Human reinforcement learning processes act on learned attentionally-filtered representations of the world

Yuan Chang Leong
Department of Psychology
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08544
yl@princeton.edu

Yael Niv
Department of Psychology
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08544
yael@princeton.edu

Abstract

Reinforcement learning (RL) models are often applied to study human learning and decision-making. However, simple RL algorithms do not fare well in explaining learning behavior in real world situations where the environment is high-dimensional and the relevant states are not known. As a solution, we propose that RL processes act on an attentionally-filtered representation of the environment. This improves the computational efficiency of RL by constraining the state-space that the learning agent has to consider. We further propose that the attention filter is learned and is dynamically modulated according to the outcomes of ongoing decisions. To test our hypotheses, we had participants perform a decision-making task with multi-dimensional stimuli and probabilistic awards. Model-based analysis of participants' choices suggests that participants prefer strategies that favor computational efficiency at the expense of statistical optimality. To better study the dynamics of attention, we had a group of participants perform a variant of the task in which they had to select the dimensions they wanted to view before making their choice. We treated the viewed dimensions as a proxy for participants' attention filter. Our models fit the data better when learning was restricted to attended dimensions, suggesting that participants do indeed constrain choice and learning to a subset of dimensions. Finally, attention dynamics themselves were best explained by a model that preferentially attended to dimensions with features that have acquired high value over the course of learning. This result provides evidence that the attention filter is dynamically modulated as participants receive feedback from ongoing decisions.

Keywords: attention, state representation, function approximation, active-sensing, human decision-making

Acknowledgements

We are grateful to Angela Radulescu, Reka Daniel and Andra Geana for their invaluable input to this project. This work was supported in part by NIH grant R01MH098861 and by an Alfred P. Sloan Research Fellowship and a New Scholar award from the Ellison Medical Foundation to YN.

1 Introduction

The framework of reinforcement learning (RL) has had a tremendous impact on the fields of psychology and neuroscience. In particular, the temporal difference (TD) learning model has helped advance our understanding of animal and human learning by providing a mathematically precise definition of how an agent learns the association between predictive stimuli and rewards [1]. In laboratory controlled experiments, where the state-space is well-defined, TD learning has indeed provided a good account for both behavioral and neural data [2, 3]. Real-world learning, however, takes place in a highly complex and multidimensional environment, which poses a challenge to the TD learning framework.

In particular, it is a well-known problem in operations research and machine learning that the number of states of a task increases exponentially with increasing number of dimensions on which these states are defined. This is known as the *curse of dimensionality* [4]. Since TD learning assigns values to states (or state-action pairs), the amount of experience required to arrive at an approximately correct value for all states increases with the number of states. Yet both animals and humans can solve complicated learning problems with limited experience.

We propose that efficient learning is possible because people employ selective attention as a learning strategy [see also 5, 6]. The role of selective attention in regulating cognitive processes is well established [7]. Here, we hypothesize that attention facilitates learning by carving out the state-space that RL operates on. We further postulate that this attention filter is learned, and as such, is dynamically modulated by the outcomes of ongoing decisions.

To test our hypotheses, we had human participants perform 3-armed bandit tasks with multidimensional stimuli and probabilistic rewards. We found that participants opt for computationally efficient strategies at the expense of statistical optimality. In addition, we found that participants' attentional-selection strategy was best described by a model that allocates attention according to the learned values of stimuli. As the values are updated with ongoing learning, the attentional filter is also modulated. The current results support our hypothesis that learning is constrained by attention, but we also learn what to attend to.

2 Experimental Tasks

2.1 Faces/Houses/Tools (FHT) Task

The FHT task was designed to be a simplified analog of many real world problems, where people have to make decisions based on multidimensional information under conditions in which most dimensions are uninformative to the decision at hand. On each trial, participants chose between three bandits, each described by features from three different dimensions: a face, a house and a tool (Figure 1a). Stimuli on each trial were generated by randomly recombining features from each dimension. In any one 'game' only one dimension (e.g., tools) was relevant to determining reward and only one target feature in that dimension (e.g., saw) was associated with a high reward probability ($p = 0.75$). Choosing stimuli that did not contain this feature yielded a reward with only $p = 0.25$. Participants were not told which dimension was relevant, and were tasked with getting as much reward as possible. Eighteen participants performed the FHT task and were paid \$12-\$15 according to their performance. Each participant played 56 games of 25 trials each.

2.2 Active Sensing Faces/Houses/Tools (asFHT) Task

To study attention processes in the FHT task, we had a separate group of participants play a variant of the task where they had to select, via button presses, the dimensions they would like to view before making their choices. On each trial, they could choose to view as many dimensions as they wish, before selecting a stimulus (Figure 1b). Recent work suggests that attentional sampling is an active process that shares many similarities with sensorimotor sampling routines [8]. As such, it is reasonable to assume that the dimensions participants chose to view would also be the dimensions they would have chosen to attend to. Nineteen participants performed the asFHT task and were paid \$12-\$15 according to their performance. The experiment began with 10 games of the FHT task to allow participants to familiarize themselves with the task structure. Participants then played 25 games (25 trials each) of the asFHT task.

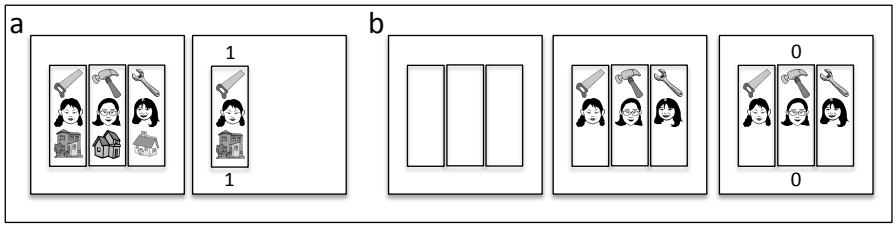

Figure 1: Stimulus displays for *a.* FHT Task and *b.* asFHT task.

3 Model-Based Analysis

3.1 Models of Choice Behavior

We analyzed participants' choice behavior in the FHT task using three computational models[1]. The first model is a *Bayesian model* that uses Bayes' rule to infer the posterior distribution that each feature f in dimension d is the target feature given all previous data $D_{1:t}$ and rewards $r_{1:t}$

$$p(d, f|D_{1:t}, r_{1:t}) \propto p(r_t|d, f)p(d, f|D_{1:t-1}, r_{1:t-1}). \tag{1}$$

This distribution is then used to compute the probability of reward for each stimulus S_i [5, 6]:

$$V_t(S_i) = p(r_t = 1|S_i, D_{1:t-1}) = \sum_d p(d, f \in S_i|D_{1:t-1}, r_{1:t-1})\rho_h + (1 - p(d, f|D_{1:t-1}, r_{1:t-1})\rho_l \tag{2}$$

where ρ_h is the probability of reward for the target feature and ρ_l is the probability of reward for a non-target feature. This model incorporates all available information across all dimensions and features in an statistically accurate manner. As such, it assumes statistically optimal learning about all features at once, akin to a diffused focus of attention.

The second model is a *function approximation* (FA) model that learns a weight w for each of the nine features. It then computes the value of stimulus S_i as the average of feature weights of this stimulus:

$$V_t(S_i) = \frac{1}{n}\sum_{d=1}^{n} w_t(d, f \in S_i) \tag{3}$$

The feature weights for the chosen stimulus c_t are updated according to TD learning [1]:

$$\delta_t = r_t - V_t(c_t) \tag{4}$$

$$w_{t+1}(d, f \in c_t) = w_t(d, f \in c_t) + \frac{1}{n}\eta\delta_t \qquad \forall d = 1, 2, 3 \tag{5}$$

where δ_t is the prediction error for trial t, n is the number of dimensions and η is the learning rate. The FA model is less statistically optimal than the Bayesian model as it learns point estimates of feature weights and does not maintain the full posterior distribution over all features. It also learns only about chosen features, and thus assumes a stronger focus of attention than the Bayesian model.

Finally, the *Decay model*, is identical to the FA model, but in addition, in this model weights of unchosen features decay to zero:

$$w_{t+1}(d, f \notin c_t) = (1 - \eta_k)w_t(d, f \notin c_t) \qquad \forall d = 1, 2, 3 \tag{6}$$

where η_k is the decay rate. The Decay model is the least statistically optimal model of the three as it loses information about unchosen features on each trial. However, it assumes the strongest attention focus since only features that have been consistently attended to and chosen can acquire weights significant enough to influence choice.

3.2 Models of Attention

In general, there are seven possible combinations of viewed dimensions on each trial: one dimension only (three possible options), a combination of any two dimensions (three possible combinations) or all three dimensions.

To model the dimensions participants chose to view on each trial, in the *Dimension Value* model we assumed that attention is costly. As such, participants must balance the benefits of attending to a dimension with the associated cost. The model thus computes the value of each attention combination a_t as the utility of attending to the dimensions in that combination minus a cost that scales with the number of attended dimensions:

$$V_t(a_t) = \sum_{d \in a_t} U_t(d) - nJ \tag{7}$$

where $U(d)$, the utility of attending to dimension d, was determined by summing the weights of features along d, such that the utility of viewing a dimension would be higher when the model has learned high weights for features in that dimension. In equation (7), n is the number of dimensions attended to and J is a cost penalty per dimension attended.

As a baseline for comparison with this model, we tested a model in which participants narrow their focus of attention over time regardless of the specific utility of each dimension. In this *Game Horizon* model, the utility of attending to each dimension is fixed (i.e., $U_t(d) = K$), but the cost of attending to dimensions increases over the course of the game:

$$V_t(a_t) = nK - nJT \tag{8}$$

where n is the number of dimensions viewed in option a_t, and T is the trial number in the current game.

[1]Choice behavior in the asFHT task was analyzed using the same three models. However, the models were modified such that both value computation and update depended only on viewed dimensions.

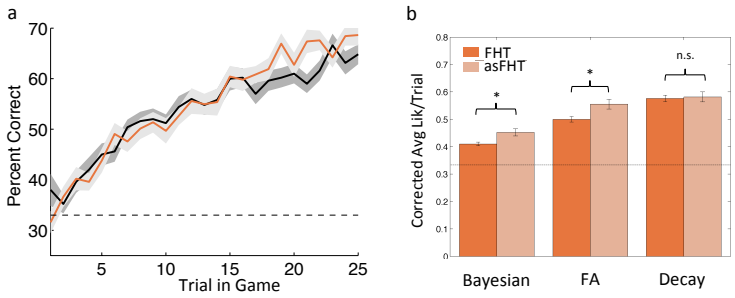

Figure 2: Participants' choice behavior: *a.* Learning curves for the FHT task (black) and asFHT task (orange). *b.* Model performance on data from FHT Task (dark shading) and asFHT Task (light shading). Models were evaluated using corrected average likelihood per trial. Error bars denote SEM; dashed lines denote chance level (33%); * denotes $p < 0.05$.

3.3 Model Comparison and Parameter Estimation

For all models, a "softmax" policy was used to compute the probability of making a particular choice[2], $\pi(c)$, or allocation of attention, $\pi(a)$, on each trial:

$$\pi_t(x) = \frac{e^{\beta V_t(x)}}{\sum_i e^{\beta V_t(i)}} \tag{9}$$

where $x = c$ or a, i enumerates all actions available to the model on that trial, and β is an inverse temperature parameter that determines the balance between exploration and exploitation.

Model parameters were optimized by finding participant-specific parameters that maximized the log likelihood of the participant's data given the model. These parameters were then used to compute the Bayesian Information Criterion (BIC) approximation of model evidence [9], E_M:

$$E_M \approx log(p(D|M, \hat{\theta}_M)) - \frac{||\hat{\theta}||}{2} log\ N \tag{10}$$

where $p(D|M, \hat{\theta}_M)$ is the likelihood of data D given model M and parameters $\hat{\theta}_M$, $||\hat{\theta}||$ is the number of free parameters in the model and N is the number of data points (trials). To provide a more intuitive measure of model evidence, we divided the total score for each model by the number of trials for which a participant provided a response, and exponentiated it, to yield a complexity-corrected average likelihood per trial that varies between 0 and 1.

4 Results

4.1 Choice Behavior

We first evaluated participants' performance by calculating the percentage of trials on which participants chose the stimulus containing the target feature. Figure 2a shows performance as a function of trial within a game. A between-subjects repeated measures ANOVA did not find a main effect of task type ($F(24, 1) = 0.32, p = 0.57$), indicating that there was no significant difference between participants' learning of the two tasks.

As is clear from Figure 2b, all models of choice behavior performed considerably better than chance (one-tailed t-tests, $p < 0.001$). For both tasks, the Decay model was best supported by the data (two-tailed t-tests, $p < 0.001$). The Bayesian and FA models fit data from the asFHT task better than that from the FHT task (two-tailed t-tests, $p < 0.05$). There was no significant difference between the fits of the Decay model for the two tasks ($t(35) = 0.72, p = 0.24$).

4.2 Attention

In the asFHT Task, the number of viewed dimensions decreased over the course of each game (trial 1: Mean = 2.34, SE = 0.17; trial 25: Mean = 1.67 , SE = 0.11; Figure 3a). Both the Game Horizon model and the Dimension Value model predicted participants' focus of attention better than chance (one-tailed t-tests, $p < 0.001$). However, the Dimension Value model performed significantly better than the Game Horizon model ($t(18) = 9.1, p < 0.001$, Figure 3b).

[2]For the attention models, feature weights were first generated using the best-fitting model of choice behavior.

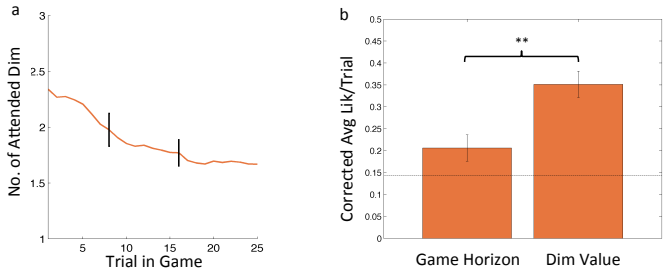

Figure 3: Dynamics of attention processes in the asFHT Task. *a.*. Average number of attended dimensions over the course of a game. *b.* Model performance on predicting participants' focus of attention. Error bars indicate SEM; dashed line indicates chance level (0.14%); ** indicates $p < 0.001$.

5 Discussion

Our experimental tasks were designed to mimic some aspects of the cluttered multidimensional state in which real-world learning and decision making are embedded in. While participants could learn to solve these tasks, an analysis of their choice behavior revealed that they were not doing so in a statistically optimal manner. Consistent with previous work [5, 6], we found that participants' choice behavior was best explained by strategies that were computationally efficient, but statistically suboptimal. In particular, the best-fit "Decay model" was an RL model with a function approximation architecture that also decays weights of unchosen features. This model learns only about chosen features, and loses useful information on each trial. However, it places low demands on computational resources by relying on algorithms that are less computationally costly and thus can easily scale up to many dimensions. Such a strategy may reflect a necessary compromise between optimal learning and computational demands given limited resources.

In some senses, the Decay model also assumes the narrowest focus of attention. We had hypothesized that attention might serve to further reduce computational demands by carving out a suitable state-space for efficient learning. The asFHT task was designed to directly test this hypothesis. Interestingly, despite the fact that participants limited their viewing to only a subset of dimensions, comparison of learning curves indicated no difference in learning between this and the FHT task in which all dimensions were always available. Moreover, incorporating information about participants focus of attention significantly improved the model fits for both that Bayesian model and the FA model. That is, even though participants in the FHT task were presented with all dimensions, they might have been learning and making choices based only on the subset of dimensions that they were attending to. Finally, it is noteworthy that performance of the Decay model was not significantly different between the two tasks. We suggest that this is because the Decay model implicitly implements a selective attention component for both tasks: due to the weight decay, weights of features that are not consistently chosen (presumably because they are not being attended to) decay to zero and thus choice in this model comes to be driven by features that have been consistently attended to. This interpretation is compatible with a role for explicit attentional mechanisms in learning and decision-making, though further work needs to be conducted to formally explore the relationship between the Decay model and attention.

Lastly, we were interested in understanding how the focus of attention changes with learning. Here we found that participants' focus of attention depended on the value of attending to the different dimensions. In the beginning of each game, when participants had no information about any of the dimensions, it was worthwhile to attend to multiple dimensions. However, as they learned more about each dimension, it was more efficient to attend only to dimensions with high-value features. Taken together, these results demonstrate an intricate relationship between learning and attention—attention constrains what we learn about, but we also learn what to attend to.

References

[1] Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction. Cambridge*, MA: MIT Press.
[2] Niv, Y. (2009). *Reinforcement learning in the brain. Journal of Mathematical Psychology*, 53(3), 139154.
[3] Balleine, B. W., Daw, N. D., & O'Doherty, J. (2009). Multiple forms of value learning and the function of dopamine. In P. W. Glimcher, C. F. Camerer, C. Fehr, & R. A. Poldrack (Eds.), *Neuroeconomics: Decision Making and the Brain* (pp. 367387). London: Academic Press.
[4] Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press
[5] Gershman, S.J., Cohen, J.D., and Niv, Y. (2010). Learning to selectively attend. In *32nd Annual Conference of the Cognitive Science Society*, Portland
[6] Wilson, R. C., & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in human neuroscience*, 5, 189.
[7] Miller, E K, & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24, 167202.
[8] Schroeder, C. E. et al. (2010). Dynamics of Active Sensing and perceptual selection. *Current opinion in neurobiology*, 20(2), 1726.
[9] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461- 464.