stage states (like the pattern of response to rare transitions in the task of Daw et al. (2011)) differentiates model-based from model-free strategies to the RL problem. The utilization of functionally equivalent start states, and the pattern of generalization across them, parallels generalization to equivalent associations in the acquired equivalence task.

2.2. Statistical analysis

For parameter estimation in our models of each task (and their relationship) we utilized Markov Chain Monte Carlo (MCMC) methods as implemented in the Stan modeling language (Stan Development Team); this software uses a variant of Hamiltonian Monte Carlo called the No-U-Turn sampler. Given an arbitrary generative model for data (here, a computational model of task performance) dependent on free parameters, together with the data themselves, this method permits samples to be drawn from the posterior probability distribution of parameter values, conditional on the observed data. Using these samples—and notably, the range over which they are distributed, which reflects uncertainty about their estimated values—we can construct confidence intervals (sometimes called "credible intervals" in this setting) over the likely values of the free parameters (Kruschke, 2010).

Because MCMC methods can be applied to arbitrary generative models, they are particularly suitable to analyses, like ours, of data with complex, hierarchical structure (here, the nesting structure implied by the inclusion of two tasks for each participant, each with its own free parameters and each with multiple trials) that are not readily treated by traditional estimation methods (Gelman & Hill, 2007). In particular, the ultimate analysis in the current study involves correlating parameter estimates across both tasks; using MCMC to compute the correlation within a larger, hierarchical model of the full dataset correctly takes account of the uncertainty in each estimate in a way that estimating both tasks separately and then correlating the point estimates would not.

For each model, we produced three chains of 50,000 samples each, retaining only every 10th sample to reduce autocorrelation. The first 1000 samples from each chain were discarded for burn-in, leaving 4900 samples per chain. We verified the convergence of the chains by visual inspection, and additionally by computing for each parameter the "potential scale reduction factor" $\widehat{R}$ of (Gelman & Rubin, 1992). For all parameters, we verified that $\widehat{R} < 1.1$, a range consistent with convergence (Gelman, Carlin, Stern, & Rubin, 2003).

We computed confidence intervals on the parameter estimates of interest using the quantiles of the samples, and in particular, identifying the region from the 2.5th to the 97.5th percentile as the 95% interval. Because this interval contains 95% of the mass of the posterior distribution, it indicates a 5% likelihood that the true value of the parameter lies outside the region, and is in this sense comparable to a traditional confidence interval.

2.3. Acquired equivalence model

We analyzed the acquired equivalence test data in a hierarchical logistic regression predicting task accuracy for novel trials (where generalization is counted as an accurate choice), and on familiar trials, using a binary indicator (0, 1: novel, familiar) to characterize the increment or decrement on these trials relative to the novel baseline. This model predicts the probability of accurate choice on each test trial t as

$$P(acc_t = 1) = logit^{-1}\left(\beta^{nov}_{sub} + \beta^{fam}_{sub} * familiar_t\right) \quad (1)$$

For each subject *sub*, this model has two free parameters, an intercept (novel trials, β^{nov}) and slope (familiar trials, β^{fam}), which estimate, respectively, the average performance on novel trials and the increment or decrement in performance on familiar trials relative to novel ones.

2.4. Reinforcement learning model

As is standard in such tasks (Daw, 2011), we characterized each subject's trial-by-trial choices in the RL task by fitting them with an RL model that learns values for actions from the sequence of rewards received by the subject, and chooses probabilistically according to these values. The model we used is based on the hybrid model of Glascher et al. (2010) and Daw et al. (2011) specialized to the design of the task at hand. The model differs from that of Daw et al. (2011) in several ways. First, we do not model learning of the transition structure in the current task as it is deterministic, fixed, instructed, and practiced before subjects begin. Second, update terms in the equations are rescaled by learning rates to facilitate parameter estimation (see below). Third,

the mixture of model-based and model-free is here implemented as an algebraic equivalent of that used previously in order to independently model the relationship of each strategy to acquired equivalence (see also Otto, Raio, Chiang, Phelps, & Daw, 2013).

The hybrid RL model assumes that subjects' choices are drawn from a weighted combination of both model-based and model-free approaches, with the relative weighting (and other free parameters governing learning) estimated separately for each subject. These approaches make different predictions about behavior in the sequential learning task. The model-based learner treats the stage-one stimuli in terms of the transitions they produce. As a result, learning is effectively generalized across the two different (but functionally equivalent) start states, A, B and A', B'. The model-free learner, in contrast, learns action values on the basis of their outcomes, and thus cannot generalize between the start states. We describe our modeling of each approach for the current task in turn.

2.4.1. Model-free component

The model-free system, SARSA (λ) (State-Action-Reward-State-Action temporal difference learning with eligibility trace parameter λ; Rummery & Niranjan, 1994), learns a value Q^{MF} for each of the two actions a in each of the four states s (each at one of the two stages i) by incrementally updating each action value in accordance with the outcomes received. Under this updating scheme, the model will not generalize across equivalent actions in the two start states, as the value of each is determined by the different history of rewards received from selecting the different actions. In particular, for each subject, the chosen action value in each state is updated at each stage i of each trial t as

$$Q^{MF}_{(s[i,t],a[i,t])} = Q^{MF}_{(s[i,t],a[i,t])} + \alpha_{sub}\delta_{i,t} \tag{2}$$

where

$$\delta_{i,t} = \left(r_{i,t} + Q^{MF}_{(s[i+1,t],a[i+1,t])}\right) \Big/ \alpha_{sub} - Q^{MF}_{(s[i,t],a[i,t])} \tag{3}$$

Parameter α ($0 \geqslant \alpha \geqslant 1$) is the learning rate that controls how rapidly action values, Q^{MF}, are updated. The quantity δ is the prediction error—the discrepancy between received and estimated reward. Note that Eq. (3) rescales the reward and next-state Q-value by the subject's learning rate, α_{sub} (Camerer, 1998; Den Ouden et al., 2013). This does not change the data likelihood, but has the effect of rescaling the inverse temperature parameters β_{sub} in the choice rule, Eq. (6) below. This reduces their correlation with α_{sub} and facilitates group-level modeling. The prediction error equation specializes differently to the first and second stages of the trial. As no rewards are delivered on arrival at the second stage, the first-stage reward $r_{1,t} = 0$, and the first-stage prediction error $\delta_{1,t}$ consists only of the Q^{MF} terms. In contrast, following the second stage, the $r_{2,t}$ is the reward actually delivered (1 or 0), but the term $Q^{MF}_{(s[i+1,t],a[i+1,t])} = 0$, as there are no further stages in the trial.

We additionally employed an eligibility trace parameter λ for each subject permitting the update of the action selected in the first stage via second stage δ values:

$$Q^{MF}_{(s[1,t],a[1,t])} = Q^{MF}_{(s[1,t],a[1,t])} + \lambda_{sub}\delta_{2,t} \tag{4}$$

A rescaling of this update by $1/\alpha_{sub}$ is again required to match the rescaling of the reward terms mentioned above; this results in the omission of the learning rate α_{sub} which would otherwise appear here. Altogether, this portion of the model has two free parameters, α_{sub} and λ_{sub}, with separate values for each subject.

2.4.2. Model-based component

In contrast to the incremental updating undertaken by a model-free system, a model-based system prospectively computes the values of actions from a "world model" detailing which states are likely to follow each candidate action, and what rewards are likely to be received there. In the current task, this corresponds to computing the value for actions at the first stage by retrieving the (model-free) estimates of the immediate rewards at the corresponding second-stage states. Thus, for each stage-one state s and action a, the model-based Q value is:

$$Q^{MB}_{s,a} = \max_{a' \in a1,a2} Q^{MF}_{S(s,a),a'} \tag{5}$$

where $S(s,a)$ is the stage-two state that would be produced by choosing action a in the stage-one state s. Because the model-based system utilizes the task transition structure to evaluate start state actions in terms of the end states they produce, it effectively generalizes across the equivalent start states in the task. In contrast to previous work utilizing this model (Daw et al., 2011; Glascher et al., 2010), here we do not model learning of the transition structure, as subjects are instructed that the transitions are fixed and deterministic and additionally complete practice trials navigating through the stages before beginning the task.

2.4.3. Choice rule

The value estimates of the model-based and model-free systems were combined in the choice rule, each weighted by free parameters β^{MB}_{sub} and β^{MF}_{sub}. These characterize the extent to which each strategy influenced choices, and allow the use of either strategy to vary across subjects. (Note that this formulation is algebraically equivalent to the one used in Glascher et al. (2010) and Daw et al. (2011), under the change of variables $\beta^{MB} = w\beta$ and $\beta^{MF} = (1-w)\beta$.)

The probability of selecting action a in the first stage of a trial was given by a softmax (logistic) rule:

$$P(a_{1,t} = a|s_{1,t}) = \frac{\exp\left(\beta^{MF}_{sub}Q^{MF}_{s[1,t],a} + \beta^{MB}_{sub}Q^{MB}_{s[1,t],a}\right)}{\Sigma_{a'}\exp\left(\beta^{MF}_{sub}Q^{MF}_{s[1,t],a'} + \beta^{MB}_{sub}Q^{MB}_{s[1,t],a'}\right)} \tag{6}$$

At the second stage, model-based and model-free learning coincide, and the choice rule is in terms of the model-free value estimates weighted by a single parameter β^{stage2}_{sub}:

$$P(a_{2,t} = a|s_{2,t}) = \frac{\exp\left(\beta^{stage2}_{sub}Q^{MF}_{s[2,t],a}\right)}{\Sigma_{a'}\exp\left(\beta^{stage2}_{sub}Q^{MF}_{s[2,t],a'}\right)} \tag{7}$$

Altogether, the choice rule has three free parameters per subject, β^{MF}_{sub}, β^{MB}_{sub}, and β^{stage2}_{sub}.

2.5. Group-level modeling and estimation

We have so far described single-subject models of the two tasks, each with a number of free parameters that are instantiated individually for each subject. To estimate their parameters, we specified a hierarchical ("random effects") model, in which each subject-specific parameter (e.g., β^{MB}_{sub}) was assumed to be drawn from a population-level distribution. For each parameter with infinite support (the βs), this was a Gaussian distribution characterized by two group-level free parameters, a mean and a standard deviation. For the parameters that ranged between 0 and 1 (α and λ), the group-level distributions were beta distributions, $Beta(a,b)$.

We jointly estimated the posterior distribution over the individual and group-level parameters using MCMC as described above; this requires further specifying prior distributions ("hyperpriors") on the parameters of the group level distributions. We took these to be uninformative. Specifically the priors for the mean and standard deviation on normal distributions were normal (mean = 0,

standard deviation = 100) and a half-cauchy (location = 0, scale = 2.5), respectively. For α and λ, we estimated the group-level parameters a and b using a change of variables that characterizes the distribution's mean $P_1 = \frac{a}{a+b}$ and spread $P_2 = \frac{1}{\sqrt{a+b}}$, the latter approximating its standard deviation. This allowed us to take as uninformative hyperpriors the uniform distributions $P_1 \sim U(0,1)$ and $P_2 \sim U(0,\infty)$ (the latter improper) (Gelman et al., 2003).

2.6. Cross-task modeling

We combined the hierarchical models of group level behavior in both tasks to investigate the correlation between generalization in the acquired equivalence task and both model-based and model-free choice in the RL task. Parameters were estimated within and across each task in a single statistical model, which incorporates the models described above together with terms modeling their relationship.

To characterize cross-task correlation, the generative model allowed for the possibility that the parameter controlling novel generalization performance in the acquired equivalence task might also affect performance in the reinforcement learning task, with the extent of this cross-task interaction governed by a free parameter. Specifically, acquired equivalence generalization estimate, β_{sub}^{nov} (having first been mean-subtracted across subjects) was taken as a covariate in modeling the reinforcement learning data. This measure of acquired equivalence entered into the choice rule (Eq. (6)) for the RL task by adding weight to the β_{sub} terms,

$$P(a_{1,t}=a|s_{1,t}) = \frac{\exp\left[\left(\beta_{sub}^{MF}+\beta^{crossMF}\beta_{sub}^{nov}\right)Q_{s[1,t],a}^{MF}+\left(\beta_{sub}^{MB}+\beta^{crossMB}\beta_{sub}^{nov}\right)Q_{s[1,t],a}^{MB}\right]}{\Sigma_{a'}\exp\left[\left(\beta_{sub}^{MF}+\beta^{crossMF}\beta_{sub}^{nov}\right)Q_{s[1,t],a'}^{MF}+\left(\beta_{sub}^{MB}+\beta^{crossMB}\beta_{sub}^{nov}\right)Q_{s[1,t],a'}^{MB}\right]} \quad (8)$$

These added terms, $\beta_{sub}^{crossMF}$ and $\beta_{sub}^{crossMB}$ characterize the cross-task correlation between generalization in the acquired equivalence task and the tendency for subjects to make model-free or model-based choices, respectively. We estimated them each with uninformative Gaussian priors, $\beta^{cross} \sim N(0,100)$.

3. Results

29 Healthy volunteers performed a memory task and an RL task. The memory task (Fig. 1A, Table 1), involved paired-associate learning between faces and locations, with an acquired equivalence structure, such that the faces were implicitly grouped in two pairs with equivalent scene associations. This allowed us to test, in novel probe trials, to what extent subjects generalized associations that had been learned about one member of each pair to its partner.

The RL task (Fig. 1B) was a two-stage Markov decision task. On each trial subjects made an initial choice between two options, which determined which of two additional choices they would face next. These second stage choices were rewarded stochastically (with virtual coins), and subjects learned by trial and error how to make the sequences of choices most likely to be rewarded. The structure of the task allows characterization of decision strategy as model-free (e.g. temporal difference learning), in which the long-run values of the first stage options are learned directly, or model-based, in which the values of the first stage options are computed prospectively in terms of the value of the second-stage choices they lead to.

Having characterized individual differences in generalization and in model-based learning, we tested the extent to which these two behavioral tendencies were correlated across subjects.

3.1. Acquired equivalence

First, we assessed performance in the acquired equivalence test phase. In this phase, subjects are shown familiar and novel combinations of stimuli, and asked to indicate the correct face-scene associations without feedback. Familiar trials are repetitions of trials seen during training and serve as a measure of learning. Novel trials are new combinations of previously seen stimuli, the overlapping relationships of which permit measurement of the extent to which subjects generalize according to the equivalence structure implicit in the training stimuli.

We used logistic regression to measure the frequency of correct responses on familiar and novel trials (where the "correct" response in the novel trials is the one consistent with the acquired equivalence structure). The parameters were taken as random effects, instantiated once per subject, and we report the group-level means. The group level mean of the regression coefficient for novel trials, β^{nov} was positive (mean: 1.15, 95% CI on the mean: 0.42, 1.9), where 0 would indicate chance performance, indicating an overall tendency toward above-chance performance on novel acquired equivalence transfer trials. The coefficient β^{fam}, which estimates the performance on familiar associations as an increment (or decrement) over that for novel probes, was also positive and nonzero (mean: 1.3, CI: 0.84, 1.79), indicating better test phase performance on training trials than transfer trials. Thus, at the group level, both retention of familiar training trials and generalization on novel trials was greater than chance, though considerable variability between individuals was observed (Fig. 2), consistent with prior reports (Shohamy & Wagner, 2008).

3.2. Sequential RL task

The RL task was designed to dissociate model-based from model-free learning strategies. Previous studies using a similar task have demonstrated evidence that behavior is governed by a combination of both approaches.

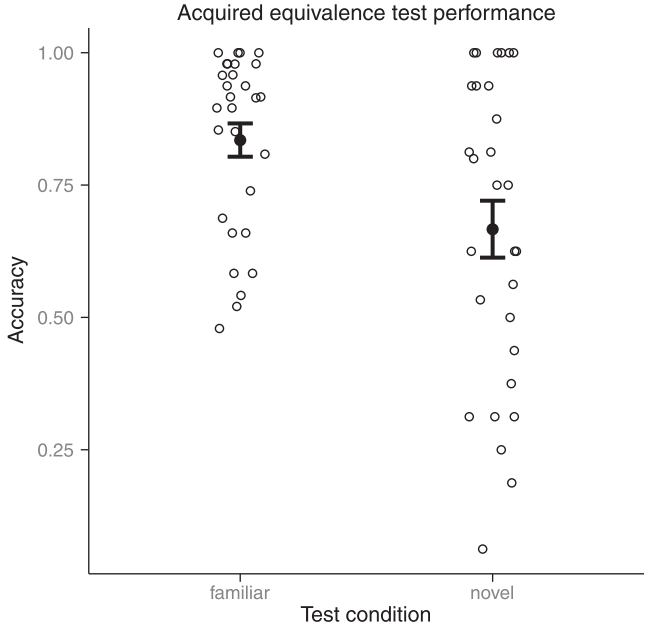

Fig. 2. Variability in acquired equivalence test phase performance summarized as subject and group means for visualization. In the test phase, subjects indicated face-scene associations for familiar and novel stimulus pairings in the absence of feedback. Novel trials measure generalization to new associations on the basis of overlapping learned associations. Filled circles indicate sample mean, and error bars reflect standard error. Individual subjects illustrated by unfilled circles.

In this task, model-based and model-free strategies are dissociable by their predictions as to how choice preferences at the first stage should be learned from feedback. The key difference is that the a model-based learner will generalize feedback across the two first-stage states, where a model-free learner will not. Specifically, the model-based approach evaluates first-stage actions in terms of the value of the second-stage choices to which they lead. Effectively, this approach does not distinguish between the corresponding actions at the two different stage-one start states, as they are identical in terms of the transitions they afford (A and A' transition to one of the stage-two states; B and B' transition to the other). Because the start states are functionally equivalent, they should not affect choice behavior in any way. If a subject using this strategy obtains reward after initially selecting action A on an A, B trial, this will increase the probability of choosing A again, and also of choosing A', since the probability of selecting A' on an A', B' trial that follows should be identical to selecting A if another A, B trial followed instead.

The model-free approach, in contrast, learns unique value estimates for each of the actions in each of the states. By separately estimating the values of the equivalent stage one actions, A will differ from A' and B from B' as their values are updated at different times, from different stochastic sequences of outcomes. A subject utilizing this strategy in the circumstance described above would have an increased chance of choosing A if A was ultimately rewarded, but this tendency would not carry over to A', B' trials. Instead, choice in that start state would be based on the value estimates made from the unique history of A', B' trials.

We fit subjects' trial-by-trial choices to an RL model that allowed for the possibility that choices could be influenced by either model-based or model-free values learned from previous outcomes (Daw et al., 2011; Glascher et al., 2010). Both approaches significantly affected choices at the group level (β^{MF} mean: 0.49, CI: 0.24, 0.75; β^{MB} mean: 0.96, CI: 0.61, 1.33). Fig. 3 illustrates the effects of reward and change of start state on stage-one choice. A higher probability of repeating the stage-one choice made in the previous trial is observed following rewards than punishments. Both the model-based and model-free strategies predict this effect when the same start state repeats from one trial to the next (e.g. A, B followed by A, B). However an increase is also observed when the start states change from one trial to the next (e.g. A, B followed by A', B'), as uniquely predicted by the model-based strategy.

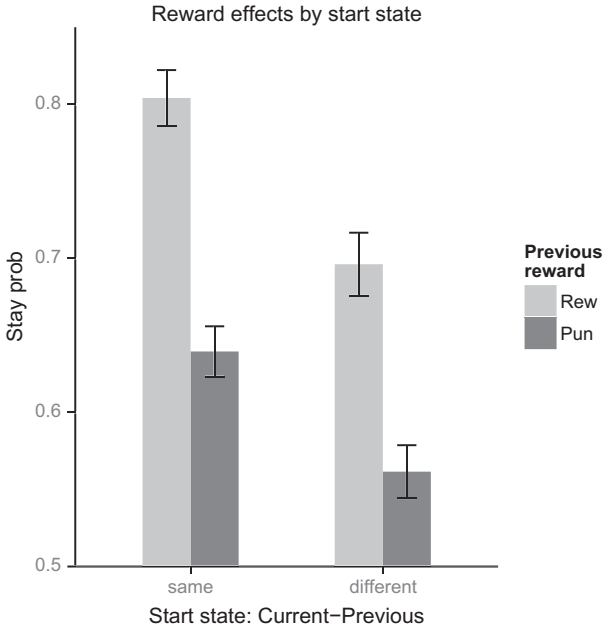

Fig. 3. Signatures of model-based and model-free strategy in the sequential reinforcement learning task summarized as probability of persisting with previous stage-one choice. Subjects tended to stay with their prior first stage choice when it led to a rewarded second stage choice (Rew; light gray) relative to a punished one (Pun; dark gray). This effect is observed when the trial start state was the same as on the previous trial (left bars), a tendency attributable to both model-free and model-based accounts. The effect was also observed when the trial start state was different than that on the previous trial (right bars). Only the model-based strategy explains this latter effect. Error bars reflect within-subject standard error of the mean.

3.3. Cross-task correlation

Our main aim was to assess the relationship between acquired equivalence and model-based choice in the multi-stage RL task. To investigate this relationship we fit data from both tasks simultaneously (see Table 2 for summary statistics), and estimated whether the degree of generalization in the acquired equivalence task predicted the degree of use of the model-based strategy in the RL task.

In accordance with our hypothesis, there was a positive relationship between generalization and model-based learning ($\beta^{crossMB}$, mean: 0.22, CI: 0.03, 0.43; Fig. 4B), indicating that acquired equivalence was positively predictive of model-based choice. Also in accord with our hypothesis, no significantly nonzero relationship was observed between acquired equivalence and model-free choice ($\beta^{crossMF}$ mean: 0.1, CI: −0.05, 0.26; Fig. 4C). However, the comparison of these effects to each other, while producing a positively skewed difference between the model-based and model-free effects, included zero in its confidence interval (mean: 0.12, CI: −0.14, 0.39).

Fig. 4A illustrates this effect in the raw choices; plotted is the degree to which reward (vs nonreward) predicts sticking with a choice either within the same start state (as predicted by both model-based and model-free RL) or across different states (a measure specific to model-based RL). Subjects with high acquired equivalence scores show about equal effects in either case, whereas subjects poorer at acquired equivalence show a reduced learning effect mainly when the start state is different, indicating a lesser proportion of model-based RL. Together, these results are consistent with the hypothesis that model-based learning and acquired equivalence share a common cognitive basis.

One possibility is that this relationship is driven by some relatively generic source of variation, such as a hypothetical group of poorly performing or undermotivated subjects who were deficient at both tasks. We addressed this possibility by repeating the analysis on a subset of high-performing subjects. Specifically, we included only subjects who exceeded median performance on the familiar test trials of the acquired equivalence task. This criterion allows for the assessment of correlation between generalization and reinforcement learning strategy exclusive of subjects who performed poorly in simple stimulus–response learning. Despite

Table 2
Group level parameter estimates for cross-task model. Medians, and 95% confidence intervals indicated.

Parameter	2.5%	50%	97.5%
β^{MB}	0.64	0.95	1.28
β^{MF}	0.26	0.5	0.75
β^{stage2}	0.49	0.8	1.11
α	0.69	0.81	0.89
λ	0.33	0.44	0.54
β^{nov}	0.41	1.14	2.0
$\beta^{crossMB}$	0.03	0.22	0.43
$\beta^{crossMF}$	−0.05	0.1	0.26

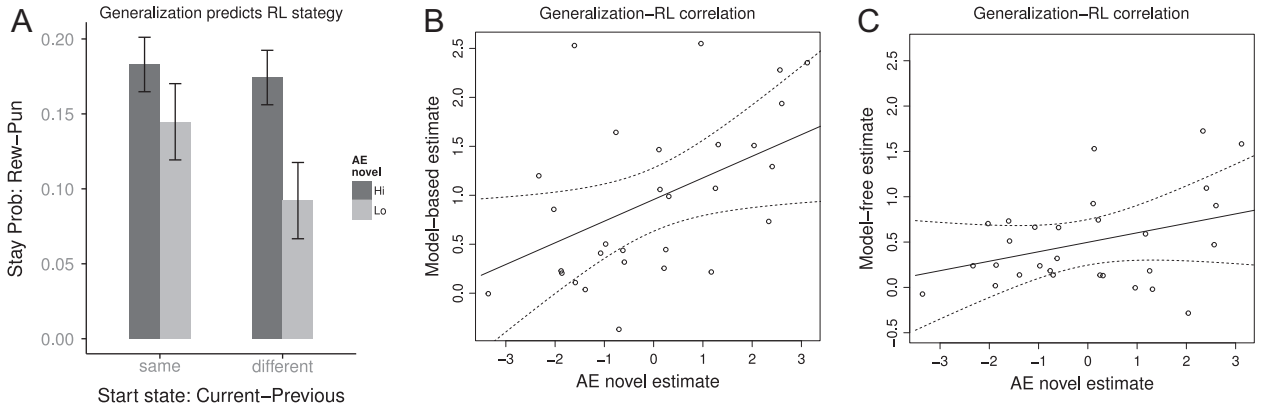

Fig. 4. Generalization in acquired equivalence predicts model-based choice in the sequential reinforcement learning task. (A) Effects summarized as difference scores (previous reward–previous punishment) in probabilities of staying with the first stage choice made on the previous trial. Subjects divided by median split on generalization scores (novel trials in acquired equivalence test phase. High generalization: dark bars, Low generalization: light bars). Left bars: Reward affected stay probability similarly for subjects with high and low generalization when the start state matched that on the previous trial (predicted by both model-free and model-based strategies). Right bars: Reward effects were larger for subjects with high relative to low generalization scores on trials where the start state differed from that on the previous trial (predicted exclusively by the model-based strategy). Error bars reflect within-subject standard error of the mean. (B) and (C) Correlation between model-based (B) and model-free (C) coefficient estimates in RL model and acquired equivalence generalization score. Solid lines indicate group level linear effects, with 95% confidence curves illustrated in dashed lines. Dots represent individual subject estimated effects.

the loss of power inherent in eliminating half of the group, the same relationship between acquired equivalence and model-based (but not model-free) choice was observed even in these high-performing subjects ($\beta^{crossMB}$ mean: 0.26, CI: 0.076, 0.49; $\beta^{crossMF}$ mean: 0.09, CI: −0.18, 0.38; difference CI: −0.15, 0.53).

4. Discussion

Although it has repeatedly been demonstrated that humans and animals can make decisions guided by a world model—that is, information about the contingency structure of the environment, as with a map of a spatial task or the abstract state transition graph in our task here—the cognitive and neural systems supporting this behavior are not well understood. Motivated by the observation that a world model consists of a set of relations between stimuli, we tested the hypothesis that model-based RL would be supported by more general cognitive mechanisms for relational memory. Consistent with this hypothesis, we found that generalization ability in a relational memory task—a hallmark of flexible, hippocampally dependent relational learning—correlated with the use of a model-based choice strategy in a separate RL task.

Although the present experiment was motivated by the appropriateness of relational memory for storing the world model, the two tasks studied here may be related not just by the sorts of memories on which they draw, but also the sorts of computations they perform on them. Indeed, evaluating actions in model-based RL and generating the response to novel probes in acquired equivalence or similar generalization tasks both require a similar process of combining information from multiple learned associations to arrive at an answer. The mechanisms by which both processes occur are under active investigation, with parallels between the two sets of hypotheses.

For acquired equivalence (and related relational memory tasks), there are two major classes of models that differ in their claims about when the generalization takes place. Some researchers have suggested that generalization depends on active generalization at the time of probe (Greene et al., 2006), e.g., in the current task, when probed about the associations of face A', subjects actively interrogate the associations learned about the related face, A. Others have asserted that generalization is instead a by-product of reactivating overlapping episodes during encoding (Shohamy & Wagner, 2008). By this "integrative' encoding" account, acquired equivalence training results in representations of related stimuli (here, equivalent faces A and A') that are coactivated by a process of pattern completion or associative spreading during initial training. In this way, both stimuli come to be, in effect, represented by a common code and any learning about one (that A is associated with some particular scene) is also encoded for its equivalent partner (A'). Generalization for the novel probes then simply consists of retrieving these pre-generalized associations at test time, without the need for active generalization.

Although an RL task does not generally block training vs probes separately, decisions and feedback about them are separate in time, and influences of world model information on decision behavior might be supported by computations at either timepoint, analogous to the two accounts of acquired equivalence. The standard form of model-based RL evaluates actions at decision time by an iterative search of the world model; this parallels the view of acquired equivalence in which active generalization occurs when novel pairings are presented. In contrast, a number of variant algorithms pre-compute this search; for example, DYNA and prioritized sweeping (Moore & Atkeson, 1993; Sutton, 1990), though these have received less attention in psychology and neuroscience (Gershman, Markman, & Otto, 2012). Given evidence of neural overlap between the approaches in human behavior (Daw et al., 2011), strategies in which model-based and model-free methods interact (e.g. DYNA) are of particular interest. Future work should further explore these hybrid methods. One alternative to prospective model-based RL is the "successor representation" (Dayan, 1993), which codes options in terms of their consequences, producing overlapping representations between different states and actions with common consequences (e.g., here, A and A') and allowing feedback to generalize from one to the other without explicit search at decision time, analogous to the integrative encoding view of acquired equivalence. This scheme would produce behavior similar or identical to traditional model-based prospective computation on the current task and many others.

The current experiment contained one feature that might have detected such a representational effect on model-based planning. Of the equivalent top state actions in the RL task, only one set (A, A') was learned about in the acquired equivalence task. This design feature permitted us to investigate whether the correlation between generalization and model-based RL was larger on the trained (A, A') than untrained (B, B') stimuli. Such a result would have indicated that rather than simply exercising a common system or cognitive capacity as model-based learning, the

equivalence training itself specifically promoted model-based choice, e.g. by promoting representational overlap between A and A'. No significant difference was found (analyses not reported). However, this null finding cannot rule out the possibility that model-based RL was supported by representational overlap; for instance B and B' may have themselves rapidly developed representational similarity during the early trials of the sequential task. Clarification of the specific computational mechanism underlying the cross-task relationship observed here awaits further research.

Although the present study collected no neural measurements, the results are nevertheless suggestive about the underlying neural substrate. In particular, the hippocampus has been implicated widely in relational memory (Cohen & Eichenbaum, 1993; Eichenbaum, Yonelinas, & Ranganath, 2007), and specifically in supporting flexible generalization in acquired equivalence (Myers et al., 2003; Shohamy & Wagner, 2008) and similar tasks, such as associative inference and sensory preconditioning (Preston et al., 2004; Wimmer & Shohamy, 2012; Zeithamova, Schlichting, & Preston, 2012). Thus, the finding that generalization in acquired equivalence predicts model-based learning suggests hippocampal involvement also in the latter function. Although the hippocampus surfaces relatively rarely in the literature on RL and decision making (compared to striatal and ventromedial prefrontal areas), it has been implicated in several studies with elaborated task preparations (Lee, Ghim, Kim, Lee, & Jung, 2012; Tanaka et al., 2004; Wimmer, Daw, & Shohamy, 2012; Wimmer & Shohamy, 2012), including those formally examining model-based decisions (Bornstein & Daw, 2013; Simon & Daw, 2011). Moreover, hippocampal involvement is plausible in light of the broader pattern of previous results, which implicate it both in representing models and utilizing those representations for planning. The concept that hippocampus supports a cognitive map for spatial navigation (Burgess, Maguire, & O'Keefe, 2002; O'Keefe & Nadel, 1979) links the structure to some of the earliest demonstrations of model-based behavior, which involved planning routes in spatial mazes (Tolman, 1948). Hippocampal representations have been shown to run ahead of animals' locations in spatial mazes, which has been interpreted as a direct neural correlate of model-based evaluation of candidate trajectories. (Johnson & Redish, 2007; Pfeiffer & Foster, 2013). Notably, this region does not merely represent space, but also appears to be involved in learning statistical relationships between stochastic events in general (Bornstein & Daw, 2012; Schapiro, Kustner, & Turk-Browne, 2012; Staresina & Davachi, 2009; Turk-Browne, Scholl, Chun, & Johnson, 2009; Wimmer & Shohamy, 2012) and with drawing on memory for imagining future events (Addis, Wong, & Schacter, 2007; Hassabis, Kumaran, Vann, & Maguire, 2007). Such knowledge of the transition structure that unites parts of the world—both spatial and non-spatial—is integral in model-based RL.

In summary, the current data add to a growing body of work indicating that multiple memory systems underlie value-based decision making. Our results suggest that a common computation supports both generalization in acquired equivalence and model-based RL. We surmise that the hippocampus may be a common link supporting these heterogeneous behavioral phenomena. The emerging correspondence between memory and decision systems has important consequences for our understanding of the latter in particular, since the hippocampal relational memory system provides a concrete and relatively well characterized foundation for understanding the otherwise rather mysterious mechanisms for model-based decision making.

Acknowledgment

The authors are supported by NINDS Grant R01NS078784-01.

References

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

Addis, D. R., Wong, A. T., & Schacter, D. L. (2007). Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. *Neuropsychologia, 45*, 1363–1377.

Bornstein, A. M., & Daw, N. (2013). Cortical and hippocampal correlates of deliberation during model-based decisions for rewards in humans. *PLoS Computational Biology., 9*(12).

Bornstein, A. M., & Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience, 35*, 1011–1023.

Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature, 379*, 255–257.

Burgess, N., Maguire, E. A., & O'Keefe, J. (2002). The human hippocampus and spatial and episodic memory. *Neuron, 35*, 625–641.

Camerer, Ho (1998). Experience-weighted attraction learning in coordination games: Probability rules, heterogeneity, and time-variation. *Journal of Mathematical Psychology, 42*, 305–326.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system*. MIT Press.

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science, 210*, 207–210.

Corbit, L. H., & Balleine, B. W. (2000). The role of the hippocampus in instrumental conditioning. *The Journal of Neuroscience, 20*, 4233–4239.

Corbit, L. H., Ostlund, S. B., & Balleine, B. W. (2002). Sensitivity to instrumental contingency degradation is mediated by the entorhinal cortex and its efferents via the dorsal hippocampus. *The Journal of Neuroscience, 22*, 10976–10984.

Corkin, S. (1968). Acquisition of motor skill after bilateral medial temporal-lobe excision. *Neuropsychologia, 6*, 225–264.

Davis, H. (1992). Transitive inference in rats (rattus norvegicus). *Journal of Comparative Psychology, 106*, 342–349.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. R. Robbins (Eds.), *Decision making, affect, and learning: Attention and performance XXIII* (pp. 3–38). Oxford University Press.

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

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

Daw, N., & Shohamy, D. (2008). The cognitive neuroscience of motivation and learning. *Social Cognition, 26*, 593–620.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation, 5*, 613–624.

Den Ouden, H., Daw, N. D., Fernandez, G., Elshout, J., Rijpkema, M., Hoogman, M., Franke, B., & Cools, R. (2013). Dissociable effects of dopamine and serotonin on reversal learning. *Neuron, 80*(4), 1090–1100.

Dickinson, A. (1980). *Contemporary animal learning theory*. Cambridge University Press.

Dickinson, A., Smith, J., & Mirenowicz, J. (2000). Dissociation of pavlovian and instrumental incentive learning under dopamine antagonists. *Behavioral Neuroscience, 114*, 468–483.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology, 22*, 1075–1081.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual review of Neuroscience, 30*, 123–152.

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

Gabrieli, J. D. (1998). Cognitive neuroscience of human memory. *Annual Review of Psychology, 49*, 87–115.

Gabrieli, J. D., Corkin, S., Mickel, S. F., & Growdon, J. H. (1993). Intact acquisition and long-term retention of mirror-tracing skill in Alzheimer's disease and in global amnesia. *Behavioral Neuroscience, 107*, 899–910.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–511.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2012). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*.

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

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Science, 108*, 15647–15654.

Greene, A. J., Gross, W. L., Elsinger, C. L., & Rao, S. M. (2006). An FMRI analysis of the human hippocampus: Inference, context, and task awareness. *Journal of Cognitive Neuroscience, 18*, 1156–1173.
Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Science USA, 104*, 1726–1731.
Heindel, W. C., Salmon, D. P., Shults, C. W., Walicke, P. A., & Butters, N. (1989). Neuropsychological evidence for multiple implicit memory systems: A comparison of Alzheimer's, Huntington's, and Parkinson's disease patients. *The Journal of Neuroscience, 9*, 582–587.
Hikida, T., Kimura, K., Wada, N., Funabiki, K., & Nakanishi, S. (2010). Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron, 66*, 896–907.
Honey, R. C., & Hall, G. (1989). Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behaviour Processes, 15*, 338–346.
Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *The Journal of Neuroscience, 27*, 12176–12189.
Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science, 273*, 1399–1402.
Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory, 1*, 106–120.
Kruschke, J. K. (2010). *Doing Bayesian data analysis, a tutorial introduction with R*. Academic Press.
Lee, H., Ghim, J. W., Kim, H., Lee, D., & Jung, M. (2012). Hippocampal neural correlates for values of experienced events. *The Journal of Neuroscience, 32*, 15053–15065.
Martone, M., Butters, N., Payne, M., Becker, J. T., & Sax, D. S. (1984). Dissociations between skill learning and verbal recognition in amnesia and dementia. *Archives of Neurology, 41*, 965–970.
Meeter, M., Shohamy, D., & Myers, C. E. (2009). Acquired equivalence changes stimulus representations. *Journal of the Experimental Analysis of Behavior, 91*, 127–141.
Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience, 16*, 1936–1947.
Moore, A. W., & Atkeson, C. G. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning, 13*, 103–130.
Myers, C. E., Shohamy, D., Gluck, M. A., Grossman, S., Kluger, A., Ferris, S., et al. (2003). Dissociating hippocampal versus basal ganglia contributions to learning and transfer. *Journal of Cognitive Neuroscience, 15*, 185–193.
Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology, 19*, 1–32.
O'Keefe, J., & Nadel, L. (1979). The hippocampus as a cognitive map. *Behavioral and Brain Sciences, 2*, 487–533.
Otto, A. R., Gershman, S. J., 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., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Science USA, 110*, 20941–20946.
Pfeiffer, B. E., & Foster, D. J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. *Nature, 497*, 74–79.
Preston, A. R., Shrager, Y., Dudukovic, N. M., & Gabrieli, J. D. E. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus, 14*, 148–152.
Reynolds, J. N. J., & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks, 15*, 507–521.
Rummery, G., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems. CUED/F-INFENG/TR 166*. Cambridge University Engineering Department.
Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. *Current Biology, 22*, 1622–1627.
Shohamy, D., Myers, C. E., Grossman, S., Sage, J., Gluck, M. A., & Poldrack, R. A. (2004). Cortico-striatal contributions to feedback-based learning: Converging data from neuroimaging and neuropsychology. *Brain, 127*, 851–859.
Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal–midbrain encoding of overlapping events. *Neuron, 60*, 378–389.
Simon, D. A., & Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *The Journal of Neuroscience, 31*, 5526–5539.
Skatova, A., Chan, P. A., & Daw, N. D. (2013). Extraversion differentiates between model-based and model-free strategies in a reinforcement learning task. *Frontiers in Human Neuroscience, 7*, 525.
Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99*, 195–231.
Stan Development Team. Stan: A C++ Library for Probability and Sampling, Version 1.3.
Staresina, B. P., & Davachi, L. (2009). Mind the gap: Binding experiences across space and time in the human hippocampus. *Neuron, 63*, 267–276.
Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience, 16*, 966–973.
Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning, GTE Laboratories Incorporated* (pp. 216–224). Morgan Kaufmann.
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience, 7*, 887–893.
Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*, 189–208.
Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology, 4*, 257–275.
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.
Tsai, H. C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., de Lecea, L., et al. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science, 324*, 1080–1084.
Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2009). Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. *Journal of Cognitive Neuroscience, 21*, 1934–1945.
van der Meer, M. A. A., Johnson, A., Schmitzer-Torbert, N. C., & Redish, A. D. (2010). Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron, 67*, 25–32.
Wimmer, G. E., Daw, N. D., & Shohamy, D. (2012). Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience, 35*, 1092–1104.
Wimmer, G. E., & Shohamy, D. (2012). Preference by association: How memory mechanisms in the hippocampus bias decisions. *Science, 338*, 270–273.
Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron, 75*, 418–424.
Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience, 19*, 181–189.
Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience, 22*, 513–523.
Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience, 6*, 70.