RLAttn: An actor-critic model of eye movements during category learning

Jordan I. Barnes[1] **(jordanb@sfu.ca)**
[1] Simon Fraser University, Department of Psychology
8888 University Dr., Burnaby, BC V5A1S6 Canada

Caitlyn McColeman[1] **(caitlyn_mccoleman@sfu.ca)**

Ekaterina Stepanova[2] **(erstepan@sfu.ca)**
[2] Simon Fraser University, Cognitive Science
8888 University Dr., Burnaby, BC V5A1S6 Canada

Mark R. Blair[1,2] **(mblair@sfu.ca)**

R. Calen Walshe (r.c.walshe@sms.ed.ac.uk)
University of Edinburgh, Department of Psychology
8888 University Dr., Burnaby, BC V5A1S6 Canada

Abstract

Here we introduce a simple actor-critic model of eye movements during category learning that we call RLAttn (Reinforcement Learning of Attention). RLAttn stores the rewards it receives for making decisions or performing actions, while attempting to associate stimuli with particular categories. Over multiple trials, RLAttn learns that a large reward is most likely when the values of the relevant stimulus features have been revealed by fixations to them. The model is able to approximate human learning curves in a common category structure while generating fixation patterns similar to those found in human eye tracking data. We additionally observed that the model reduces its fixation counts to irrelevant features over the course of learning. We conclude with a discussion on the effective role eye movements might play in bridging structural credit assignment and temporal credit assignment problems.

Keywords: Reinforcement learning; category learning; computational cognitive modeling; eye tracking; actor-critic; attention

Introduction

Researchers have known for decades that appropriate selective attention is needed to facilitate learning (Shepard, Hovland & Jenkins, 1967) due in part to evidence showing that deficits in attention impair learning (Filoteo, Maddox, Ing & Song, 2007). However, we are only just beginning to understand how selective attention itself is learned. Gottlieb (2012) has advanced the thinking about this issue and has outlined some of the important interactions between learning, attention and eye movements. This work strongly motivates thinking of eye movements as both aiding in learning, as well as implementing actions, by virtue of the rewards obtained by them. In general, the process of discovering optimal behaviours, given particular rewards, is known as reinforcement learning (Sutton & Barto, 1998), and a number of different methods have been developed that properly apportion reward given a sequence of behaviours.

Of the several different kinds of reinforcement learning approaches, one in particular, the actor-critic method, has stood out as having a plausible mapping on to the mid-brain mesencephalic dopaminergic system (Holroyd & Coles, 2002). In this account, a recurrent loop between the anterior cingulate cortex and the basal ganglia works to produce action signals that are then critiqued based on differences in expected versus acquired reward. With uniform expectations, those differences are merely proportional to the size of the reward. As experience reveals the utility of particular actions, any differences in expectations modify the size of future weight adjustments; should something good come along when something very bad was expected, this difference would have a larger reinforcing effect on the decision taken to get it than a more expected result. To reinforce the sequence of actions, a small amount of that reward, known as the temporal difference error, is passed back to the preceding actions that got to the present decision. This is done not in one step but as a function of what the system can expect by taking that particular action the next time it finds itself in that particular state. Once the chain is in place, following the sequence of actions simply collects the expected reward.

While there is considerable evidence supporting the notion that eye movements are sensitive to rewards (Hikosaka, Sakamoto & Usui, 1989), models of categorization and attention have not comprehensively investigated the implications of this. Most of the learning models built to explore various category structures employ methods meant to optimally assign blame for classification errors based on physical features of the input without regard for the temporal nature of information acquisition.

In this paper we introduce an actor-critic reinforcement learning model of eye-movements in the context of category learning that we call RLAttn (Reinforcement Learning of Attention). We demonstrate that the model qualitatively mimics several aspects of human eye-movement data, including the overall number of fixations and the relative number of fixations to relevant and irrelevant features. Finally, we consider the similarities and differences between RLattn and existing reinforcement learning models, each of which, including RLAttn, has its own strengths and weaknesses.

Structure of the model

In reinforcement learning, the agent improves its performance by interacting with the environment in order to achieve a goal. The agent is the learner, and the environment is the set of all possible actions or interactions that the agent faces. The actor-critic method of reinforcement learning that RLattn uses is based on the method of temporal differences (Sutton & Barto, 1998). This is a class of dynamic programming methods that breaks up a problem into a set of possible states (S) with associated actions (A) leading from one state into another ideally with optimal transition probabilities, without storing the entire history. The agent acts in the environment in discrete time steps. During each time step (t), the agent receives some information about the state of the environment.

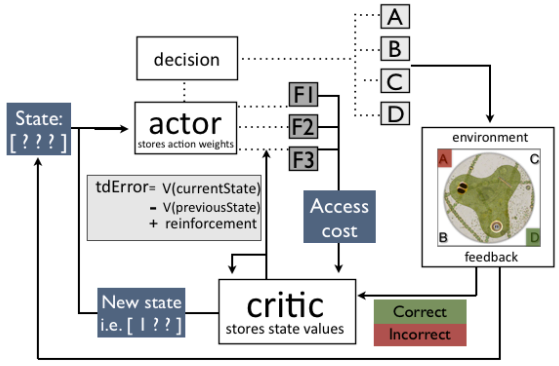

Figure 1: Starting a trial from the state (k=27) of unknown feature values [? ? ?], RLAttn selects an action, collects the associated rewards or punishments and transitions to a different state.

The possible states in the category learning environments in the present investigation are based on three stimulus features and a category decision action. All three features can hold one of two possible values on any given trial: either 1 or 2. Additionally, each of the three features can be unknown, which we coded as 3. As such, there are three possible states resulting from each of the three features: known (1 or 2) or unknown (3). Additionally, the three possible feature states over the three different features yield $3^3 = 27$ possible combinations, and the addition of a category decision state make a total of S=28 states, $k \in S$. The decision state is one in which the agent makes a category choice. As an example, state k=27 is the starting state on each trial and represents the state of not knowing anything about the features. It encodes the state of knowledge [3 3 3], which is to say that none of the three features have been fixated. From this state, the agent might decide to look at feature 1. Having fixated the feature, the agent is in one of two possible initial "feature 1 known" states, which, depending on the value of that feature, might move it to state k=25, (1 3 3), or state k=26, (2 3 3). The values of features 2 and 3 are both coded to 3 because they are still unknown after the first fixation.

RLAttn selects an action based on the action probabilities in its current state, where $A=Q_{k,t}$ The possible actions for the agent in our environment are to fixate one of the three features, or to make a category decision. Upon selecting an action, the agent is brought into a different knowledge state, unless opting to fixate the same feature again. For that particular time step t, the agent is given a reward r as a consequence of its chosen action.

RLAttn's environment has three possible sources of reward: access cost, and correct or incorrect decision rewards. In modelling human gaze data, access cost is a punishment that has been used to represent the bio-mechanical energy cost of making an eye movement (Nelson & Cottrell, 2007). Access costs, in the form of temporal delays, have been experimentally shown to influence patterns of human eye movements in category learning tasks (Meier & Blair 2013; Wood, Fry & Blair, 2010). In RLAttn, access costs are a small penalty for each eye-movement. Correct and incorrect rewards are relatively larger rewards/punishments that are collected as function of whether or not the model made the correct decision. The record of the reward for each action a in each state is stored in a Q-matrix:

$$Q_{k,t}(a) = Q_{k,t}(a) + \alpha\delta_t \qquad (1)$$

where δ_t is the temporal difference error calculated by differences in the state value record V_k. and α is a learning rate. The Q-matrix has A columns and S rows; in our case 4 (one for each of the three features, and one for making a category decision) columns and 28 rows. The final, 28th row, contains the set of possible category choices, which also happens to be A=4 in this particular case (one action a for each of the four possible categories shown in Table 1). The reward, r_t, for a decision action, $a \in A$, is stored here. The relationship between the state $Q_{k,t}$ and the action a is called the policy. The policy is a mapping from each of the states and the selection of a possible action based on a set of corresponding probabilities. Typically a policy might be set to greedy, i.e. select the most reinforced action, or be varying probabilistic to explore the space.

Over time, the agent learns that it is more preferable to be in some states as opposed to others. This is controlled

by the value vector V. This value vector has one column which stores the value of the full state, and so contains S rows. In practice, the value of a state under a policy is the expected return starting from state *k* and following the policy to select an action, which in our case is the Luce decision rule. The values within V are updated by:

$$\delta_t = V_{k+1} + r_{t+1} - c - V_k \quad (2)$$

where V_k is the value of the current state, r_{t+1} is the reward, modulated by an access cost *c*, and V_{k+1} is the value of the state that the agent is in after its action.

$$V_k = V_k + \alpha\delta_t \quad (3)$$

Given this formalization, it still has to be decided how the action probabilities for particular unvisited states in the Q matrix are initialized. We currently opt for simple generalization rules in RLAttn. If a particular knowledge state *k* has never been visited, all actions are equiprobable, however the decision action is defined by $V_{k=28}$. When initializing the category selection probabilities from state $Q_{k=28,t}$ the average of all $Q_{k,T}$ is taken.

The action probabilities are transformed by a modified Luce decision rule, such that:

$$p(a, Q_{k,t}) = \frac{e^{Q_{k,t}(a)/\tau}}{\sum_{b \in A} e^{Q_{k,t}(b)/\tau}} \quad (4)$$

where $Q_{k,t}$ is the agent's current state, *e* is Euler's constant, *a* is the action whose odds of selection are being transformed, *b* is a member all possible actions for that state, and τ is a temperature constant, set to 1 for RLAttn. Over multiple time steps, the probability of selecting the action with the highest reward is higher than selection all of the other possible actions. In general, the agent's goal is to maximize the reward that it receives overall - not just the immediate reward. The Luce decision rule acts as the policy for the model, and is among the simplest strategies for defining a policy (Sutton & Barto, 1998).

Human Data

In order to assess the performance of RLAttn we fit data (Figure 2) from McColeman, Barnes, Chen, Meier, Walshe and Blair (2014). In this study, participants had to learn to sort images of fictitious micro-organisms into four possible categories. The images each contain three spatially separated features, and the values of two of the features indicate to which category the image belongs. The remaining feature is irrelevant (Table 1). The data comes from 19 learners with high quality gaze data in the perfect feedback condition of an experiment manipulating feedback validity. All participants come from Simon Fraser University's Research Participation pool, and everyone received partial course credit for their participation.

Eye tracking data were converted into fixations using a modified version of the Salvucci-Goldberg dispersion algorithm (2000). Additional methodological details are available along with the probability of fixating the irrelevant feature, fixation durations, attention change and error bias in McColeman, Barnes, Chen, Meier, Walshe & Blair (2014). These data are available on the Simon Fraser University Summit Repository system[1].

Table 1: Category Structure.

Feature 1	Feature 2	Feature 3	Category
0	0	0/1	A
0	1	0/1	B
1	0	0/1	C
1	1	0/1	D

Accuracy

As can be seen in Figure 1, people quickly learn the category structure and maintain high accuracy for the duration of the experiment. Trials prior to achieving 9 correct answers consecutively are deemed pre-learning, and those after 9 correct answers in a row, the criterion point, are post-learning.

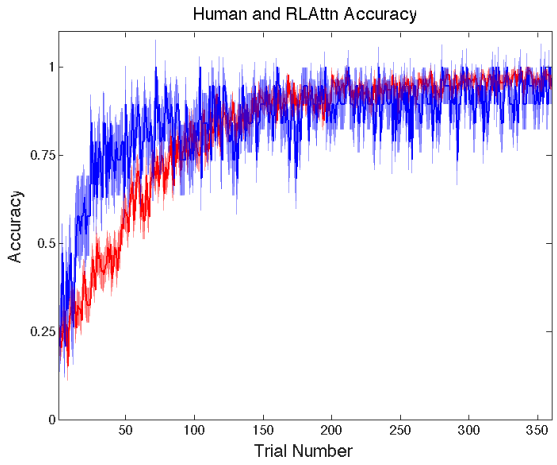

Figure 1: Accuracy is shown with standard error represented as the shaded region around the mean accuracy line.

Fixation count

As with the participants' accuracy, the fixation counts were also reported in McColeman, Barnes, Chen, Meier, Walshe & Blair (2014). Figure 2 depicts an example of one kind of attentional optimization, whereby participants reduce the overall number of fixations they make over the course of the experiment.

[1] http://summit.sfu.ca/item/12720

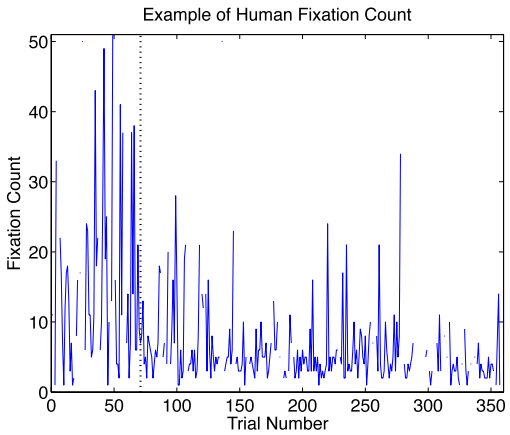

Figure 2: Fixation counts relative to the criterion point, shown as a dashed vertical black line, in the participant most similar to the mean criterion point (in this case trial 71 compared to the mean criterion point of all participants of trial 75). Trials with gaze quality are not included, and are visible as gaps in the plot.

Model Data

The model was fit by minimizing the difference between its accuracy and the human subject accuracy (see Figure 1) on a trial-by-trial basis. The fitting procedure was initiated with a simple grid search of several different levels of the two free parameters, learning rate, and reward. The best fitting parameters were chosen as the seeds for a simplex based minimization method implemented in MATLAB, named fminsearch (Lagarias, Reeds, Wright & Wright, 1998). Because RLAttn is a stochastic model we ran the model 3 times before returning the average fit value from these runs to fminsearch. Once the best fitting parameters were found, we ran the model another 5 times under these parameters to generate 5 simulations for each of our 19 human subjects upon which to base our analysis. Occasionally RLAttn would enter a pathological state, such as endlessly fixating features without making a decision. If such a state was entered using a set of best fitted parameters, the simulation was dropped. In this case we lost 3 simulations leaving 92 simulations. Despite a particularly conservative fit calculation, where we calculate the match in accuracy on each trial between the human and the model data, RLAttn matched of 81% (SD = 0.12) of the total trials.

Accuracy

Overall RLAttn matches the qualitative features of human learning. As can be seen in Figure 1, RLAttn took slightly longer to attain the same level of accuracy as humans, likely due to the level of randomness in the Luce decision rule, governed by τ, which was not fit as a free parameter in this instantiation of RLAttn. Humans may also be generalizing their category knowledge more efficiently than the model.

Fixation count

As with the human data, the mean fixation counts decrease over the course of the experiment (Figure 4.).

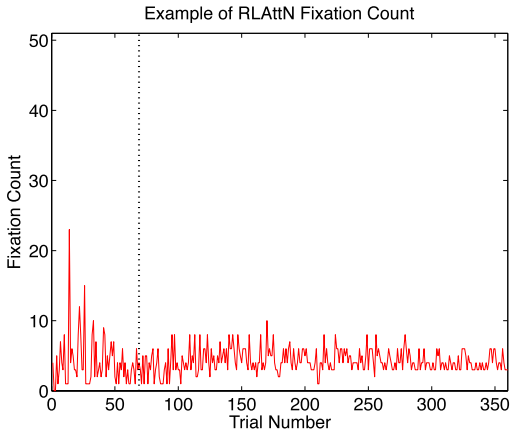

Figure 3: Fixation counts relative to the criterion point, shown as a dashed vertical black line, in the best fitted model for the same subject as that in Figure 2, with criterion point most similar to that individual from the simulation distribution (trial 76 for this simulation). Although the trials with fixation counts >20 are less in the model for this simulation (fixation counts were not fit directly), the means are very similar ($\mu_{Subject}$= 3.86 and μ_{RLattn} = 3.46).

In line with previous attentional efficiency results (McColeman et al, 2014), the decrease is disproportionately associated with the irrelevant feature. As the values of particular eye movement decisions are refined, access cost begins to outweigh the expected reward from looking at irrelevant information.

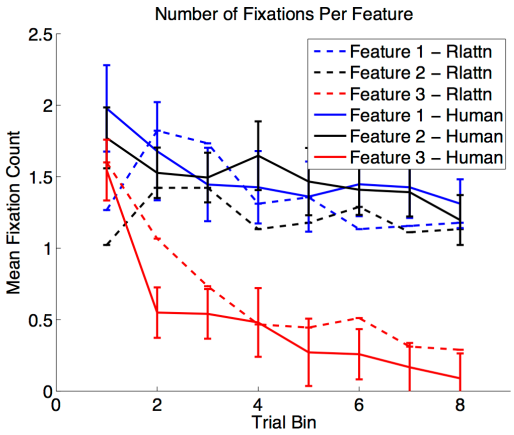

Figure 4: The mean fixation counts to each feature are comparable between the model and the human participants over the course of the experiment. Further, RLAttn is exhibiting the right kind of reduction in fixation count as it reduces its fixations to the irrelevant feature.

Model Comparison

Here we compare RLattN with two related models that attempt to address eye movements and attention during category learning. The first is a Bayesian model developed by Nelson and Cottrell (2007) which links the probability of making a query to a stimulus dimension to the probability that a particular dimension will improve a category decision. This estimation is added to the mutual information between that stimulus dimension and the probability of a correct category determination, while the energy cost of making an eye movement is subtracted. If the mutual information between a dimension and the correct category decision is low, a function of the feature being irrelevant, than the cost of making a movement will outweigh the expected information gain of looking at that feature and no movement will be made. In this sense, the model is an attempt to directly account for attentional optimization results which show reduced fixations over the course of learning to irrelevant dimensions. Apart from being easily interpretable, the use of mutual information as a metric for deciding an action is supported by research showing that human behaviours are often taken to reduce uncertainty within a task (Renninger, Verghese & Coughlan, 2007).

Although RLAttn employs the idea of access cost and can be interpreted as being influenced by probability gain, there are a number of interesting differences in model behaviours and assumptions. For instance, Nelson and Cottrell note that human learners often do not query all stimulus dimensions even prior to understanding the category structure (Rehder & Hoffman, 2005). In RLAttn this kind of counter-intuitive behaviour is seen as a direct consequence of the agent not knowing the relative values of looking at information as opposed to making category decisions. Not until negative rewards have had a chance to discourage this kind of ignorant decision making will the participant settle in to a more consistent and useful fixation pattern; this could be thought of as a simple rule testing mechanism but more empirical research on this question would be needed. Further, the inclination to make a particular motor movement is not simply guided by probability gain concerns (Meier & Blair, 2013) but also by the reinforcement history (Holroyd & Coles, 2008). Overall, we see the use of mutual information as an important consideration to a more comprehensive model of eye movements and see RLattN as addressing a separate set of concerns pertaining to the reward responsive nature of the saccadic system.

One of the most influential models of attentional learning in categorization is ALCOVE (Kruschke, 1992). In ALCOVE, attentional weight to particular stimulus dimensions (represented in the present work as features) is tuned by errors during learning. Whereas ALCOVE was developed to model human learning based on a psychological reworking of the back-propagation algorithm, Jones and Cañas (2010) extended these ideas in Q-ALCOVE using reinforcement learning principles. They did this by using temporal-difference error to solve what would normally be a structural credit assignment problem but posed in such a way as to allow a temporal credit assignment solution. To understand this, consider the categorization problem they develop. The stimuli they use have two parts, presented sequentially, where the action taken by the participant when viewing the first part modifies the available actions in the second part, thereby affecting the elicited reward. Learning as a function of temporal difference error would predict that higher valued actions taken in the first part of stimulus viewing would be reinforced by the later reward and, perhaps unsurprisingly, this is what was found in human participants doing the same task (Cañas & Jones, 2010).

The insightful part about this work is the way in which it uses the structure of the task to separate stimulus dimensions in time to circumvent what could easily have been a contemporaneous conjunctive rule problem, thus allowing the temporal difference error to determine the correct structural relationships. There is very clearly an important intuition about human learning being developed in this model but it is never quite explicitly stated: all structural problems have temporal contingencies as a consequence of serial selective attention. The primary difference between RLAttn and Q-ALCOVE is that we see the embodied sequential actions of the eyes as the primary conduit of these contingencies. Either way, this contribution to the literature deserves to be recognized as an advancement of category learning models towards the general fact that any category decision is the result of a serious of previous sub-decisions that share in rewards.

Discussion

We have presented an actor-critic reinforcement learning model of eye movements during category learning that is able to approximate human learning curves while also improving their attentional efficiency based on previous rewards. In addition to these qualities, the model offers a reason for the seemingly odd behaviour of participants to guess categories without fully exploring a stimulus, as had been previously reported: until punished, decision actions in a particular state are actions like any other. Thinking of behaviour in this way allows us to understand why people are sometimes prone to making objectively non-optimal decisions. We see the creation of models like RLAttn as a useful first step towards bridging neurophysiological research on reward processing, particularly with respect to eye movements and attention, with the well-studied category structures used in the category learning literature. To our knowledge, only a few models have been presented that have attempted to address these issues (see Barnes, Walshe, Blair & Tupper, 2013, in addition to the models looked at here) which is surprising given the longstanding interest in psychology in category learning and computational modeling.

Finally, the relationships between category learning and sequencing behaviours are important to understand for a

number of reasons. Not only do humans solve both structural and temporal credit assignment problems (the difference between what causes something and when should something be done) but deficits in areas implicated with reinforcement learning, like the basal ganglia, are observed to impair both category learning and sequencing behaviours (Seger, 2006). The solution provided here, which is to store information needed to solve classification problems over multiple, serially accessed states, fits will with the 'just-in-time' representations posited by Ballard and colleagues (1997). That is to say that working memory resource constraints point to the need to dynamically retrieve information as it is needed and actions like eye movements offer a method of pointing to the information needed. It seems plausible to think that computer models that learn category structures based on these principles may one day contribute solutions to broader problems in cognitive science and psychology.

Acknowledgments

We would like to thank the members of the Cognitive Science Lab at SFU and our funding bodies: the National Science and Engineering Research Council (NSERC), and the Canadian Foundation for Innovation (CFI).

References

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *The Behavioral and brain sciences*, *20*(4), 723–42.

Barnes, J.I., Walshe, R.C., Tupper, P.F., & Blair, M.R. (2013) A dynamic neural field model of eye movements during category learning tasks. *Learning to Attend, Attending to Learn: Neurological, Behavioural, and Computational Perspectives*.

Cañas, F., & Jones, M. (2010). Attention and reinforcement learning: Constructing representations from indirect feedback. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Filoteo, J. V., Maddox, W. T., Ing, A. D., & Song, D. D. (2007). Characterizing rule-based category learning deficits in patients with Parkinson's disease. *Neuropsychologia, 45*(2), 305–20.

Gottlieb, J. (2012). Attention, Learning, and the Value of Information. *Neuron,* 76, 281-295.

Hikosaka, O., Sakamoto, M., & Usui, S. (1989) Functional properties of monkey caudate neurons. I. Activities related to saccadic eye movements. *Journal of Neurophysiology*, *61*, 780-798.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & psychophysics*, 57(6), 787–95.

Holroyd, C. B., & Coles, M. G. H. (2002). The Neural Basis of Human Error Processing: Reinforcement Learning, Dopamine, and the Error-Related Negativity. *Psychological Review, 109*(4), 679–709.

Holroyd, C. B., & Coles, M. G. H. (2008). Dorsal anterior cingulate cortex integrates reinforcement history to guide voluntary behavior. *Cortex. 44*(5), 548–59.

Jones, M. & Cañas, F. (2010). Integrating Reinforcement Learning with Models of Representation Learning *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Kruschke, J K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22–44.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E (1998), Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions. *SIAM Journal of Optimization*, *9*(1), 112-147.

McColeman, C. M., Barnes, J., Chen, L., Meier, K., Walshe, R. C., & Blair, M. (2014). Learning-induced changes in attentional allocation during categorization: a sizable catalog of attention change as measured by eye movements. *PLoS ONE 1*(9).

Meier, K. M., & Blair, M. R. (2013). Waiting and weighting: Information sampling is a balance between efficiency and error-reduction. *Cognition*, *126*(2), 319–25.

Nelson, J. D., & Cottrell, G. W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, *70*(13-15), 2256–2272.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive psychology*, *51*(1), 1–41.

Renninger, L. W., Verghese, P. & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. Journal of Vision. 7(3), 1–17.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. *Proceedings of the symposium on Eye tracking research & applications*, 71–78.

Seger, C. (2006). The basal ganglia in human learning. The *Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, *12*(4), 285–90. doi:10.1177/1073858405285632

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Wood, M.J., Fry., M., & Blair, M.R. (2010). The price is right: A high information access cost facilitates category learning. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*. 236-241.