Episodic Contributions to Model-Based Reinforcement Learning

Oliver Vikbladh (omv208@nyu.edu)
Center for Neural Science, NYU, 4 Washington Place
New York, NY, 10003, USA

Daphna Shohamy (ds2619@columbia.edu)
Dept of Psych and MBBI, Columbia, 1190 Amsterdam Ave
New York, NY, 10027, USA

Nathaniel Daw (ndaw@princeton.edu)
PNI and Dept of Psych, Princeton University
Princeton, NJ, 08540, USA

Abstract

Much research indicates that organisms can plan actions using maps or models of the environment, and that such model-based (MB) learning trades off against simpler model-free (MF) mechanisms. However, standard models of both use incremental learning rules that extract statistical summaries from experience. Another possibility is that individual events are stored as episodic memories and later sampled to guide choice. Such episodic evaluation may confound standard tests for model use, since individual trajectories contain the same information as the map.

We examined the contribution of episodic memory to MB vs MF learning using a a task that combines 2-step MDP dynamics with trial-unique memory cues that also predict reward. The task reveals whether episodic information influences choices via MB or MF evaluation, and also whether these effects trade off against what has previously been interpreted as incrementally learned estimates.

Subjects displayed standard, putatively incremental MF and MB strategies, but also strong MB planning using individually cued episodes. Furthermore, on trials that contained episodic cues, incremental MB planning was reduced. This tradeoff suggests that previous interpretations of choices reflecting running averages may reflect covert retrieval of episodes, which are replaced by cued episodes when these are provided.

Keywords: episodic memory; sampling; model-based; reinforcement-learning; third way

Introduction

In addition to maintaining incrementally learned averages of past events, which are thought to guide choice, organisms can store individual events (e.g. trials) as separate episodic memories. These episodic memories may also guide later decisions (Lengyel and Dayan 2007; Gershman & Daw, 2016) and could contribute covertly even in tasks without explicit single-trial cuing (Bornstein et al in press). Values computed directly from such traces could share many of the advantages of computation from a cognitive map or world model, and indeed might confound tests that purport to show that organisms plan actions using a map or model of the environment.

We created a novel task that combines two-step MDP dynamics of the sort previously used to distinguish MB from MF learning (Daw et al, 2011), with single trial memory cues (unique objects) that also predict reward (Duncan et al 2016). If seemingly incremental RL strategies – specifically, we hypothesized, MB strategies – indeed reflect covert episodic retrieval, we expected these to be attenuated as we redirected a potential episodic controller by cuing subjects to retrieve a specific episode.

Methods

80 subjects were recruited via Amazon's Mechanical Turk**.** On each of 200 trials, subjects chose between two fractals (Figure 1). Each fractal tended to lead to one of two second-level states (with 70% probability). The second-level states were defined by categories (objects vs. animals), such that upon visiting a state, an exemplar from that category was displayed, followed by a reward of 0-8 units. The rewards were determined by a binary latent variable at each state (high or low, corresponding to two different distributions (Figure 2). On each trial, there was a 10% chance that each state's reward distribution flipping between low and high.

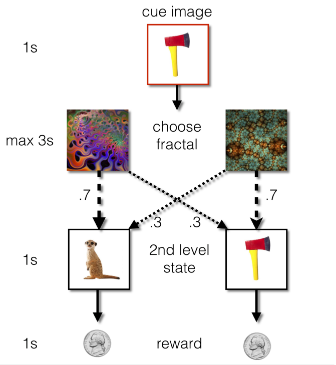

Figure 1: Trial Structure

On 2/3 of trials, before subjects made their initial choice between fractals, a cue image of an object or an animal was display. Previously encountered objects were repeated as cues at most once, on roughly half of trials with cues and 13-47 trials after their initial occurrence. The cue signified two promises: first, if, on that trial, the subject encountered the second-level state corresponding to its category, this would be the specific object encountered there. Second, if the subject had previously encountered that object, it would be paired with the same reward amount as before, notwithstanding the current reward distribution (high or low) of its category. Repeated cues were scheduled to comply with this episodic promise, but also respect the (orthogonal) average state values of the categories.

The embedding of choices within a two-step MDP allowed us (as previously; Daw et al., 2011) to examine MB or MF valuation – i.e. whether fractals' action values were tied to their own reward history vs derived indirectly from the associated categories' reward histories (computed via a one-step model). We refer to these strategies as "incremental" MB and MF.

Episodic cues in the current task permit two new strategies, notably an "episodic MB" strategy, where on trials with a repeated cue, one would retrieve the reward associated with the associated cue, assign that value to the associated category, and evaluate the fractals via a one-step model as above. In principle, an "episodic MF" strategy is also available (reminiscent of the Lengyel and Dayan, 2007, episodic controller), which retrieves the entire trajectory associated with the cue, and treats it as a TD(1) estimator for the value of the fractal previously chosen with it.

Results

We fit a full computational model to our data. Incremental MB and MF, and episodic MB all had significant effects on choice (non zero softmax temperatures; Ps<.001), but incremental MF did not (P=.20) (Figure 1). We also analyzed the change in reliance on incremental strategies when a repeated cue was presented vs. when it was not (Figure 3). We found a significant decrease in incremental MB behavior (p=0.02) but not incremental MF behavior (p=0.46) (figure 7), suggesting the two MB strategies trade off in some way. (We verified this result was not due to the presence of a cue per se, by further separating the effects for trials with no-cue vs a non- repated one, but these did not differ.)

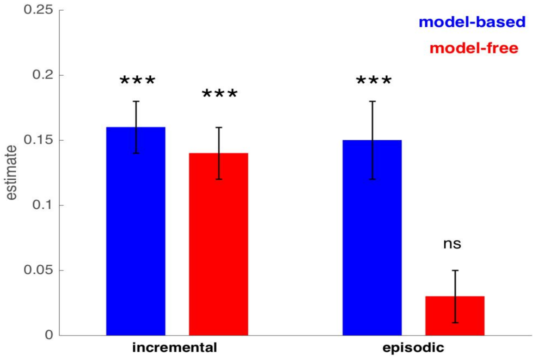

Figure 2: Incremental and episodic model-based and model-free estimates according to our model.

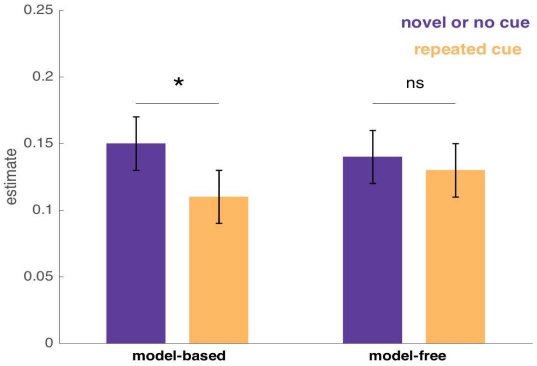

Figure 3: Incremental MB and MF estimates on trials with and without repeated cues.

Discussion

Our results indicate that in addition to using standard "incremental" MB and MF value estimates, humans can also plan in a MB fashion by retrieving memories of reward received during specific episodes in the past. The results here are of particular interest because aspects of the data are suggestive of a specific role for episodic memories in the MB, rather than the MF value estimates. For instance, we found that on trials when subjects were cued with an object from the past, they exhibited less "incremental" MB planning. This trade-off may reflect that these strategies share something in common. One possibility is that, even "incremental" MB learning over recent trials actually reflects covert retrieval of episodes with a recency bias, previously mistaken for incremental learning. Such retrieval then would be redirected toward specific trials in our cued condition, producing the tradeoff.

References

Bornstein AM, Khaw MW, Shohamy D, Daw ND (in press). What's past is present: Reminders of past choices bias decisions for reward in humans. *Nat. Comm.*

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*(6), 1204-1215.

Duncan K, Gerraty RT, Doll, BB, Daw ND, Shohamy D 2016 Disentangling the contributions of episodic memory and incremental learning to value-based decisions. Abstract *Society for Neuroscience*

Gershman SJ and Daw ND 2017. Reinforcement learning and episodic memory in humans and animals. *Annual Review of Psychology* 68 101-128

Lengyel M,Dayan P 2007. Hippocampal Contributions to Control: The Third Way." *NIPS*. Vol. 20.