A Memory-Augmented Reinforcement Learning Model of Food Caching Behaviour in Birds

Johanni Brea (johanni.brea@epfl.ch)
EPFL, Switzerland

Wulfram Gerstner (wulfram.gerstner@epfl.ch)
EPFL, Switzerland

Abstract

Birds of the crow family are well known for their complex cognition. In laboratory experiments it has been observed that jays adapt food caching strategies to anticipated needs and rely on a memory of the what, where and when of previous caching events for cache recovery. While this behaviour is well studied, little is known about the algorithms and neural processes that produce this behaviour. We present a computational model and propose a neural implementation of food caching behaviour. Our model features latent hunger variables for motivational control, an associative memory for snapshots of the sensory states during caching events, a system memory consolidation for flexible decoding of the age of a memory, a stimulus-driven retrieval mechanism, and reward-modulated update of retrieval and caching policies during inspection of caches. We show that our model is in quantitative agreement with the results of 22 behavioural experiments. Our methodology of a formalization of experimental protocols via a domain-specific language is transferable to other domains and may serve as a tool to design new experiments and foster collaboration between experimentalists and theoreticians. Our model is an example of a structured reinforcement learning algorithm that could have evolved in species that operate in partially observable environments.

Keywords: reinforcement learning; associative memory; episodic-like memory; food caching behaviour

Introduction

Birds of the crow family (*corvidae*) have been proposed as animal models for human cognitive neuroscience, because of their remarkably complex cognition (Clayton & Emery, 2015). In the wild, nutcrackers and jays are famous for caching acorns, nuts or pine seeds but also perishable items like insects and scraps of meat in thousands of small cracks or in loose soil, for hours, days or months. Recovery of their own caches is highly probable ($50-99$ %), clearly dependent on visual cues, independent of olfactory cues and unlikely to be explained by random search at preferred locations (Vander Wall, 1990). In laboratory experiments, jays were found to rely on episodic-like memories to retrieve from the most promising cache sites (Clayton & Dickinson, 1998, 1999a, 1999c, 1999b; de Kort, Dickinson, & Clayton, 2005) and adapt their caching strategy to anticipated future needs: jays decrease the amount of cached food items at sites where food was abundantly available (Raby, Alexis, Dickinson, & Clayton, 2007; Correia, Dickinson, & Clayton, 2007; Cheke & Clayton, 2011) or where they experienced pilfering or degradation of the cached food items (Clayton, Dally, Gilbert, & Dickinson, 2005; de Kort, Correia, Alexis, Dickinson, & Clayton, 2007).

The interpretation of these results is controversial. On one side, they have been interpreted as evidence for 'mental time travel' in animals (Raby et al., 2007; Correia et al., 2007; Cheke & Clayton, 2011), challenging, first, the hypothesis that this ability to 're-experience' the personal past and 'pre-experience' a potential personal future is uniquely human (Suddendorf & Corballis, 2007) and, second, the Bischof-Köhler hypothesis that animals' apparently future-oriented actions are driven only by current needs (Suddendorf & Corballis, 1997). On the other side, a mnemonic-associative account has been formulated that explains this behaviour with a re-evaluation of previous actions at the time of cache recovery (Clayton et al., 2005; Dickinson, 2011).

We want to shed light on this controversy with a computational model. From a traditional computational neuroscience perspective the question is, whether the birds' food caching behaviour can or cannot be explained with standard concepts like model-free reinforcement learning with reward-modulated synaptic plasticity and Hopfield-network-like associative memories (Brea & Gerstner, 2016).

Methods

All considered experiments investigate the caching and cache recovery behaviour of a single jay. A typical experiment proceeds as follows. 1) A few hours of food deprivation raise the motivation of the bird. 2) The experimenter adds some food items and caching trays at specific positions in the bird's cage and waits for a short interval. 3) The caching trays and remaining food items are removed and counted. 4) While trays are outside the cage, the experimenter may or may not remove ('pilfer') or degrade the food items cached in some of the trays. 5) After some waiting interval, the caching trays are returned to the cage and the number of times the bird inspects the returned trays inspections are counted during a recovery interval. 6) After another waiting period, steps 2–5 are repeated with some variations.

We formalized all experimental protocols using actions ADD, REMOVE, WAIT, DEGRADE, PILFER, COUNT_ITEMS and (UN)COVER_TRAY, that operate on objects of type `CachingTray`, `InspectionObserver` and `FoodItem`. A full

experimental protocol is a function that takes as input models parametrized by θ_E and returns a summary of observed quantities $\hat{S}_E(\theta_E)$.

For each experiment E we extracted a result summary S_E that consists of at least 5 (Raby07 planning) and at most 212 (Clayton99B exp1) observed quantities, collected from figures and text of the respective publications, e.g. ANOVA tests and means and variances of the number of cached items. Since we do not have analytical expressions for the likelihood function of our model parameters, we resorted to likelihood-free methods (Gutmann & Corander, 2016), i.e. we searched by repetitive simulation of each experiment for θ_E that maximizes the probability $P(\Delta(S_E, \hat{S}_E(\theta_E)) < \varepsilon)$ of the difference $\Delta(S_E, \hat{S}_E(\theta_E))$ between experimental S_E and simulated $\hat{S}_E(\theta_E)$ results being smaller than bandwidth ε. To find approximately the maximum likelihood estimate, we used a differential evolution optimizer (Fendt, 2017).

Results

Description of the model

We propose a memory-augmented reinforcement learning model in continuous time to describe caching behaviour. The model's internal state consists of 1) hunger variables, 2) an associative memory of caching events and 3) weights that influence the caching and the retrieval preference (Fig. 1).

Action selection. As soon as food items or caching trays are available the model bird chooses between immediate eating of a food item, caching a food item in one of the available trays, inspection of a tray or doing something else. Action i is sampled with probability $p_i / \sum_j p_j$, where p_i is a hunger modulated preference of action i. Each action is followed by a random timeout interval. The preference to immediately *eat* a certain food item depends on its type. To choose what to *cache* where, the model birds compute caching preferences that depend though plastic weights on cache site features and food type. When caching trays are available, the model bird may *inspect* them. Available trays can trigger the recall of a snapshot memory, if a tray's features coincide with remembered features. The preference to inspect a given tray is high when items of the associated food type are desirable at the current state of hunger and expected to be palatable at the current age of the memory.

Dynamics of hunger variables and associative memory. We model hunger with multiple variables to capture specific satiety. Eating one type of food with a high fat concentration, may decrease the first hunger variable more than the second one, whereas another type of food with a high protein concentration, decreases the second variable more than the first one. At the moment when a food item is cached, a snapshot is taken which associates the features of the cache site (where) with the food types (what) of items cached at this position. Through system memory consolidation (Sekeres, Winocur, & Moscovitch, 2018), each snapshot moves over time to other regions in the brain, which allows a flexible readout of the age of a memory (when). Items are removed from memory when

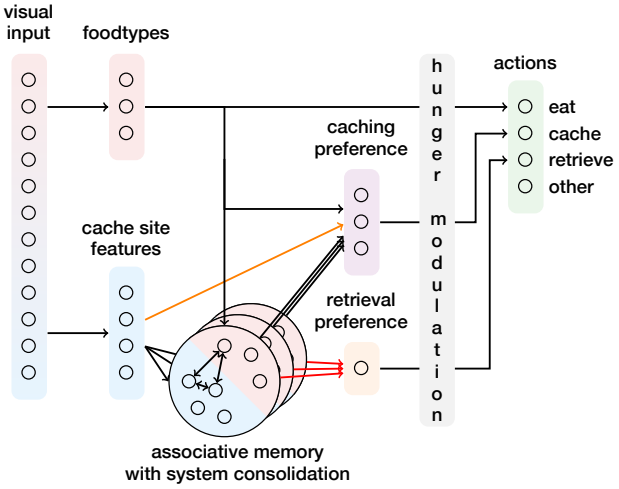

Figure 1: **The model.** Given the **visual input**, the model bird identifies the presence of food items of certain **food types** and the locations of cache sites with certain **cache site features**. Once a food item and a putative cache site is identified, the bird compares the **caching preference** with its desire to immediately consume the food item, where both preferences are modulated by the current state of **hunger**, and selects an **action** thereafter. If the bird decides to cache the food item, an association of cache site features and food types (bidirectional arrows) gets stored in an **associative memory**; the content of this memory undergoes **system consolidation**, i.e. over the course of days and weeks the memory contents are moved to other locations in the brain (depicted as multiple discs). Cache site features can trigger the associative recall of food items. Its **retrieval preference** will depend on the age of the memory that can be decoded from its current state of consolidation. After a retrieval attempt, the retrieval weights (**red arrows**) and caching preference weights (**orange arrows**) are updated in a way that depends on whether a palatable food item could be retrieved or not.

the last item in a cache gets recovered or the cache is found empty. We do not discretize time but respect in all experimental protocols the actual durations and integrate the dynamics with an event-based numerical integrator.

Reinforcement learning of weights. The update of the caching and retrieval weights follows a multi-factor Hebbian learning rule $\frac{d}{dt} w_{ij} = M \cdot e(\mathrm{post}_i, \mathrm{pre}_j)$, where e is an eligibility trace on the timescale of seconds, triggered by pre- and postsynaptic activity, and M is a modulating factor that depends on the outcome of cache inspection. Because associative recall of snapshot memories reactivates the relevant pre- and postsynaptic neurons, this standard reward modulated rule enables delayed reinforcement learning of weights that determined caching decisions potentially long ago. If inspection leads to successful recovery of a fresh and desirable food item, the preference to cache at sites with similar features increase. If the bird unexpectedly recovers a degraded

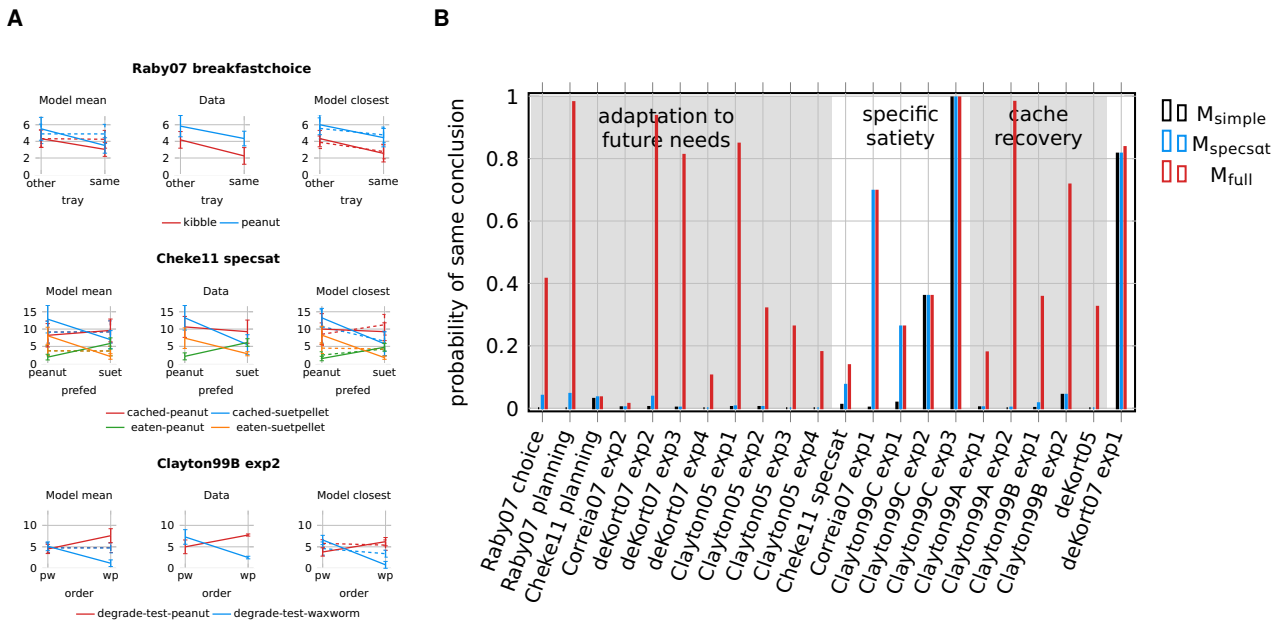

Figure 2: **The experimental results are considerably more probable under the full model with hunger dynamics and delayed reinforcement learning than under simpler models. A** Example results. Experimental data (middle column) and results of the full model are shown with solid lines, results of a simpler model with dashed lines; left column: average mean and average standard deviation over 1000 simulated repetitions of the same experiment; right column: closest simulated result to experimental data out of 1000 simulations. **B** For each experiment, the probability of reaching the same conclusion as in the experiments was computed as the fraction of simulations that had the same significance levels ($\alpha = 0.05$) on the most relevant statistical tests reported in each study. The experiment "deKort07 exp1" was a control experiment that does not require any memory or learning. The simple model reaches also similar levels as the more complex models for experiments "Clayton99C exp2/exp3" because of a between group design. For most other experiments that test specific satiety, model $\mathrm{M}_{\mathrm{specsat}}$ is on the same level as the full model and better than the simple model $\mathrm{M}_{\mathrm{simple}}$. On all other experiments the full model reaches the same conclusion as the experiments with a higher probability.

food item, the palatability of this type of food is lowered for the given age of memory and the preference to cache at sites with similar features is lowered. If no food item can be recovered, e.g. because the cache site got pilfered, but the bird has a snapshot memory for the current cache site features, the preference to cache at sites with these features is lowered. Additionally, the preference to cache a certain type of food at a certain location decreases in places where it is abundant, for example there is no need to cache pine seeds on a pine tree.

Simulation results

We compared our full model to two simpler versions. The simplest model $\mathrm{M}_{\mathrm{simple}}$ has food-type-independent, fixed preferences for immediate eating, caching, retrieval and other actions, but no hunger modulation or memory. The second model $\mathrm{M}_{\mathrm{specsat}}$ has a food-type-specific policy and hunger modulation, but no associative memory and no update of the caching preference weights (red arrow in Fig. 1). Each model was fitted to each experiment. We find that the specific satiety model $\mathrm{M}_{\mathrm{specsat}}$ suffices to reproduce with a high probability the findings of experiments that test only specific satiety, but for experiments that test cache recovery or adaptation to future needs the full model has a higher probability of reaching the same conclusions as the ones reported in the experimental papers (Fig. 2).

Discussion

We developed a memory-augmented reinforcement learning model in continuous time, where action selection and reward signals in simulated birds depend on motivational states in form of hunger variables and a list of snapshot memories that associate the location and food items of past caching events. We found that both the motivational state and the snapshot memories are necessary to reach a high probability of explaining the observed experimental data across all experiments.

To simulate the experiments, we expressed them in a domain-specific, model-independent formal language. This approach should not only simplify model comparison but also ease the communication between experimentalists and theoreticians and help designing new experiments.

The snapshot memory in our model could be implemented

with an associative neural network (Hopfield, 1982), where special care would need to be given to readout for cache retrieval and deletion of memories after emptying a cache. The reinforcement learning rule to update caching preferences could be implemented with synaptic plasticity rules, where neuromodulators signal the condition at retrieval (Gerstner, Lehmann, Liakoni, Corneil, & Brea, 2018). There is no need to maintain an explicit long synaptic eligibility trace over days for our reinforcement learning rule, since eligibility is mediated by the associative recall of snapshot memories.

The model presented here is a reinforcement learning model with structured state representation and memory to cope with partial observability. The state is factored into visual input, hunger variables and memory state. The hunger variables account for motivational effects in a similar way as e.g. proposed by (Niv, Joel, & Dayan, 2006). Read and write access to memory is specialized to the task of caching. Reinforcement learning models with more flexible forms of explicit memory (e.g. (Graves et al., 2016)) would eventually also learn to cache and retrieve efficiently, but would potentially require many more caching-retrieval events until they reach high efficiency. It is not unlikely that evolution has led to highly specialized memory systems given that the caching behaviour of juvenile birds starts already at 6 weeks of age (Stotz & Balda, 1995). In our model, the retrieval of snapshot-memories is unidirectional in that cache site features are exclusively used as keys to find the associated food items but currently desired food items are never used as keys to search for associated cache site features.

Our model belongs in the class of model-free reinforcement learning, since the simulated birds do not learn the transition structure of the environment (Sutton & Barto, 2018). The simulated birds are memory-augmented stimulus-response machines that do not perform any offline planning or imagination of what it could be like in an alternative situation than the currently perceived one. While the associative recall of food items might be interpreted as a first step towards mental time travel, it is implemented here as a simple pattern completion. Our model is similar to the mnemonic-associative account (Dickinson, 2011), where policies are updated at the time of cache recovery. We conclude that traditional concepts of computational neuroscience are sufficient to explain these experiments on food caching behaviour in birds, but new experiments might falsify our model by providing unequivocal evidence for mental time travel or more flexible memory usage.

Acknowledgments

This work was supported by the Swiss National Science Foundation (Grants 200020_165538 and 200020_184615). J.B. thanks Vasiliki Liakoni, Samuel Muscinelli and Valentin Schmutz for helpful discussions and feedback.

References

Brea, J., & Gerstner, W. (2016). Does computational neuroscience need new synaptic learning paradigms? *Current Opinion in Behavioral Sciences*, *11*, 61–66.

Cheke, L. G., & Clayton, N. S. (2011). Eurasian jays (garrulus glandarius) overcome their current desires to anticipate two distinct future needs and plan for them appropriately. *Biology Letters*, *8*(2), 171–175.

Clayton, N. S., Dally, J., Gilbert, J., & Dickinson, A. (2005). Food caching by western scrub-jays (aphelocoma californica) is sensitive to the conditions at recovery. *Journal of Experimental Psychology: Animal Behavior Processes*, *31*(2), 115–124.

Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, *395*(6699), 272–274.

Clayton, N. S., & Dickinson, A. (1999a). Memory for the content of caches by scrub jays (aphelocoma coerulescens). *Journal of Experimental Psychology: Animal Behavior Processes*, *25*(1), 82–91.

Clayton, N. S., & Dickinson, A. (1999b). Motivational control of caching behaviour in the scrub jay, aphelocoma coerulescens. *Animal Behaviour*, *57*(2), 435–444.

Clayton, N. S., & Dickinson, A. (1999c). Scrub jays (aphelocoma coerulescens) remember the relative time of caching as well as the location and content of their caches. *Journal of Comparative Psychology*, *113*(4), 403–416.

Clayton, N. S., & Emery, N. (2015). Avian models for human cognitive neuroscience: A proposal. *Neuron*, *86*(6), 1330–1342.

Correia, S. P., Dickinson, A., & Clayton, N. S. (2007). Western scrub-jays anticipate future needs independently of their current motivational state. *Current Biology*, *17*(10), 856–861.

de Kort, S. R., Correia, S. P. C., Alexis, D. M., Dickinson, A., & Clayton, N. S. (2007). The control of food-caching behavior by western scrub-jays (aphelocoma californica). *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(4), 361–370.

de Kort, S. R., Dickinson, A., & Clayton, N. S. (2005). Retrospective cognition by food-caching western scrub-jays. *Learning and Motivation*, *36*(2), 159–176.

Dickinson, A. (2011). Goal-directed behavior and future planning in animals. *Animal Thinking*, 79–92.

Fendt, R. (2017). *BlackBoxOptim.jl*. Retrieved from https://github.com/robertfeldt/BlackBoxOptim.jl

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility traces and plasticity on behavioral time scales: Experimental support of neohebbian three-factor learning rules. *Frontiers in Neural Circuits*, *12*, 53.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476.

Gutmann, M. U., & Corander, J. (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *Journal of Machine Learning Research*, *17*(125), 1-47.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, *79*, 2554–2558.

Niv, Y., Joel, D., & Dayan, P. (2006). A normative perspective on motivation. *Trends in Cognitive Sciences*, *10*(8), 375–381.

Raby, C. R., Alexis, D. M., Dickinson, A., & Clayton, N. S. (2007). Planning for the future by western scrub-jays. *Nature*, *445*(7130), 919–921.

Sekeres, M. J., Winocur, G., & Moscovitch, M. (2018). The hippocampus and related neocortical structures in memory transformation. *Neuroscience Letters*, *680*, 39–53.

Stotz, N. G., & Balda, R. P. (1995). Cache and recovery behavior of wild pinyon jays in northern arizona. *The Southwestern Naturalist*, *40*(2), 180–184.

Suddendorf, T., & Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genetic, Social, and General Psychology Monographs*, *123*, 133–167.

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral and Brain Sciences*, *30*(03).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (second ed.). MIT Press, Cambridge, MA.

Vander Wall, S. B. (1990). *Food hoarding in animals*. The University of Chicago Press.