Title

Episodic Control as Meta-Reinforcement Learning

Permalink

https://escholarship.org/uc/item/7zj8r415

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 40(0)

Authors

Ritter, S
Wang, JX
Kurth-Nelson, Z
et al.

Publication Date

2018

Episodic Control as Meta-Reinforcement Learning

S Ritter[1,2*], **JX Wang**[1*],
Z Kurth-Nelson[1,3], **M Botvinick**[1,4]

[1]DeepMind, London, UK
[2]Princeton Neuroscience Institute, Princeton, NJ
[3]MPS-UCL Centre for Computational Psychiatry, London, UK
[4]Gatsby Computational Neuroscience Unit, UCL, London, UK
{ritters, wangjane, zebk, botvinick} @google.com

Abstract

Recent research has placed episodic reinforcement learning (RL) alongside model-free and model-based RL on the list of processes centrally involved in human reward-based learning. In the present work, we extend the unified account of model-free and model-based RL developed by Wang et al. (2017) to further integrate episodic learning. In this account, a generic model-free "meta-learner" learns to deploy and coordinate all of these RL algorithms. The meta-learner is trained on a broad set of novel tasks with limited exposure to each task, such that it *learns to learn* about new tasks. We show that when equipped with an episodic memory system inspired by theories of reinstatement and gating, the meta-learner learns to use the same pattern of episodic, model-free, and model-based RL observed in humans in a task designed to dissociate among the influences of these learning algorithms. We discuss implications and predictions of the model.

Keywords: Reinforcement learning; model-based; deep learning; meta-learning; episodic memory

Introduction

Nearly every decision an intelligent organism makes is informed by its memory of the results of its past decisions. To be successful, agents must distill the results of past decisions into memories, then make use of those memories to make better decisions in the future. Accordingly, much effort has been directed toward understanding 1) what humans and animals store after a sequence of actions and rewards, and 2) how they use that stored information to appraise the value of future actions.

Model-free and model-based reinforcement learning (RL; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Sutton & Barto, 1998) offer distinct solutions to these two problems. Model-free RL stores statistics about the relationship between states, actions and rewards, and appraises actions by calculating how frequently they led to reward. Meanwhile, model-based RL stores estimated state-state transition probabilities, and appraises actions by using this model to simulate sequences of states to predict future reward. Signatures of both model-free and model-based learning appear in behavior and in the brain (e.g. Daw et al., 2011), and a venerable tradition holds that they are implemented by dissociable neural systems (for review see Dolan & Dayan, 2013).

However, the recent theory of meta-reinforcement learning (meta-RL) proposed that model-free learning, model-based learning, and their sometimes complex interaction could all be explained by a simple unified mechanism (Wang, Kurth-Nelson, Tirumala, Soyer, et al., 2017; Wang, Kurth-Nelson, Tirumala, Leibo, et al., 2017). In meta-RL, the weights of a recurrent neural network (RNN) are trained by model-free learning on a series of interrelated tasks, and given the reward signal as part of its input. Remarkably, this leads to the emergence of an independent RL algorithm implemented in the activation dynamics. Through its recurrence, the network has access to the history of observations, actions, and rewards. It learns to distill this history into its activation dynamics (a form of *working memory*) and use this to select rewarding actions. The end result is a learned reinforcement learning algorithm that operates even with the weights of the RNN frozen. This meta-learned algorithm can itself be model-based even though it was acquired through model-free learning (Wang, Kurth-Nelson, Tirumala, Soyer, et al., 2017).

While meta-RL provides a full account of incremental learning as it is carried out in working memory, it does not account for the episodic learning processes to which attention has recently been called (Gershman & Daw, 2017). In addition to learning by incrementally storing recent sequences of behavior in working memory, humans appear to learn by storing summaries of individual episodes for long periods of time, then retrieving them when similar contexts are encountered. For example, cues triggering episodic memory retrieval impact reward-based learning, both for good and for ill (Bornstein, Khaw, Shohamy, & Daw, 2017; Wimmer, Braun, Daw, & Shohamy, 2014), and distinctive aspects of episodic memory function contribute to decision-making behavior (Vikbladh, Shohamy, & Daw, 2017; Bornstein & Norman, 2017). Such observations, along with some fundamental computational insights, have recently landed episodic learning a spot beside incremental model-free and model-based reinforcement learning on the list of processes centrally involved in decision making (RL; Gershman & Daw, 2017).

In the present work, we develop a natural extension to meta-RL that enables it to integrate episodic learning. The resulting theory explains how incremental and episodic learning, as well as the coordination between them can be meta-learned through purely model-free RL. The episodic meta-RL theory proposes the following:

1. Meta-RL's working memory is supplemented by a non-

[*] These authors contributed equally to this work.

parametric long-term memory which itself stores working memory states.

2. Each state is paired with a perceptual context embedding that is later used to retrieve the working memory state when similar perceptual contexts are encountered.
3. The retrieved states are then gated into the working memory using a parameterized function, whose parameters are optimized toward the same model-free objective that trains working memory.

This proposal is inspired in part by evidence that episodic memory retrieval in humans operates through reinstatement, triggering patterns of neural activity related to those that were induced by the original encoding of the relevant episode (see, e.g., Xiao et al., 2017), and evidence that reinstatement occurs not only in perceptual systems, but also recreates patterns of activity in neural circuits supporting working memory (see Hoskin, Bornstein, Norman, & Cohen, 2017; Cohen & O'Reilly, 1996). Our implementation of this proposal draws additional inspiration from recent work on differentiable memory systems (e.g., Graves et al., 2016), especially that of Pritzel et al. (2017), which makes use of context-based retrieval for RL.

To empirically test this model, in this work we compare its behavior to that of humans observed by Vikbladh et al. (2017) in a task designed to dissociate the effects of multiple types of incremental and episodic learning. Vikbladh and colleagues found evidence of the use of a model-based form of episodic memory, whereby traces of specific episodes are retrieved from long-term memory based on visual similarity, then used along with knowledge of the transition structure of the environment to select actions. This episodic model-based learning was present in conjunction with incremental model-free and incremental model-based learning. In the following sections, we describe the task in detail and demonstrate that meta-RL with episodic memory replicates the qualitative pattern of human behavioral results observed by Vikbladh et al. To conclude, we consider directions for future work, including testable predictions of the theory.

Task

The task we study is a version of the two-step task (Daw et al., 2011) augmented with episodic cues to previous trials. The task structure, which was inspired by Vikbladh et al. (2017), is diagrammed in Figure 1. Each trial consisted of two stages. On the first stage, state S_0, the agent was either presented with the "no-cue" stimulus (a vector of all -1's) if this was an uncued trial, or with a binary vector associated with a previously seen second-stage context if this was a cued trial (see Figure 1). In response, the agent chose either a_1 or a_2 and transitioned into one of two second-stage states S_1 or S_2 with probabilities $P(S_1|a_1) = P(S_2|a_2) = 0.9$ (common transition) and $P(S_1|a_2) = P(S_2|a_1) = 0.1$ (uncommon transition). On the second stage, the agent was presented with a stimulus representing the context of that second-stage state, followed by a final step in which it was shown the reward outcome. The context thus represented the conjunction of having transitioned into that particular state and obtaining that particular reward.

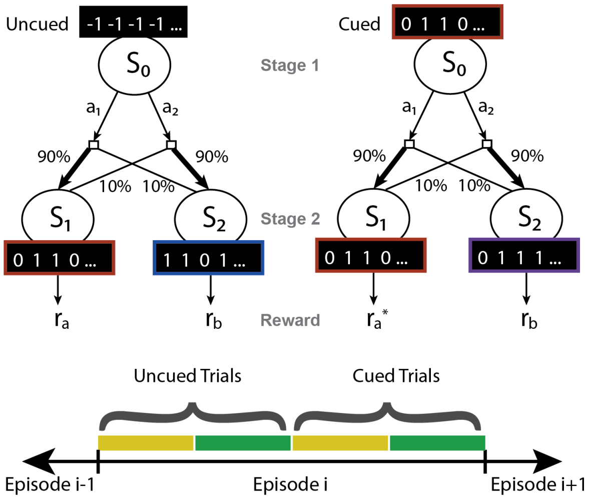


Figure 1: Contextual two-step task modeled after Daw et al. (2011) and Vikbladh et al. (2017). (top) Two trial types are shown: uncued and cued. All trials start in state S_0 at the first stage, at which point agents are presented with either a "no-cue" stimulus or are cued with a second-stage stimulus seen on a previous trial. Transition probabilities after taking actions a_1 or a_2 are depicted in the graph. On uncued trials, S_1 and S_2 result in Bernoulli rewards with probabilities r_a and r_b. On cued trials, transitioning into the same state as the trial being cued results in the exact sampled reward as before, r_a^*. (bottom) Trials within an episode are split into 4 blocks, with block 3 consisting of cued trials which are cued with stimuli from block 1, and block 4 cued from block 2.

The transition probabilities P were fixed across episodes. On uncued trials, the states S_1 and S_2 yielded Bernoulli probabilistic rewards of 0 or 1 according to $[r_a, r_b] = [0.9, 0.1]$ or $[0.1, 0.9]$, with the specific reward contingencies having a 10% chance of randomly switching at the beginning of each trial. On cued trials, if the agent transitioned into the same state as on the trial being cued (i.e. the context is the same), the agent was given the exact same reward as before. If the agent transitioned into the other state, the reward was determined as on uncued trials.

The first half of every episode (50 trials) consisted of all uncued trials, and the second half consisted of only cued trials, with trials 51-75 being cued with stimuli from trials 1-25 and trials 76-100 cued with stimuli from trials 26-50, randomly sampled without replacement. This was done to reduce autocorrelation in the reward probabilities by enforcing a minimum of 25 trials between seeing the stimulus on the second stage and being cued with it on the first stage.

The agent was trained for 10,000 episodes of 100 trials each, and evaluated with weights fixed on 500 further episodes.

Learning Algorithms

The two-step task with episodic cueing is designed to dissociate among the influences of four different learning strategies on choice. First, the incremental model-free strategy prescribes taking the same action that was taken on the last trial if it was rewarded, and taking the opposite action if it was not

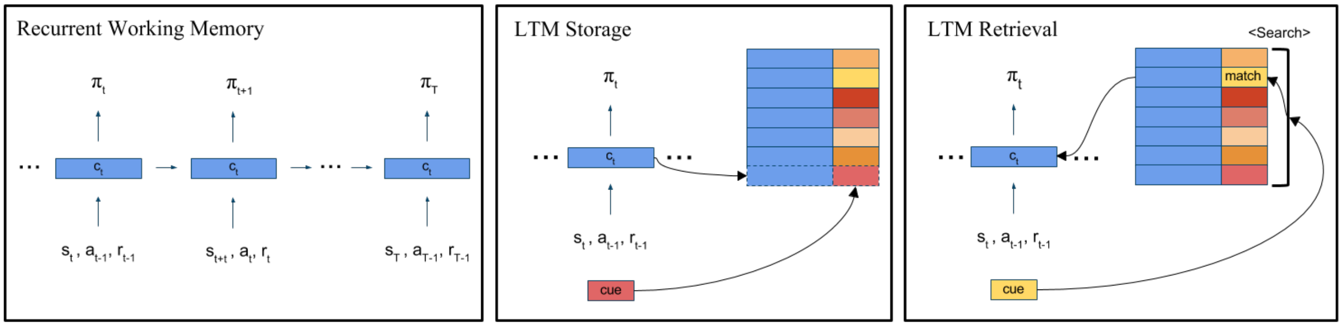


Figure 2: (Left) A high-level schematic of the recurrent network (LSTM) that comprises the episodic meta-RL (EMRL) agent's working memory. On each time step the LSTM receives an environment state, the action taken on the previous trial, and the reward received on the previous trial. The LSTM encodes this information incrementally into its cell state c, and then outputs a policy and value estimate (not shown). (Middle) The storage operation to long-term memory at a single LSTM time step. Storage is triggered when reward is received at the end of each two-step trial, at which point the agent appends the contextual cue along with its cell state to a non-parametric store of such items. (Right) The long-term memory retrieval operation which occurs on every time step. A search is carried out over the cues stored in long-term memory for the closest match to the current contextual cue. The working memory activations associated with the closest match are retrieved and reinstated to the working memory state.

rewarded, regardless of whether the transition on the previous trial was common or uncommon. In contrast, the incremental model-based strategy prescribes staying only if the previous trial was rewarded *and* the the previous transition was common. If the previous transition was uncommon and the trial was rewarded, the agent will take the opposite action. Episodic model-free and model-based strategies operate like their incremental counterparts, but with respect to the trial associated with the cue rather than the immediately previous trial.

Model

In the episodic meta-RL model, vectors represent working memory states, and the function that updates these states and selects actions based on them takes the form of a recurrent neural network, specifically a long short-term memory network (LSTM; Hochreiter & Schmidhuber, 1997). To implement context-based reinstatement of the activations in this working memory, we append to this architecture a long-term memory of previous working memory states, searchable by a second column containing representations of perceptual context. This could for instance be the agent's visual representation of the current environment state or an externally provided stimulus. Optimization was performed by an implementation of asynchronous actor-critic (Mnih et al., 2016).

The episodic meta-RL (EMRL) agent writes its current working memory state and perceptual representation to the long-term memory array when appropriate. In our experiments with the two-step task, the agent writes when it receives reward at the end of the each trial.

The agent reads from this long-term memory array on every time step by searching for the perceptual representation in the array with the smallest cosine distances to that of the current state, allowing it to retrieve the associated working memory state. The retrieved activations are next passed through a learned gating function that arbitrates among the influences on the current working memory state of 1.) current perceptual inputs 2.) the previous working memory state and 3.) the working memory state retrieved from long-term memory.

This gating mechanism is a natural extension to the standard LSTM working memory, which uses gates to arbitrate between current inputs and previous memory state:

$$c_t = i \circ c_{in} + f \circ c_{prev}$$

c_t is the current working memory state, c_{in} is the agent's representation of its current input, c_{prev} is the working memory activation from the previous timestep, and $\circ$ signifies element-wise multiplication.

The gates i and f are values between zero and one that allow (or disallow) inputs and and past working memory activations into the current state. These gates are computed accordingly:

$$i = \sigma(W_{xi}x + W_{hi}h + b_i)$$

$$f = \sigma(W_{xf}x + W_{hf}h + b_f),$$

where x is the perceptual input, h is a function of the previous working memory state, and the weight matrices W and bias vectors b contain learned parameters.

To reinstate working memory activations retrieved from long-term memory without losing the current contents of working memory, our architecture adds the retrieved activations to the current working memory state, after passing them through a gate that is computed in a manner exactly analogous to the standard LSTM gates:

$$c_t = i \circ c_{in} + f \circ c_{prev} + r \circ c_{ltm}$$

$$r = \sigma(W_{xr}x + W_{hr}h + b_r)$$

c_{ltm} contains the retrieved activations from long-term memory. This reinstatement gate r is intended to learn to allow activations from long-term memory into working memory when they are useful, but not when they will interfere with the maintenance of important information in working memory. See Figure 2 for depictions of the architecture. To illustrate how this architecture works in practice, consider the episodic two-step task, wherein the working memory keeps track of the reward probabilities at each outcome state. In order to infer these quantities, it must maintain information about its past actions and states. When the agent receives reward in the final step of a two-step trial, it will save the activations of its current working memory - which encode the agent's outcome state and reward - to its long-term memory. These are saved along with a representation of the context stimulus, which models the participant's visual representation of the object images or fractals in the human experiments (Daw et al., 2011; Vikbladh et al., 2017).

At the beginning of a future two-step trial, the agent will encounter this same stimulus. It will then search its long-term memory for matches for that stimulus, and will retrieve the hidden state from the past trial. Crucially, this hidden state will encode the state the agent encountered at the end of the last exposure to that stimulus as well as the reward received and the action taken. Possessing this crucial episodic information, the agent is able to exploit the structure of the episodic two-step task. Specifically, it can learn to implement model-based or model-free valuation with respect to trial information retrieved from long-term memory.

Results

After training, we assessed meta-RL's performance on a set of evaluation episodes having the same structure as the training episodes. However, to isolate the behavior of the *learned* learning algorithm operating in the activation dynamics, all data shown in the Results were obtained with the network's weights frozen.

The learned algorithm obtained more reward on cued than uncued trials (Figure 3a; $p < 10^{-10}$ by Fisher exact test), suggesting it could make use of the information carried by the episodic cue. We also compared against a control agent (meta-RL; MRL) that was trained and tested in exactly the same way, but did not have access to an episodic memory (r-gate was always fixed to zero). The agent with episodic memory (EMRL) performed significantly better on cued trials than MRL (Figure 3a; $p < 10^{-16}$ by Fisher exact test). For comparison, random behavior in this task yields 0.5 reward per trial, while optimal behavior achieves 0.756 on average.

Next, we asked whether, on uncued trials, EMRL exhibited the canonical pattern of model-based behavior first described in (Daw et al., 2011). In the two-step task, if action a_t was taken on timestep t, followed by a common transition and resulting in a high reward, then both model-based and model-free learners are expected to increase their preference for a_t. However, if after action a_t, an *uncommon* transition was observed, followed by high reward, a model-free learner will still increase its preference for a_t (since it was rewarded after taking this action), while a model-based learner will decrease its preference for a_t (by using its knowledge of the transition structure of the task to infer a higher value for the other action). We formally tested for these patterns of behavior by performing an ANOVA on the probabilities of repeating the previous action, with two binary factors: whether the previous trial was rewarded, and whether the previous trial had a common transition. A main effect of previous trial being rewarded would indicate a model-free strategy, while an interaction between previous trial being rewarded and previous trial being common would indicate a model-based strategy.

On uncued trials (Figure 3b), we found a strong effect of the interaction term ($F(1,1853) = 9134$, $p \approx 0$), indicating that the learned algorithm correctly exploited the transition structure of the task when no episodic information was available, replicating our previous work (Wang, Kurth-Nelson, Tirumala, Soyer, et al., 2017). On cued trials (Figure 3c), we also found an effect of the interaction term ($F(1,1893) = 295$, $p \approx 0$), suggesting that EMRL partially attempted to continue to use the incremental strategy even though it had no reward benefit on cued trials.

Next, most centrally, we asked whether – on cued trials – EMRL could apply model-based reasoning to information retrieved from episodic memory (Figure 3d). We performed the same ANOVA described above, but using as factors: whether the past trial was rewarded, and whether the past trial had a common transition. Since our task guaranteed receiving the same reward if the agent reached the same state as the past trial, the agent should prefer to take the opposite action as on the past trial if it experienced an uncommon transition and received reward on that trial. We indeed found a strong effect of the interaction term in this analysis ($F(1,1850) = 5975$, $p \approx 0$). This pattern mimics the behavior of humans on the task from which ours was directly inspired (Vikbladh et al., 2017). Note that we only performed this analysis on cued trials because the factors would be undefined on uncued trials.

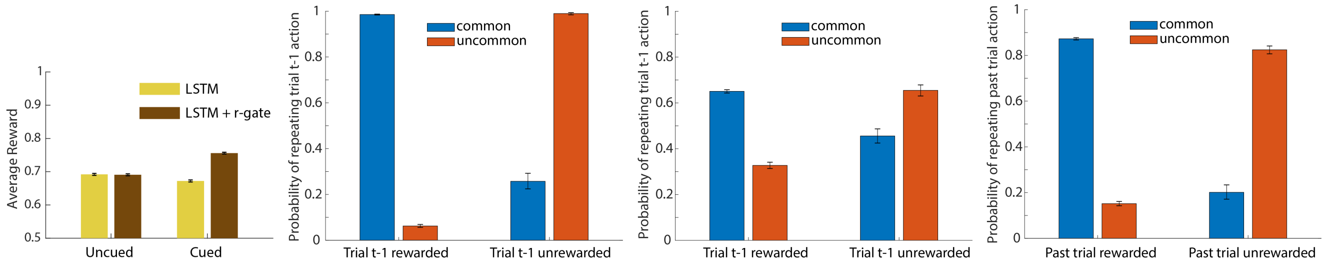
To supplement this analysis, we also fit a probabilistic choice model to EMRL's behavior. In this model, action probability was the softmax of the weighted sum of four choice values:

$$P(a_0) = \frac{1}{1 + \exp(-(\beta_{if} V_{if} + \beta_{ib} V_{ib} + \beta_{ef} V_{ef} + \beta_{eb} V_{eb}))}$$

where, for example, V_{if} is the difference in incremental model-free values: action a_0 minus action a_1. All incrementally learned values were updated with the same learning rate α, for a total of five parameters. These were estimated by maximum likelihood on the concatenated data of all 500 episodes (with incrementally learned values reset to 0.5 at the beginning of each episode). Cued and uncued trials were fit separately. Mirroring the human data in Vikbladh et al. (2017), we found a contribution of all systems except episodic model-free, and a reduction in the incremental model-based parameter on cued trials (Figure 4).

Analysis of Reinstatement Gate Activations

We performed a preliminary analysis of the activations of the reinstatement gates (see Figure 5). First, we plotted the time-course of mean r-gate values averaged over 500 episodes. We split these time courses based on the stage in the two-step trial



(a) Performance

(b) Incremental Uncued

(c) Incremental Cued

(d) Episodic

Figure 3: EMRL exhibits both incremental and episodic model-based behavior. (a) Average reward obtained by MRL and EMRL on cued and uncued trials. EMRL, but not MRL, makes use of episodic memory on cued trials to earn more reward. (b) Proportion of uncued trials in which EMRL repeated the action it took on the previous trial ($t-1$), split by whether it received reward on $t-1$ and whether the transition on $t-1$ was common. The interaction between those two factors is a sign of model-based learning, as in Daw et al. (2011). (c) Same as b, but for cued trials. (d) Same as b, but split by whether EMRL received reward on trial k and whether the transition on k was common. k refers to the past trial in which the cue was first encountered.

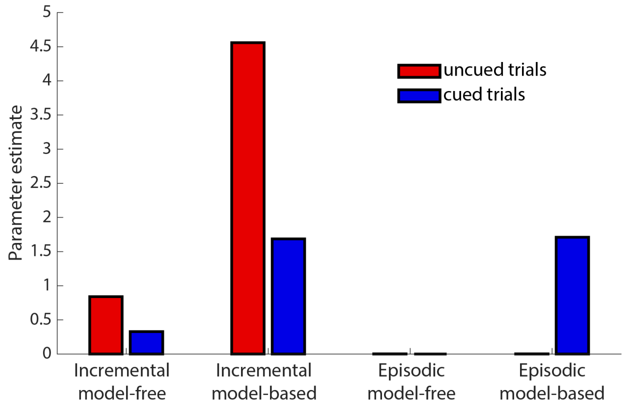
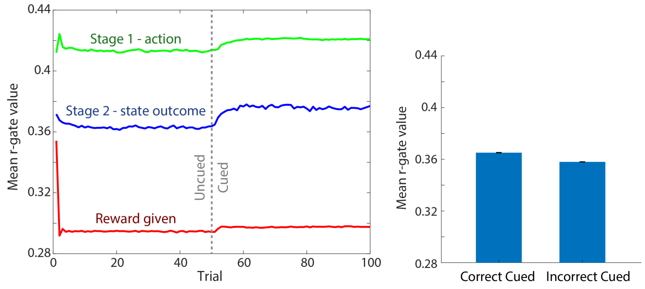


Figure 4: Parameter estimates in a model with weighted contributions of four decision systems. Model was fit to EMRL's actions separately on cued and uncued trials. All systems except episodic model-free have positive contributions to EMRL's behavior. The contribution of the incremental model-based system is reduced on cued trials relative to uncued trials.



(a) r-gate throughout an episode

(b) r-gate correct vs incorrect

Figure 5: (a) Time course of the mean values of the reinstatement gate averaged over 500 episodes, split up by stage of the trial. (b) Mean values of the reinstatement gate on cued trials on which the agent selected the optimal and on cued trials on which it selected the opposite action, averaged over all units.

(see Figure 1). We first observed that the r-gate was more open during cued trials compared to uncued trials, consistent with the presence of useful episodic information on cued trials. Next, we observed that the r-gate was most open during the first stage of each trial. We believe this makes sense because it is critical to base action on the information retrieved from long-term memory.

Next, we compared the mean r-gate values on correct cued trials versus cued trials where the agent made an error, and found the r-gate was significantly more open on correct trials. We speculate that fluctuations in r-gate openness resulting from the multiplexing of episodic memory control with current policy control might have driven some behavioral errors. Future work should analyze this phenomenon in detail.

Discussion

The experiments in this work establish that when trained on a task distribution with both incremental and episodic reward structure, episodic meta-RL learns to simultaneously execute incremental model-based, incremental model-free, and episodic model-based learning strategies. Like the participants in the study by Vikbladh and colleagues, our agent exhibits these three learning strategies, but does not exhibit episodic model-free learning. This striking match in behavior provides support for episodic meta-RL as a model of human decision making as it is implemented by working memory and episodic memory structures.

Episodic meta-RL thus provides an empirically supported unified account of incremental and episodic learning processes, whereby a single model-free learning mechanism learns to execute and deploy a variety of learning algorithms observed in humans. In addition to this support from behavioral data, the model accords in principle with a large neuroscientific literature which supports the notion that episodic memory retrieval recreates patterns of activity in neural circuits supporting working memory (see Cohen & O'Reilly, 1996; Lewis-Peacock & Postle, 2008; Staresina, Henson, Kriegeskorte, & Alink, 2012; Hoskin et al., 2017; Xiao et al., 2017). Further, the gating system that allows reinstated activations into working memory is formally equivalent to those in the LSTM that inspired neuroscientific theory which posits that multiple such gating mechanisms operate in prefrontal cortex (Chatham & Badre, 2015).

The key takeaway from the success so far of the episodic meta-RL model is a proof of the sufficiency of a small set of well motivated architectural components, when trained to optimize a specific objective function, to produce the complex pattern of learning processes observed in humans. The architecture components are: 1) a recurrent working memory with 2) a non-parametric store of working memory activations that can be retrieved by context and reinstated through 3) a learned gating system. The objective function is total reward achieved on a distribution of learning tasks.

Predictions

This model makes a number of predictions which may be tested through further empirical work:

- The pattern reinstatement seen during episodic RL tasks (Bornstein & Norman, 2017) should be observed in cases where the patterns in question are linked with working memory, rather than only immediate perception.
- There is already strong neurobiological evidence for gating to regulate the flow of information into and out of working memory (Chatham & Badre, 2015). Analogous experiments should find evidence for a similar mechanism to gate reinstated activations from long-term memory into working memory.
- In a novel episodic RL task, we predict that relevant cues will trigger episodic recall, but this recall initially will not trigger reinstatement in the populations of cells responsible for working memory. As task experience increases, improvements in behavioral performance will be correlated with increased functional coupling between episodic and working memory. However, if DA-dependent plasticity is blocked during training, this increase in functional coupling will be slowed.

Summary

In summary, this work presents a new model that explains the collage of learning processes observed in humans during decision making as an interplay between working and episodic memory that is itself learned through training to maximize reward on a distribution of learning tasks. Future work may test the predictions made by this model and test the model's ability to replicate additional sources of empirical data.

References

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*.

Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*.

Chatham, C. H., & Badre, D. (2015). Multiple gates on working memory. *Current Opinion in Behavioral Sciences*.

Cohen, J. D., & O'Reilly, R. C. (1996). A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. In *Prospective memory: Theory and applications.*

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.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual review of psychology*, *68*.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... others (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hoskin, A. N., Bornstein, A. M., Norman, K. A., & Cohen, J. D. (2017). Refresh my memory: Episodic memory reinstatements intrude on working memory maintenance. *bioRxiv*, *170720*.

Lewis-Peacock, J. A., & Postle, B. R. (2008). Temporary activation of long-term memory supports working memory. *Journal of Neuroscience*, *28*(35), 8765–8771.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *Proc. of int'l conf. on machine learning, ICML.*

Pritzel, A., Uria, B., Srinivasan, S., Puigdomènech, A., Vinyals, O., Hassabis, D., ... Blundell, C. (2017). Neural episodic control. *arXiv preprint arXiv:1703.01988*.

Staresina, B. P., Henson, R. N., Kriegeskorte, N., & Alink, A. (2012). Episodic reinstatement in the medial temporal lobe. *Journal of Neuroscience*, *32*(50), 18150–18156.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge.

Vikbladh, O., Shohamy, D., & Daw, N. (2017). Episodic contributions to model-based reinforcement learning. In *Annual conference on cognitive computational neuroscience, CCN.*

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Leibo, J., Soyer, H., Kumaran, D., & Botvinick, M. (2017). Meta-reinforcement learning: a bridge between prefrontal and dopaminergic function. In *Cosyne abstracts.*

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... Botvinick, M. (2017). Learning to reinforcement learn. *Cognitive Science, CogSci*. Retrieved from `arXivpreprintarXiv:1611.05763`

Wimmer, G. E., Braun, E. K., Daw, N. D., & Shohamy, D. (2014). Episodic memory encoding interferes with reward learning and decreases striatal prediction errors. *J Neurosci*, *34*.

Xiao, X., Dong, Q., Gao, J., Men, W., Poldrack, R. A., & Xue, G. (2017). Transformed neural pattern reinstatement during episodic memory retrieval. *Journal of Neuroscience*, *37*.