Human learning in non-Markovian decision making

Aaron Clarke
Laboratory of Psychophysics
Brain Mind Institute
EPFL, Switzerland

Johannes Friedrich
Computational and Biological Learning Lab
Department of Engineering
University of Cambridge, UK
jf517@cam.ac.uk

Walter Senn
Computational Neuroscience Lab
Department of Physiology
University of Berne, Switzerland

Elisa Tartaglia
Department of Neurobiology
University of Chicago, IL

Silvia Marchesotti
Laboratory of Psychophysics
Brain Mind Institute
EPFL, Switzerland

Michael H. Herzog
Laboratory of Psychophysics
Brain Mind Institute
EPFL, Switzerland

Abstract

Humans can learn under a wide variety of feedback conditions. Particularly important types of learning fall under the category of reinforcement learning (RL) where a series of decisions must be made and a sparse feedback signal is obtained. Computational and behavioral studies of RL have focused mainly on Markovian decision processes (MDPs), where the next state and reward depends only on the current state and action. Little is known about non-Markovian decision making in humans. Here we consider tasks in which the state transition function is still Markovian, but the reward function is non-Markovian. For example, learning to go from A to B is non-Markovian when receiving a reward at B is contingent on having visited a switch-state C before arriving at B. Learning is also non-Markovian when feedback is delayed and there is no unique mapping between feedback and state-action pairs. Classical RL algorithms can be categorized into value based methods, such as temporal difference (TD) learning, and policy gradient methods. The former cannot cope with such non-Markovian conditions, whereas policy gradient methods do, but are infamous for being slow. Here, we show that humans can learn both, with non-Markovian switch states and delayed feedback. Human learning with switch-states is nearly Bayes-optimal, whereas learning with delayed feedback is Bayes-suboptimal. Stikingly, both tasks are well modeled with a spiking neural network using a cascade of eligibility traces to implement a policy gradient procedure.

Keywords: non-Markovian; decision making; policy gradient; Bayes; psychophysics

Acknowledgements

Aaron Clarke and Elisa Tartaglia were funded by the Sinergia project and by the "Perspective Researcher fellowship" of the Swiss National Science Foundation (SNSF). Johannes Friedrich was funded by Neurochoice within the SystemsX initiative (evaluated by the SNSF) and the "Perspective Researcher fellowship" of the SNSF.

1 Introduction

The bulk of research on (model-free) reinforcement learning in neuroscience has centered on temporal difference (TD) algorithms. Indeed, the phasic activity of dopamine neurons in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc) appears to mimic the error signal in the algorithm [5]. Such TD algorithms assume that the underlying decision process is Markovian, i.e. the new state depends on only the previous state and action, but not the history (formally, $p(s_{t+1}|a_t, s_t) = p(s_{t+1}|a_t, s_t, a_{t-1}, s_{t-1}, ..., a_0, s_0)$. These RL models, can go awry under non-Markovian conditions [4]. Non-Markovian learning is more complex than coping with delayed and sparse feedback. For example, if feedback signals are delayed randomly and independently, such that feedback for state t+1 may come before feedback for state t, then the learning scenario is non-Markovian. Alternatively, if an agent is rewarded for traveling from A to B via C, but not when skipping C, then this switch-state scenario is also non-Markovian. These processes are non-Markovian because the new state depends on more than the previous state. Policy gradient methods can cope with non-Markovian situations [2, 6, 3, 4] but are infamous for being slow. Here we consider a "bottom-up" model implemented with spiking neurons together with a policy gradient learning rule that depends on the pre- and post-synaptic spike-timings and compare it with "top-down" models of a Bayesian learner. In the first considered switch-state experiment the performance of both algorithms is similar and coincides with the one of humans. In the second task involving randomly delayed feedback the Bayesian model learns much faster than the spiking network, but the latter actually matches human subjects better.

2 Psychophysical Experiments

2.1 Walk-the-Dog Experiment

Participants were instructed that throughout the experiment, they would be presented with one of eight images at a time and that in order to proceed to the next image they had to make mouse click on one of three green disks presented below the image (Fig. 1**A**). At the trial sequence's end a "Yeah!" appeared instead of an image. The participants were instructed that their goal was to reach the "Yeah!" as often as possible within 10 minutes. They were further informed that the associations between disks and images would not change throughout the experiment. It was initially unbeknownst to the participants that, in order to reach the goal image, they had to pass through the image marked "Key State" on the figure in order to have access to the goal as shown in Fig. 1**A**. This constituted their learning task. If participants tried to go directly to the goal without visiting the key state image then they were re-directed to the image situated graphically to the penultimate image' s left. setup contained both recurrent and outward-bound connections. There were two runs of 10 minutes each, where participants completed as many episodes as they could in the allotted time. The number of images visited per episode was recorded as a function of episode number. From the first 10 minute run to the second, the images assigned to each node were re-randomized, but the underlying node-connection structure stayed the same. The effects of a left/middle/right disk press were randomized over images.

2.2 Delayed Feedback Experiment

Eight new participants learned image-classification pairings (with two possible classifications - left or right - per image) with *randomly delayed* feedback. This task is inherently difficult, so we built up task-difficulty gradually. The general procedures for these experiments are illustrated in Figure 2.

Part I: Training with immediate feedback: We started with a basic task that involved learning of only four images (Fig. 2**A**). Participants received immediate feedback about the correctness of their classifications after each image presentation. Feedback was provided in the form of a red or green bar presented at the bottom of the screen indicating an incorrect or a correct response respectively. This step acquainted the participants with the basic task.
Part II: Training with delayed feedback: Next, we replaced the original four images with four new images. We increased task difficulty by randomly delaying the feedback time for each image according to a gamma distribution (Fig. 2**D**) with a shape parameter of $k = 2$ and a scale parameter of $\theta = 1.5$ seconds. This scenario allowed feedback for different classifications to go out of order with respect to the image order (Fig. 2 B). The task was still to learn the correct classifications (left or right) that the computer had randomly assigned to each image. Here, it was possible that the feedback for an image could have been delayed to the point where it was presented simultaneously with a later image.
Part III: Main experiment: We increased task-difficulty by requiring participants to learn image-classification pairings for 10 images, again with two classifications per image and with randomly delayed feedback following a gamma probability density function with a shape parameter of $k = 2$ and a scale parameter of $\theta = 1.5$ (Fig. 2 C).
Part IV: Memory control-experiment: To test our participants' basic associative memory capacities we had them learn image-classification pairings for 10 new images with immediate feedback.
Part V: Swapping control-experiment: We examined participants' efficacy at reversal learning, that is, learning where some of the images switch which of their classifications are rewarded. Here, participants repeat the Memory experiment, but with half of the images having switched classification categories (right to left and visa versa).
Part VI: Replication of the main experiment: Here, we replicated the main experiment of *Part III* with 10 new images.

Our results suggest that, indeed humans are capable of learning under such conditions, however, the first run revealed a separation into learners and non-learners (Fig. 2**E**). The hypothesis that the non-learners were simply having trouble remembering the 10 image-classification pairings was ruled out by the two further control experiments with immediate feedback (Fig. 2**G**,**H**). We hypothesized that the non-learners may have been capable of learning the task, but that they needed more practice. Indeed, in the replication of the main experiment (Fig. 2**F**) all participants learned the task..

3 Modeling

3.1 Population of spiking neurons: Walk-The-Dog Experiment

Simulations were performed with a population of spiking neurons as described in [4] but using a larger network with 100 neurons and an input dimension of size 250. The increase in input dimension in comparison to [4] was compensated for by a reduction of the input Poissonian spike rate from 6 to 2 Hz. The position was encoded by 150 input spike trains and an additional 100 spike trains coded for the previous position. Figure 1 **B** and **C** plot this model's mean performance (±SEM) over five runs for the walk-the-dog experiment.

The population of spiking neurons is oblivious to the episodic structure of the task, the deterministic state transitions, the fact that only at the end of an episode reward is received and that this reward is single valued. The neural network is able to cope with a much wider range of tasks including e.g. probabilistic state transitions, probabilistic rewards of different sizes at various times and even non-episodic tasks. Indeed, the same population of spiking neurons is able to learn in the delayed feedback experiment demonstrating its multi purpose applicability.

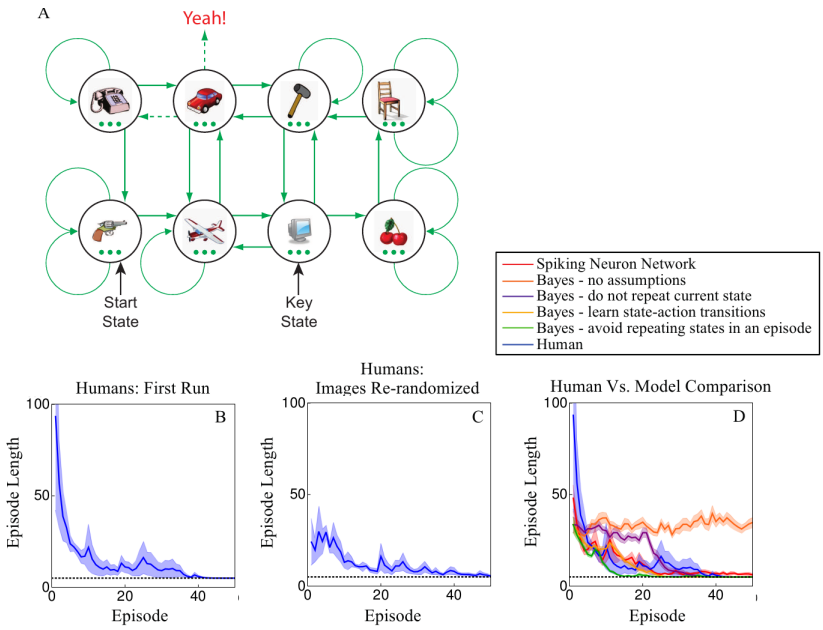

Figure 1: The 2D walk-the-dog experiment. **A**. Each image was presented in isolation with three disks below. Participants clicked on a disk to proceed to the next state. Participants started at the state in the bottom-left hand corner of the graph structure, marked here as "Start State". The green arrows indicate the three possible actions available in each state. The target "Yeah!" could be reached only from one state (the car in this example), and only if the state marked "Key State" had been visited first (the monitor in this example). If the key state had not been visited first, then the same action, that brought the participant to the target, would instead bring him or her to a different state (e.g. the telephone in this example). **B**. Human data showing the participants first run through the experiment for comparison with **C**, which shows human data from the second part of the experiment, where the images assigned to each node were re-randomized. Solid lines represent mean data averaged over three subjects and shaded regions denote ±SEM. Participants learned the task within 50 episodes and showed similar learning patterns between the first learning session and the second image-re-randomized session. **D**. Comparison of human results (blue – re-plotted from **B**) with various model implementations. The implemented models include: a population of spiking neurons (red), a Bayesian learner with no assumptions about the task structure (orange), a Bayesian learner that does not take actions leading directly back to the current state (purple), a Bayesian learner that explicitly learns the Markovian state-action transitions (yellow), and a Bayesian learner that avoids repeating states visited earlier in the current episode (green).

3.2 Bayesian learner: Walk-The-Dog Experiment

Whereas the population of spiking neurons is oblivious to the aforementioned structural aspects of the task, the human subjects are instructed about it. We incorporate this prior knowledge into a Bayesian learner, thus tailoring it to the specific task at hand. The learner tries to maximize its reward rate, i.e. it tries to find a *shortest* rewarded sequence of actions. We obtain the probabilities that a sequence of actions is in the set of shortest rewarded sequences through Monte Carlo sampling. For sequence selection we considered four different methods, with each subsequent method making more assumptions than the last. The first method learns directly using the MAP estimate over all sequences. It does not incorporate any information about states, but in a most general fashion takes into account that the next state might depend on the full history of actions and memorizes the latter completely. In the task there are recurrent actions which never change the current state. In the second method, the learner considers those as redundant and does not repeat such actions in subsequent trials.

The MAP estimate is not over all sequences any more, but sequences that are known to the learner to contain a recurrent action are excluded. This reduction in search space speeds up learning. In the third method, the learner considers not just actions but takes the states into account. It assumes that the next state only depends on the current state and the performed action, but not on the history, i.e. it assumes the Markov property for the state transition dynamics (but not the reward dynamics). The agent learns a model thereof while interacting with the environment. Based on this model the agent can infer the state sequence that corresponds to an action sequence. Moreover the learner still omits actions that are already known to the agent to be recurrent. While in the previous method the learner memorized for which history an action is recurrent, here it memorizes the next state for each action and state, in particular for which state an action is recurrent. Hence an information transfer takes place between sequences that lead to the same state, decreasing the size of the search space even more by ruling out further sequences. In the fourth method, based on its current knowledge of the environment, the agent not only omits actions that do not leave the current state, but also avoids those that lead to a state which has already been visited earlier in the current episode, hence preventing looping. If the agent performed a sequence that led to a state in which the latter restriction is not satisfiable any more, it lets go of the latter constraint and continues the sequence as in the third learning method.

Humans outperform the model-free Bayesian learner and perform almost as well as the Bayesian agent that learns a model of the environment and makes some correct prior assumptions about the task structure. Surprisingly, our network of spiking neurons with a plain policy gradient procedure learns nearly as fast as humans and the Bayesian learner.

3.3 Population of spiking neurons: Delayed Feedback Experiment

Simulations were performed with a population of spiking neurons as described in [4]. We used 100 neurons and an input dimension of size 250. Fig. 2 **E**, **F** **G** and **H** plot this model's mean performance (±SEM) over five runs for the delayed feedback experiment.

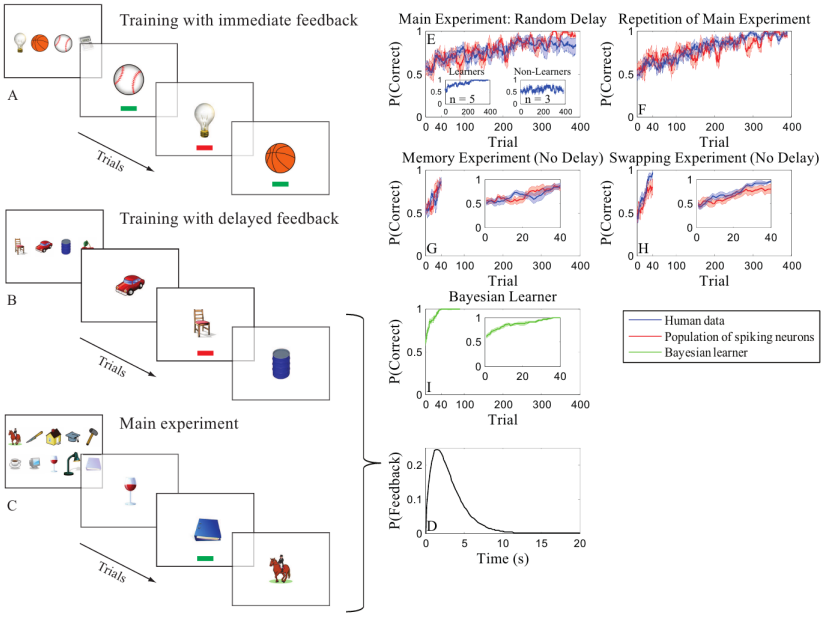

Figure 2: Procedure for Parts I (**A**), II (**B**) and III (**C**) of the delayed feedback experiment. Participants were first shown four images. Following this, participants were presented with one image at a time. Each image belonged to either the "left" or "right" category and observers were asked to learn to which category each image belonged. Responses were followed by feedback in the form of a red or green bar (for incorrect and correct responses respectively) at the bottom of the screen. Feedback for part I was immediate, while feedback for parts II and III was randomly delayed by the γ-probability density function shown in **D**. **E-I**. Delayed feedback experiment results. Proportion correct versus trial number plotted for simulation results from a population of spiking neurons (red) compared with human performance (blue) for the ten images of the experiment. **E**. Part III, random delay. Insets show separate averages for participants who could learn to do this task (learners) and those who could not (non-learners). F. Part V repetition of the main experiment. **G**. Part IVa Memory experiment with no delay. **H**. Part IVb Swapping experiment with no delay. I. Simulation results for a Bayesian learner (green). **G-I**. Insets show performance over the first 40 trials. Human performance is matched by a population of spiking neurons but is inferior to a Bayesian learner when learning stimulus-response associations with randomly delayed reinforcement.

3.4 Naïve Bayesian learner: Delayed Feedback Experiment

We make the assumption that the obtained rewards are independent when conditioned on the images' classification labels, which is known as naïve Bayes. Because the learner does not know the parametric form of the delay distribution, we use a flexible nonparametric method. More precisely, we use a Dirichlet Process (DP) as a prior over the delay distribution. Pattern label assignments are made by calculating the posterior probability distribution for the presented pattern's label, which involves marginalizing over the other patterns' labels as well as the delay distribution, and using the maximum a posteriori (MAP) estimate. As base distribution for the DP we used a Cauchy distribution, motivated by the fact that it is asymptotically scale free. The DP's concentration parameter cancels due to marginalization in the decision making. Figure 2 **I** plots the mean performance (±SEM) over 30 trials for the delayed feedback experiment.

4 Discussion

Animals and humans can learn from sparse and delayed feedback. Reinforcement learning models have provided powerful algorithms that do well at explaining the outcomes of behavioural experiments. These models cover only Markovian learning situations, that is, situations where the outcome of an action depends only on the current state. Recently it was shown that classical reinforcement learning models can be extended to cope with non-Markovian learning situations [9, 4]. These models employ a population of spiking neurons which update their weights based both on a global reward feedback signal and on feedback about the population response. Furthermore, they employ a cascade of synaptic memory traces, which allows the model to handle substantially delayed reinforcement signals (relative to action performance) and still learn. Here we ask whether humans are able to learn in non-Markovian situations and, if so, what are the characteristics of this learning process? In the walk-the-dog experiment, we created a non-Markovian environment which required observers to learn to visit a switch state that would allow them to reach the goal later on. The result was that all five observers were able to learn very quickly under this condition. For these experiments, human performance was well describes by both, the spiking neural population model and by a Bayes-optimal model.

In a second set of experiments, we tested whether humans can learn from substantially delayed feedback. In this experiment, humans learned to classify each of 10 pictures as belonging to one of two classes. Feedback about their decisions was randomly delayed by an average of 3 seconds. More importantly, there was no unique mapping between images and feedback because feedback for image one may have appeared after the feedback for image three. This is a non-Markovian situation because calculating the appropriate state- action value update requires averaging rewards over multiple past episodes. Even though this is a very challenging task, human observers coped very well. Only half of the observers managed to perform the task during the first run, but all observers were successful in the second run. Interestingly, when we swapped the image labels (left to right and right to left) the participants re-learning was rather quick, i.e. executive function seems to outperform memory here [7]. Judging from the learning speed with randomly delayed feedback, performance lies more closely in line with the spiking neuron network of [4] than with the Bayesian learner. This implies that learning with delayed feedback is less optimal than learning with a switch state. Furthermore, it is likely that this task limits the cognitive strategies that could be employed, thereby reducing human observers to low-level learning strategies such as those employed in the spiking neural network model. Several papers have looked at non-Markov decision processes in machine learning, in particular within the framework of partially observable Markov decision processes [8], but [1] and our current paper seem the first to look at them in human learning. Here we showed that humans can cope with non-Markovian situations quite well. Just recently, algorithms were put forward which can cope with non-Markovian situations. Our results are captured well by the models tested. The Bayesian model (which yields an upper bound on learning performance) provides an accurate description of participant learning in the walk-the-dog experiment, while the spiking neuron model described both experiments well. Despite the fact that the neural model does not contain cognitive strategies, whereas humans seem or claim to use both cognitive and implicit cues, the model provides fairly good approximations to human learning ability in the considered tasks. Future work should examine learning situations where cognitive strategies are indeed of avail and develop spiking neural population models that implement both low-level reinforcement learning strategies and that form higher-level hierarchical models of the state-action decision space.

References

[1] D. Badre, A.w S. Kayser, and M. D'Esposito. Frontal cortex and the discovery of abstract action rules. *Neuron*, 66(2):315–326, 2010.

[2] J. Baxter and P. Bartlett. Infinite-horizon policy-gradient estimation. *J Artif Intell Res*, 15:319–350, 2001.

[3] I. R. Fiete and H. S. Seung. Gradient learning in spiking neural networks by dynamic perturbation of conductances. *Phys Rev Lett*, 97(4):048104, 2006.

[4] J. Friedrich, R. Urbanczik, and W. Senn. Spatio-temporal credit assignment in neuronal population learning. *PLoS Comput Biol*, 7(6):e1002092, 2011.

[5] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[6] H. S. Seung. Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron*, 40(6):1063–1073, 2003.

[7] D. A. Simon and N. D. Daw. Environmental statistics and the trade-off between model-based and TD learning in humans. In *Advances in Neural Information Processing Systems 24*, pages 127–135, 2011.

[8] R. D. Smallwood and E. J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Oper Res*, 21:1071–1088, 1973.

[9] R. Urbanczik and W. Senn. Reinforcement learning in populations of spiking neurons. *Nat Neurosci*, 12(3):250–252, 2009.