Reward-sensitive attention dynamics during human reinforcement learning

Angela Radulescu
Department of Psychology
Princeton University
Princeton, NJ 08540
angelar@princeton.edu

Yuan Chang Leong
Department of Psychology
Stanford University
Stanford, CA 94305
ycleong@stanford.edu

Yael Niv
Department of Psychology
Princeton Neuroscience Institute
Princeton University
Princeton, NJ 08540
yael@princeton.edu

Abstract

Selective attention is thought to facilitate reinforcement learning (RL) in multidimensional environments by constraining learning to dimensions that are most relevant for the task at hand. But how would agents know what dimensions to attend to in the first place? Here we use computational modeling of human attention data to show that selective attention is sensitive to trial-by-trial dynamics of reinforcement. Twenty-five participants performed a decision-making task with multi-dimensional stimuli, while undergoing functional magnetic resonance imaging (fMRI) and eye-tracking. At any one time, only one of three stimulus dimensions (faces, houses or tools) was relevant to predicting probabilistic reward. Participants had to learn, through trial and error, which was the predictive dimension, and what feature within that dimension was the most rewarding. We chose this task design in order to capture real-world learning problems where only some dimensions in the environment consistently predict noisy reward. In previous work we showed that attention to different dimensions modulates learning in this task. To examine how subjects learn what to attend to, here we developed and compared different models that specify how attention changes trial-by-trial. Both the neural and eye-tracking data were best explained by an RL model that tracks feature values learned through trial-and-error, and allocates dimensional attention in proportion to the highest-valued feature along each dimension. This model outperformed models that determined attention based on choice history alone, suggesting that attention dynamically changes as a function of recent reward history. To our knowledge, ours is the first explanation of how attention measured directly and simultaneously from neural data and eye-tracking is determined. Our results establish a bidirectional interaction between attention and RL: attention constrains what we learn about, and learned values determine what we attend to.

Keywords: selective attention, reinforcement learning, human decision-making, behavioral modeling, Dimensions task

Acknowledgements

This research was supported by NIMH award R01MH098861 award. We thank Nicolas Schuck for helpful discussions.

1 Introduction

Any RL algorithm requires maintaining and updating the value of a possible set of stimuli or environment configurations (formally, states of the task). The efficiency of the learning process (in terms of how long it will take to learn a solution to the task), and the quality of the learned solution, both depend on choosing an appropriate state representation [1]. For instance, when learning to pick avocados, it is more important to consider color and firmness rather than shape or smell. This raises a fundamental question: in a world where features are abundant, how do agents choose which to focus on and learn about? One possibility is to employ selective attention to narrow down the dimensionality of the task [2] [3] [4] [5]. Selective attention prioritizes a subset of environmental dimensions for valuation and learning while generalizing over others, thereby reducing the number of different stimulus configurations that the agent must consider. Selective attention can thus be thought of as a mechanism for controlling which world features are used to determine the internal state [6]. Previous work has shown that selective attention to some dimensions but not others influences human trial-and-error learning, significantly reducing the computational cost of learning in multidimensional settings [7] [2] [8] [9].

In order to facilitate learning, attention has to be directed towards dimensions of the environment that are important for the task at hand (i.e., dimensions that predict reward). However, what dimensions are relevant to any particular task is not always known a priori, and might itself be learned through experience. In other words, one has to first learn what to attend to. We propose that a bidirectional interaction exists between attention and learning in high-dimensional environments [10] [9], and here we provide evidence for one direction of this interaction: human selective attention dynamically changes as a function of recent reward history.

To test this hypothesis, we had human participants perform a reinforcement learning (RL) and decision-making task with compound stimuli, each comprised of a face, a house and a tool, while undergoing functional magnetic resonance imaging (fMRI) scans. Using eye-tracking and multivariate pattern analysis (MVPA) of fMRI data [11], we obtained a quantitative measure of participants' attention to different stimulus dimensions on each trial. To study how attention was modulated by outcomes of ongoing learning, we built computational models of trial-by-trial changes in the focus of attention [12].

2 Participants and methods

Twenty-five participants performed the task. On each trial, participants chose between three columns, each comprised of a face, a house and a tool (Fig. 1A). Mimicking real world learning problems where only a subset of dimensions in the environment is relevant for the task at hand, at any one time, only one of three stimulus dimensions (faces, houses or tools) was relevant to predicting probabilistic reward. Participants had to learn, through trial and error, which was the predictive dimension, and what feature within that dimension was the most rewarding. We obtained two trial-by-trial measures of participants' attention to each dimension. First, we computed the proportion of time participants spent looking at each dimension on a given trial. Second, using MVPA methods we quantified face-, house- and tool-selective neural activity on each trial. Each measure provided a numeric vector of three "attention weights" that sum to 1, denoting the attention towards each of the three dimensions on that trial. Our goal was to build a model that can predict trial-by-trial fluctuations in attention weights from past choices and rewards, where each model embodies a different hypothesis about which decision variables determine changes in attention.

3 The family of models considered

All models that we tested assume that the participant maintains and updates weights $\mathbf{w}$ associated with each of the nine features (e.g. Einstein, Big Ben, etc). On every trial, the weights of chosen features $\mathbf{w}_{chosen}$ are adjusted towards a target that differs from model to model, using an error-correcting learning rule with learning rate η as a free parameter. The weights of unchosen features $\mathbf{w}_{\neg chosen}$ are decayed towards 0 with decay rate η_k as a free parameter. And the predicted attention weights ϕ_d are determined by passing the maximal feature weights in each dimension through a softmax function with gain g as a free parameter:

$$\phi_{d,t+1} = \frac{e^{g*max(\mathbf{w}_{d,t+1})}}{\sum_{j=1}^{3} e^{g*max(\mathbf{w}_{j,t+1})}}$$

3.1 Recent choice history

The *recent choice history* model adjusts weights of chosen features towards 1, therefore keeping track of which features were consistently chosen in the past trials:

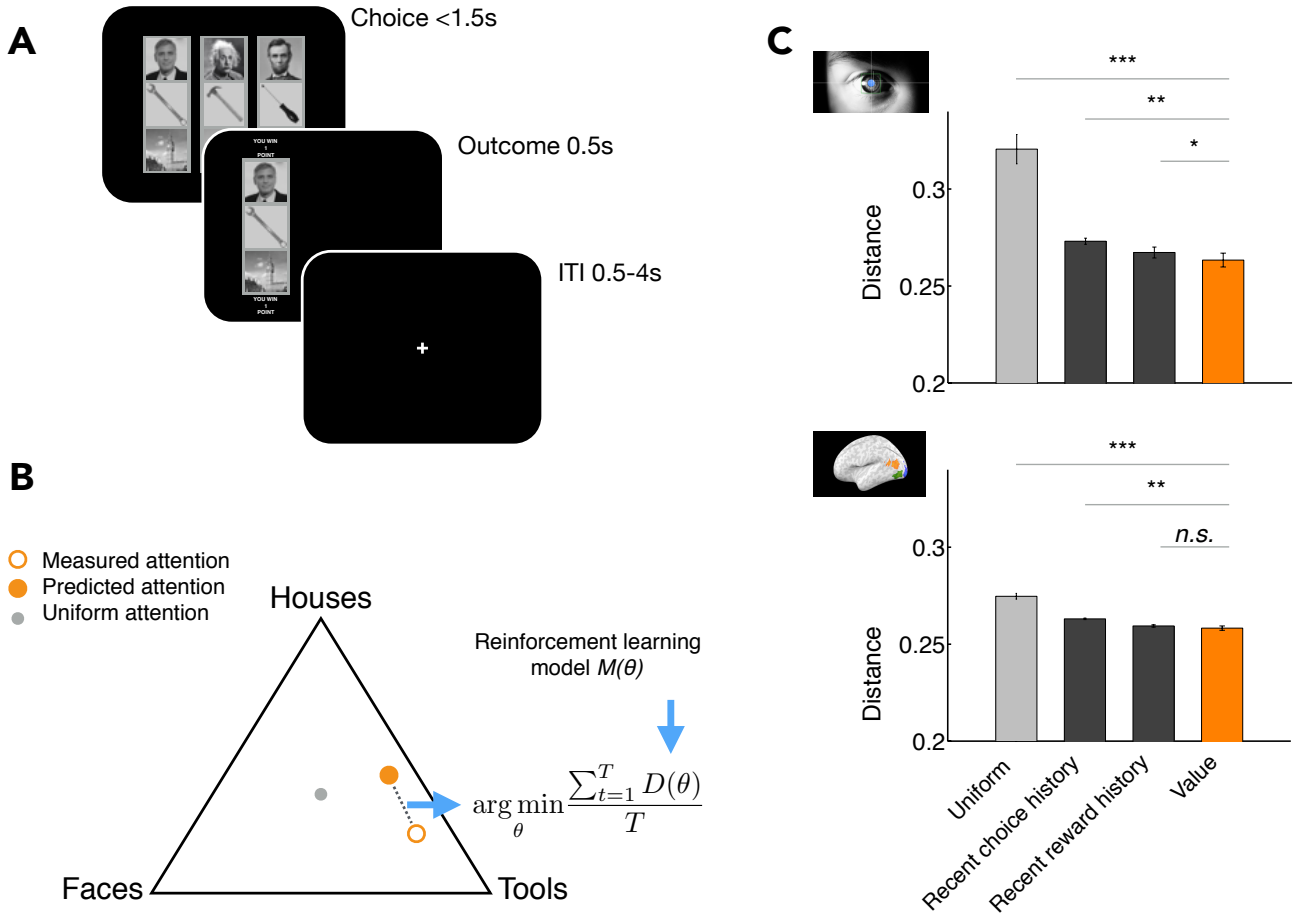

Figure 1: A. On each trial of the task, the participant was presented with three stimuli, each defined along face, landmark and tool dimensions. The participant chose one of the stimuli, received feedback and continued to the next trial in the game. B. Attention data on each trial are 3-D vectors that sum to 1, so they can be thought of as points on the 2-simplex. To compare models, we derived predictions for how attention changes from different reinforcement learning models and we used leave-one-game-out cross-validation to determine the mean distance per trial on held-out data for each model. C. Comparison of models of attention fitted separately to the eye-tracking measure (top), and the MVPA measure (bottom), according to the distance of the model's predictions from the empirical data (lower values indicate a better model). For the uniform model (light gray), we computed the average per-trial distance between the observed attention vector on each trial and [1/3 1/3 1/3]. For the remaining models, we computed the distance by repeatedly fitting the models to all games except one and testing on the holdout game. Plotted is the subject-wise average per-trial distance calculated from the holdout games. The Value model is shown in orange. Error bars indicate one SEM. ***: $p < .001$; ** $p < .01$; *$p < .05$.

$$\mathbf{w}_{chosen,t+1} = \mathbf{w}_{chosen,t} + \eta * (1 - \mathbf{w}_{chosen,t})$$

$$\mathbf{w}_{\neg chosen,t+1} = (1 - \eta_k) * (\mathbf{w}_{\neg chosen,t})$$

This model, also known as a "choice kernel", formalizes the hypothesis that attention follows choices (c.f. [13] [14]), and that choices farther in the past contribute less to what people will attend to in the future.

3.2 Recent reward history

The *recent reward history* model is identical to the *recent choice history* model above, except in that it only adjusts the weights and performs the decay when the participant receives a reward:

$$\mathbf{w}_{chosen,t+1} = \mathbf{w}_{chosen,t} + \eta * (R_t - \mathbf{w}_{chosen,t}) * R_t$$

$$\mathbf{w}_{\neg chosen,t+1} = (1 - \eta_k * R_t) * (\mathbf{w}_{\neg chosen,t})$$

This model formalizes the hypothesis that attention is dynamically modulated by recent rewards: the more consistently reward is obtained when choosing a feature in a particular dimension, the more attention will be directed towards that dimension.

3.3 Value

Finally, in the *value* model, attention tracks the within-dimension maximum of feature values, as learned through reinforcement learning with decay (see [2], where this model was developed for a different variant of the task). This model maintains value weights for each of the features, initializing them at 0 and updating them on each trial as follows: the value of the chosen stimulus is assumed to be the sum of the values of all its features; a prediction error is calculated as the difference between the obtained reward and the value of the chosen stimulus; the values of each of the chosen features are then updated based on the prediction error (with the learning rate being a free parameter), and the value of unchosen features are decayed towards 0 (with the decay rate being a free parameter).

$$\mathbf{w}_{chosen,t+1} = \mathbf{w}_{chosen,t} + \eta * (R_t - \sum_{d=1}^{3} w_{chosen,t,d})$$

$$\mathbf{w}_{\neg chosen,t+1} = (1 - \eta_k) * (\mathbf{w}_{\neg chosen,t})$$

As in the other models, the maximum value weight in each dimension is passed through a softmax function to obtain the predicted attention vector. This model therefore formalizes the hypothesis that attention follows value as learned using a simple reinforcement learning algorithm.

3.4 Model fitting and comparison

We fit the free parameters of each model for each participant separately by minimizing the distance between trial-by-trial predicted and measured attention vectors, and used leave-one-game-out cross-validation to determine the mean distance per trial on held-out data for each model. We used the root mean squared deviation (RMSD) as a distance metric. And we compared the fits to those of a zero-parameter baseline model that assumes that attention is always uniform (Fig. 1B).

4 Results and discussion

We tested whether attention allocation data could be better explained by recent choice history (i.e., attention was enhanced for features that have been previously chosen), recent reward history (i.e., attention was allocated to features that have been previously rewarded), or learned value (i.e., attention was enhanced for features associated with higher value over the course of a game). If reward history dynamically determines attention allocation in our task, we would expect the two models that include reward information to explain our data better than a model that only tracks which features the participant chose. Cross-validated model comparison revealed that attention data were best explained by a model that tracked feature values. The *value* model outperformed the *recent reward history* model for the eye-tracking measure (paired-sample t-test, t(24) = 2.77, $p < .05$, best fit for 17/25 subjects, Fig. 1C top). For the independent MVPA measure, the *value* model did not significantly improve upon the predictions of the *recent reward history* model (paired-sample t-test, t(24) = 1.02, p = 0.31, best fit for 16/25 subjects, Fig. 1C bottom). However, this model still performed significantly better than the *recent choice history* model (paired-sample t-test, t(24) = 3.83, $p < 0.001$). Our results suggest that reward history, whether counted directly or through estimation of values via reinforcement learning, determined attention allocation.

References

[1] R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, and Y. Niv, "Orbitofrontal cortex as a cognitive map of task space," *Neuron*, vol. 81, no. 2, pp. 267–279, 2014.

[2] Y. Niv, R. Daniel, A. Geana, S. J. Gershman, Y. C. Leong, A. Radulescu, and R. C. Wilson, "Reinforcement learning in multidimensional environments relies on attention mechanisms," *The Journal of Neuroscience*, vol. 35, no. 21, pp. 8145–8157, 2015.

[3] R. C. Wilson and Y. Niv, "Inferring relevance in a changing world," *Frontiers in human neuroscience*, vol. 5, 2011.

[4] F. Canas and M. Jones, "Attention and reinforcement learning: constructing representations from indirect feedback," in *Proceedings of the 32nd annual conference of the cognitive science society*, 2010.

[5] M. Jones and F. Canas, "Integrating reinforcement learning with models of representation learning," in *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pp. 1258–1263, 2010.

[6] A. K. McCallum, *Reinforcement learning with selective perception and hidden state*. PhD thesis, University of Rochester, 1996.

[7] S. J. Gershman, J. D. Cohen, and Y. Niv, "Learning to selectively attend," in *32nd Annual Conference of the Cognitive Science Society*, 2010.

[8] A. Radulescu, R. Daniel, and Y. Niv, "The effects of aging on the interaction between reinforcement learning and attention.," *Psychology and aging*, vol. 31, no. 7, p. 747, 2016.

[9] Y. C. Leong, A. Radulescu, R. Daniel, V. DeWoskin, and Y. Niv, "Dynamic interaction between reinforcement learning and attention in multidimensional environments," *Neuron*, vol. 93, no. 2, pp. 451–463, 2017.

[10] P. Dayan, S. Kakade, and P. R. Montague, "Learning and selective attention," *nature neuroscience*, vol. 3, pp. 1218–1223, 2000.

[11] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, "Beyond mind-reading: multi-voxel pattern analysis of fmri data," *Trends in cognitive sciences*, vol. 10, no. 9, pp. 424–430, 2006.

[12] N. D. Daw, "Trial-by-trial data analysis using computational models," *Decision making, affect, and learning: Attention and performance XXIII*, vol. 23, p. 1, 2011.

[13] I. Krajbich, C. Armel, and A. Rangel, "Visual fixations and the computation and comparison of value in simple choice," *Nature neuroscience*, vol. 13, no. 10, pp. 1292–1298, 2010.

[14] I. Krajbich and A. Rangel, "Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions," *Proceedings of the National Academy of Sciences*, vol. 108, no. 33, pp. 13852–13857, 2011.