An Integrated Model of Associative and Reinforcement Learning

Vladislav D. Veksler (vdv718@gmail.com)
Christopher W. Myers (christopher.myers.29@us.af.mil)
Kevin A. Gluck (kevin.gluck@wpafb.af.mil)
Air Force Research Laboratory
Wright-Patterson AFB, USA

Abstract

Any successful attempt at explaining and replicating the complexity and generality of human and animal learning will require the integration of a variety of learning mechanisms. Here we introduce a computational model which integrates associative learning and reinforcement learning. We contrast the integrated model with associative learning and reinforcement learning models in two simulation studies. The first simulation demonstrates performance advantages for the integrated model in an environment with a dynamic and diverse reward structure. The second simulation contrasts the performances of the three models in a classic latent learning experiment (Blodgett, 1929), demonstrating advantages for the integrated model in predicting and explaining the behavioral data.

Keywords: Associative Learning, Reinforcement Learning, Model Integration, Cognitive Modeling, Cognitive Systems, Latent Learning

Introduction

Integration of computational cognitive models is critical for accelerating progress in the field of cognitive modeling (Gray, 2007a). By means of integrative approaches the field can begin to predict and explain the robustness and flexibility of human behavior in complex, uncertain, non-stationary environments (or *large worlds*, Binmore, 2009). Specifically, Choi & Ohlsson (2011) assert that the integration of *learning mechanisms* is essential to improving the predictability and explanatory power of cognitive models.

There is no denying that people are adaptive – for example our memories are retrieved based on their recency and frequency of use (Anderson & Schooler, 1991), visual search is adapted to the structure of the the task environment (Shen, Reingold, & Pomplun, 2000; Myers & Gray, 2010), and problem solving strategies are adapted with increasing task experience (Siegler & Stern, 1998). People's ability to adapt allows them to persist and thrive in large worlds. If we are to build cognitive models for large worlds, we have to endow them with human learning mechanisms. Hand-coded knowledge engineering results in brittle and expensive models, and is a method that does not scale well beyond simple laboratory environments (Gluck, 2010). Our hypothesis is that models may begin to demonstrate human-like flexibility and adaptivity in large worlds through the integration of multiple human learning mechanisms.

In the current paper we present an integrated model of associative and reinforcement learning, as it is evident that humans are capable of learning both the spatiotemporal contingencies and the reward structures of their environment (Stevenson, 1954; Chun, 2000; Myers, Gray, & Sims, 2012). We demonstrate that model integration improves flexibility and adaptability, provides better predictions of behavioral data, and produces more efficient behavior in environments with diverse and dynamic reward structures when compared to each of the individual models.

In the following sections we first provide background on associative and reinforcement learning theories and models. Next we describe the integrated model. Finally, two simulations are presented. Simulation 1 contrasts the associative and reinforcement learning models with the integrated model in their ability to efficiently adjust to novel goals and diverse reward structures in a grid-navigation environment. Simulation 2 contrasts the associative and reinforcement learning models with the integrated model in their ability to predict behavioral data from a classic latent learning experiment.

Reinforcement Learning

Reinforcement learning (RL) is a formal model of action selection where the utility of different actions is learned by attending to the reward structure of the environment. It has been used in a wide array of domains, from robotics (Peters, Vijayakumar, & Schaal, 2003) and artificial intelligence (Russell & Norvig, 1995) to cognitive architectures (Fu & Anderson, 2006; Nason & Laird, 2005) and cognitive neuroscience (Holroyd & Coles, 2002).

Generally speaking, RL works in a trial-and-error fashion – attempting various actions and recording the reward gained for those actions (for a review see Sutton & Barto, 1998). More formally, given the state that an agent is experiencing, the action most likely to be chosen is the one with the highest learned utility, plus or minus some exploratory noise. The utility of any given state-action pair, *SA*, in turn, is directly proportional to the value of the reward, that the agent receives after *SA* is executed. Hence, state-action pairs are *reinforced* when they result in a reward; and the likelihoods of their future selection are directly proportional to the values of the experienced rewards.

There are several variations on how utility is learned in RL (for an introduction, see Sutton & Barto, 1998). For example, Temporal Difference RL (TDRL), a version commonly used to model human behavior (Anderson, 2007; Holroyd & Coles, 2002), propagates the received reward to past actions. Reward is discounted as a function of time, so that actions taken just prior to the reward are strengthened more than earlier actions. In this way, TDRL reinforces a sequence of actions that lead to the reward, rather than just a single state-

action pair, helping to obtain a solution in a more efficient manner.

Some RL approaches take into account transitions between *SA*, the resultant reward *R*, and the utility of the next *SA* (SARSA). SARSA models update the utility of the state-action pair executed at time t, $SA(t)$, by a function of the reward that follows it, $R(t+1)$, combined with a function of the utility of the state-action pair that follows it, $SA(t+1)$. SARSA models are not as efficient as TDRL, but are guaranteed to converge on an optimal solution.

The Model-based RL approach extends RL by learning the structure of the world beyond utilities. The term *model* in "model-based RL" refers to an agent's internal representation of the environment, and an agent developed in this framework is capable of planning its route before execution. This is extremely useful when memory and decision cycles are less expensive than actions (e.g. robotics).

One of the limitations of RL as a complete model of human decision-making becomes apparent in environments where goals change. Imagine that on your way to work each day you pass a post office. One day you need to mail a letter. At this point, an RL agent would consider, "let's try a random action, see how that works." This is because, by definition, RL models make decisions based solely on the learned state-action utilities. If the goal changes, the utilities representing the reward structure from the initial goal become irrelevant at best, or subversive at worst. Humans and animals, of course, will employ their knowledge of the environment (e.g. that there is a post office on the way to work) to make better-than-chance decisions for achieving new goals (Stevenson, 1954; Tolman, 1948; Quartermain & Scott, 1960).

The SARSA and Model-based approaches are major steps toward more flexible behavior. The SARSA approach considers the state-action-state transitions when learning utilities, but stops short of learning these transitions. The Model-based RL approach learns such transitions, but employs them strictly to enable planning. The decision process during the planning stage, however, is still based on the learned utilities. Thus, when presented with a new goal a Model-based RL agent will still begin to plan its route by considering random actions.

Associative Learning

Another class of decision models relies on associative learning. Associative learning (AL) models focus on acquiring the spatiotemporal contingencies of the environment and employing these in action-selection. The utility of any given choice is estimated as a function of previously experienced spatiotemporal proximity between this choice and the current goal. The advantage of this approach over RL is that the stored knowledge is goal-independent. Whenever a new goal is given, an AL model can employ its knowledge to make informed goal-directed decisions.

Voicu and Schmajuk (2002) implemented a computational model that learns the structure of the environment as a network of adjacent cells. Once a goal is introduced, reward signal spreads from the goal-cell through this network, such that the cells farther from the goal-cell receive less activation than those that are close. Goal-driven behavior in this model comprises moving towards the cells with the highest activation. Once this model memorizes the map of the environment, it does not need to learn the reward structure through trial-and-error; rather, the utility of each action-path is identified through spreading activation from the goal.

SNIF-ACT (Fu & Pirolli, 2007) is another model that employs associative rather than reward knowledge for action-selection. SNIF-ACT is a model of human information-seeking behavior on the World Wide Web. The World Wide Web is unpredictable in the sense that there is no way for any of its users to know what links they will encounter during web browsing. The utility of selecting a link in SNIF-ACT is not based on any prior reward, but rather on the semantic association of a link's text to the current goal (i.e., *information scent*). This mechanism allows SNIF-ACT to make non-random decisions in novel situations based on associative knowledge.

A limitation of SNIF-ACT is that it does not learn the association strengths between links and goals, but rather imports these values from an external source. The Voicu & Schmajuk model learns association strengths in a psychologically implausible manner. The Goal-Proximity Decision-making model (GPD; Veksler, Gray, & Schoelles, 2009) mends this by employing the psychologically-plausible delta learning rule (Rescorla & Wagner, 1972; Widrow & Hoff, 1960) to update association strengths. Like the other two models, GPD then estimates the utility of a path based on its association strength to the current goal. Veksler, Gray, & Schoelles demonstrate that in an environment where goals continue to change, GPD is able to replicate human performance and RL cannot.

A limitation of AL models is that no reward information is learned. In this class of models decisions are based on explicitly specified goals. Associative learning does not help to understand a diverse reward structure, where some actions may result in less reward and some in greater reward. Hence, AL models cannot explain why an organism might learn to prefer actions leading to one goal-state over another.

Integrating Associative and Reinforcement Learning

It is our opinion that AL and RL complement each other. As discussed above, RL models capture behavior based on a given reward structure. However, as agent goals change, so does the reward structure of the environment. Since RL fails to capture environmental contingencies beyond the original reward structure, it cannot predict the efficiency of human behavior in environments where goals tend to change. Contrariwise, AL models store the the spatiotemporal contingencies of the environment independent of the reward structure, and are more flexible in adapting to new goals. However, in ignoring the reward information in the environment, association-based models cannot capture people's sensitivity to the value of reward. In this section we describe how the two learning approaches can be integrated to produce more flexible behavior in environments where the reward structure is both diverse and dynamic.

Given some agent state, *S*, and a possible action, *A*, RL models learn the utility, *u*, of the *SA* state-action pair as directly proportional to the reward that has been experienced after prior executions of *SA*, and, in models like TDRL, inversely proportional to the length of time between *SA* and the

reward in prior experience. AL models do not learn the utility of *SA*, but estimate it based on the strength of association, *w*, between *SA* and the current goal, *G*, where *w* is inversely proportional to the length of time (or distance) between *SA* and *G* in prior experience. From the perspective of what is stored in model memory, the RL models store the values of *u* for each state-action pair, SA*u*, and AL models store the values of *w* for each state-action-state transition, SA*w*S.

To integrate these two models, we propose that the association strength, *w* should continue to be recorded as in the AL models, whereas the utility *u* should be recorded for each state, *S*, rather than for each state-action pair *SA*. Thus, what will be stored in memory and used for action-selection in the integrated model is both *w* and *u* for each state-action-state transition, SA*w*S*u*. The strength of association, *w*, is useful as an estimate of the probability that a state might follow a given state-action pair and the length of time of this transition. The utility, *u*, is useful as an estimate of the reward probability/value to be received after a transition occurs.

The integrated model uses the delta learning rule to update both utilities and association strengths. For each previously executed state-action pair j and each new state i, the strength of association between j and i, w_{ji}, at current time, n, is increased in the following manner:

$$\Delta w_{ji}(n) = \beta[a_i(n) - w_{ji}(n-1)] \quad (1)$$

where β is the learning rate parameter, and a_i is the activation of i ($a_i = 1$ if i is present, else 0). The utility for each new state i, u_i, at current time, n, is increased in the following manner:

$$\Delta u_i(n) = \alpha[r(n) - u_i(n-1)] \quad (2)$$

where α is the learning rate parameter, and $r(n)$ is the reward experienced at time n.

At each decision point, the utility of a given state-action pair, j, is calculated as follows:

$$U_j = \sum_{\forall i}(w_{ji} \times u_i \times \delta^t) + N \quad (3)$$

where δ is a discount parameter ($0 < \delta < 1$), t is the temporal distance between j and i, and N (exploratory noise) is a number drawn randomly from a normal distribution with a mean of zero and a standard deviation set to some parameter, σ.

In the following section the SA*w*S*u* model is examined in terms of efficiency and psychological validity within environments with diverse and dynamic reward structures.

Simulations

The following subsections compare the integrated AL+RL model (SA*w*S*u*) with AL-only and RL-only models. First, the models are evaluated based on the efficiency of finding rewarding states in a 10×10 grid. Second, the models are evaluated based on the ability to match data from a classic latent learning experiment. A single value for the discount parameter ($\varepsilon = .9$) was used for both simulations.

Simulation 1: Dynamic & Diverse Reward Structure

As pointed out in the Introduction, the strength of RL is in learning a diverse reward structure, where some actions may lead to greater reward than others; AL excels at learning the environmental structure independent of rewards, such that this knowledge may be applied in purposive behavior whenever new goals arise. However, large worlds are both diverse and dynamic. The following simulation was conducted to *highlight the conditions under which the RL and AL approaches begin to falter*, and how an integrated approach addresses these limitations.

A 10×10 navigation grid was used, where a model's state was uniquely identified as one of the cells in the grid, and the model had four possible actions from each cell – to move north, south, east, or west[1]. If an illegal move was selected (i.e. a move that would take the model off the grid), the model's state was not changed. For each model run, a model was placed in a random cell on the grid. Each time the model reached a reward state, the model would again be placed in a random cell on the grid. Before the model began the task of locating a reward, a reward of 1.0 was placed in a random cell on the grid. After 4000 steps, the reward was cleared, and placed in a different random cell. Following the next 4000 steps (8,000 total steps), the reward was cleared, and rewards of 1.0 and 0.1 were placed in two randomly selected cells. Finally, after the next 4000 steps (12,000 total steps), the reward was cleared again, and replaced with rewards of 1.0 and 0.1 in two randomly selected cells.

The integrated AL+RL model (SA*w*S*u*) was compared with Random-walk, AL (GPD), and two RL models – Temporal-Difference RL (TDRL), and Q-learning (Q-RL). As discussed above, TDRL is the version of RL most commonly used in modeling human/animal behavior. Q-RL is a popular SARSA model that is not as efficient as TDRL, but is guaranteed to converge on an optimal solution (Sutton & Barto, 1998). The results, averaged over 100 model runs for each model type, may be observed Figure 1.

AL+RL was the best overall model, averaging a total score of 1328.4 for the entire run in this environment, whereas AL (GPD), TDRL, Q-RL, and Random models scored 1107.8, 362.2, 237.1, and 66.1, respectively. Q-RL is guaranteed to converge to an optimal solution for any one reward structure in the environment, but it is too inefficient to find such a solution within the 4000 trials allotted in this task (though its efficiency improves after the first 8000 steps, where there is more than one goal-state).

TDRL, AL, and AL+RL produce indistinguishable performance until the first goal change (see Figure 1, first 4000 steps). However, once a new goal is presented, TDRL struggles to relearn the reward-structure of the environment, as all of the state-action utility values need to be relearned (these may be relearned faster if the exploratory noise was increased, but this would come at the expense of performance, even for the first goal). Q-RL struggles with the same issue, as both RL models drop to random-level performance once a new reward-structure is introduced. In contrast, AL and

[1]This is a standard simulation environment for performance examination of computational agents, as it aims to represent a generic problem space.

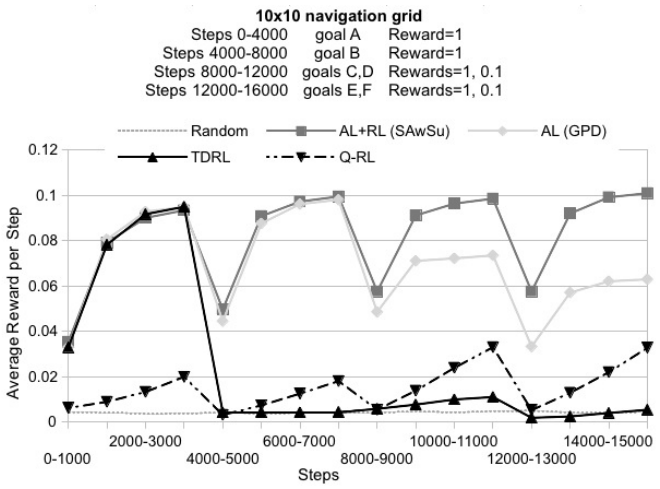

Figure 1: Simulation 1 results.

AL+RL can employ all of the associative knowledge that was gathered in the first 4000 steps, and apply it to achieving the new goal.

Where AL+RL begins to differ from AL is when the reward structure of the environment becomes more diverse. After 8000 steps, there are two rewarding states introduced into the environment, one of these having a high value (1.0) and the other having a low value (0.1). The AL+RL model learns the correct reward values of these states. AL, however, cannot distinguish between the two types of goals, as it records no information corresponding to varying reward values.

In summary, the integration of associative and reinforcement learning results in better performance than could be achieved by either model alone in an environment where the reward structure is dynamic and diverse. The AL+RL model displays more flexibility than RL in adapting to changing goals, and more flexibility than AL in adapting to a varying reward structure.

Simulation 2: Blodgett, 1929

Latent learning is a classic behavioral paradigm that focuses on performance in an environment with a dynamic reward structure, and often involves a diverse reward structure. In this paradigm, after having spent some time in an environment, subjects are presented with some goal. Upon the introduction of the goal, subjects display a higher level of performance than would be expected if they had not spent any time in the environment prior to the goal introduction. This phenomenon is observable in children, adults, and animals (e.g. Quartermain & Scott, 1960; Stevenson, 1954; Tolman, 1948).

For example, Blodgett (1929) ran three groups of rats in a maze-learning experiment. One group (the control) was rewarded upon reaching the end of the maze on every trial (R1). The second group began receiving rewards on trial 3 (R3). The third group began receiving rewards on trial 7 (R7). Results demonstrate that subjects in groups R3 and R7 began to perform at the level of control subjects immediately upon the introduction of the reward, producing much steeper error-reduction slopes in these groups than that of R1 (see Figure 2, top-left panel). An associative learning model can predict this phenomenon. Such a model would learn the structure of the maze and begin to employ its knowledge immediately once the reward is introduced.

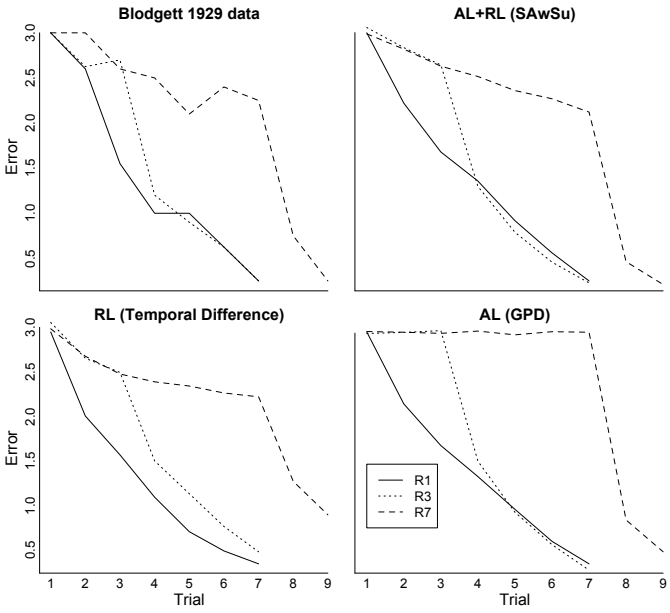

Figure 2: Maze Performance: Avg. Errors by Trial. Data adapted from Blodgett, 1929 (top-left) and simulation results from Reinforcement Learning (bottom-left), Associative Learning (bottom-right), and integrated (top-right) models.

Interestingly, groups R3 and R7 did not continue to display random-level performance until the introduction of the reward. Rather, these groups displayed a shallow error-reduction curve, indicating that there was at least some intention to complete the maze even in the "no-reward" trials ("low-reward" from hereon)[2]. An RL model can predict this phenomenon, producing a shallow learning curve for the "low-reward" trials (R3 until trial 3, R7 until trial 7), and a steeper learning curve for the high-reward trials (R1).

A model that integrates RL and AL should reproduce both (1) the better-than-random level of performance in groups R3 and R7 prior to the introduction of reward, and (2) the steep improvements in performance once this reward is introduced.

The integrated AL+RL model (SA*w*S*u*) was compared with AL (GPD) and RL (TDRL) models. A parameter search was performed, seeking the model parameters that produced the best fit (least sums of square differences) to data in the constant-reward (R1) and the "low-reward" (R7, trials 1-7) conditions. Three parameters were varied for each model: learning rate, amount of exploratory noise (σ), and the perceived low-reward (*LowR*) for finishing the maze on the "low-reward" trials. The learning rate parameter varied for RL was the utility-learning constant, α, and for AL and AL+RL it was the associative-learning constant, β (α remained unvaried for AL+RL at 1.0).

Once the best parameter values were found (RL: $\alpha = .4, \sigma = .08, LowR = .15$; AL and AL+RL: $\beta = .2, \sigma =$

[2]We interpret the shallow learning curves as resulting from a low reward, such as being taken out of the maze.

Table 1: Root Mean Square Difference to Blodgett, 1929.

	Best fit to data		Predicted	
Model	R1	R7 [trials 1-7]	R3	R7*
AL+RL	0.21	0.14	0.11	0.15
RL	0.26	0.17	0.19	0.32
AL	0.22	0.56	0.23	0.50

*Only trials 8 and 9 are predicted.

$.05, LowR = .15$), the full simulations were executed to get model predictions for R3 and for R7 after the introduction of reward (these conditions were not included during the parameter search). Results may be observed in Figure 2 and Table 1. As expected, AL and AL+RL produced steeper performance improvements than RL upon the introduction of the reward by the experimenter on trials 3 and 7. As expected, RL and AL+RL replicated the shallow error-reduction curves in trials 1-3 for condition R3 and 1-7 for condition R7, and AL did not.

AL+RL produced a better overall fit to data than did the other two models (see Table 1). The advantages become more apparent when we focus on the error-reduction after the introduction of reward. Figure 3 demonstrates model predictions for error reduction in the R3 group between trials 3 and 5, and the R7 group between trials 7 and 9. The AL model predicts too high a performance improvement (because the initial performance is underestimated), and the RL model predicts too low a performance improvement in these trials.

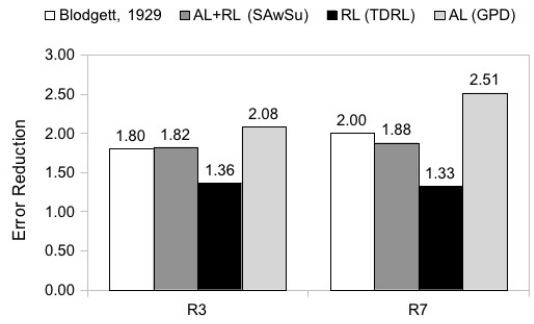

Figure 3: Error reduction after the introduction of reward in Blodgett, 1929.

Summary and Discussion

In this paper we described how two learning mechanisms widely supported in the psychological literature, reinforcement and associative learning, may be integrated. In contrast with RL-only and AL-only models, the integrated model, SA*w*S*u*, was shown to produce more efficient, higher fidelity behavior in environments where the reward structure is both diverse and dynamic.

Gläscher, Daw, Dayan, and O'Doherty (2010) propose an alternative integration of AL and RL by including a supervisory mechanism that learns to arbitrate between AL and RL. This implementation seems less parsimonious than SA*w*S*u* – it has three learning and three decision mechanisms, whereas SA*w*S*u* has two and one, respectively. Further comparison of the two approaches is warranted.

To the best of our knowledge there are no other computational frameworks that learn the reward structure and the spatiotemporal predictions of the environment, and employ both in the decision-making process. Frameworks that employ some form of Model-based planning (e.g. Daw, Niv, & Dayan, 2005; Sutton & Barto, 1998) include both AL and RL, but these tend to focus on the trade-off between planning in the head and acting in the world. Associative knowledge in this class of models is used to enable planning rather than to determine how a path of actions, whether in the head or in the world, is chosen.

The overall scarcity of decision models that employ AL and RL together is rather surprising given the long history of research on learning in experimental psychology, cognitive science, and artificial intelligence. Ohlsson (e.g. Choi & Ohlsson, 2011) has been promoting the integration of learning mechanisms, including AL and RL, and Alonso & Mondragón (2006) and Dickinson & Balleine (1993, 1994) call for AL+RL integration. None of these proposals, however, has been implemented as a computational model, and thus cannot be easily contrasted with the SA*w*S*u* implementation.

The Voicu & Schmajuk (2002) model mentioned in the Introduction, does employ AL in action-selection, and even considers variable utility of the goal state in the decision phase. However, the Voicu & Schmajuk model does not specify any way of actually learning state utilities.

Earlier versions of the ACT-R integrated cognitive architecture included both RL and AL (see Anderson, 1993; Anderson & Lebiere, 1998). However, according to Anderson (2001), the particular form of associative learning implemented in ACT-R turned out to be "disastrous," and produced "all sorts of unwanted side effects" (p. 6). Thus, as it stands, the implementation of associative learning in ACT-R 6 has been reduced to a single equation that relates the fan effect to spreading activation. This limits AL to chunks that have a direct symbolic relationship, where associative strengths can only decrease as more knowledge enters the system and the "fan" of associations to each chunk increases.

The current effort to integrate AL and RL is in accord with the many calls for the integration of cognitive mechanisms within a unified computational framework (e.g. Gray, 2007b; Choi & Ohlsson, 2011). However, the current work presents the integration of only two learning mechanisms, addressing only some of the complexities of large worlds. In the pursuit of models that can produce persistent, adaptive, and flexible behavior in large worlds, it is required that we address how a model like SA*w*S*u* might be incorporated into a broader cognitive architecture such as ACT-R. Further integration of AL and RL with other cognitive mechanisms is the necessary next step for this research.

Acknowledgements

This research was performed while the author held a National Research Council Research Associateship Award with the Air Force Research Laboratory's Cognitive Models and Agents Branch.

References

Alonso, E., & Mondragón, E. (2006). Associative Learning for Reinforcement Learning: where animal learning and machine learning meet. In *Proceedings of the 5th symposium on adaptive agents and multi-agent systems.*

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (2001). Activation, Latency, and the Fan Effect. In *Eighth annual act-r workshop.* Pittsburgh, PA.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford University Press.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the Environment in Memory. *Psychological Science*, *2*(6), 396–408.

Binmore, K. (2009). *Rational Decisions*. Princeton University Press.

Blodgett, H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology*.

Choi, D., & Ohlsson, S. (2011). Effects of multiple learning mechanisms in a cognitive architecture. In *Proceedings of the thirty-third annual meeting of the cognitive science society.*

Chun, M. M. (2000). Contextual cueing of visual attention. *Trends in Cognitive Sciences*, *4*(5), 170–178.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711.

Dickinson, A., & Balleine, B. (1993). Actions and responses: The dual psychology of behaviour. Spatial representation: Problems in philosophy and psychology. In N. Eilan, R. McCarthy, & B. Brewer (Eds.), *Spatial representation: Problems in philosophy and psychology.* (pp. 277–293). Oxford University Press.

Dickinson, A., & Balleine, B. (1994, March). Motivational control of goal-directed action. *Animal Learning & Behavior*, *22*(1), 1–18.

Fu, W. T., & Anderson, J. R. (2006). From recurrent choice to skilled learning: A reinforcement learning model. *Journal of Experimental Psychology: General*, *135*(2), 184–206.

Fu, W. T., & Pirolli, P. (2007). SNIF-ACT: A Cognitive Model of User Navigation on the World Wide Web. *Human Computer Interaction*.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595.

Gluck, K. (2010). Cognitive architectures for human factors in aviation. In E. Salas & D. Maurino (Eds.), *Human factors in aviation, 2nd edition* (pp. 375–400). New York, NY: Elsevier.

Gray, W. D. (2007a). Composition and control of integrated cognitive systems. In W. D. Gray (Ed.), *Integrated models of cognitive systems.* New York: Oxford University Press.

Gray, W. D. (Ed.). (2007b). *Integrated models of cognitive systems*. New York: Oxford University Press.

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.

Myers, C. W., & Gray, W. D. (2010). Visual scan adaptation during repeated visual search. *Journal of Vision*, *8*(10).

Myers, C. W., Gray, W. D., & Sims, C. R. (2012). The insistence of vision: Why do people look at a salient stimulus when it signals target absence? *Visual Cognition*, *9*(19), 1122–1157.

Nason, S., & Laird, J. I. (2005). Soar-RL: Integrating reinforcement learning with Soar. *Cognitive Systems Research*, *6*, 51–59.

Peters, J., Vijayakumar, S., & Schaal, S. (2003). Reinforcement Learning for Humanoid Robotics. In *Humanoids2003, third ieee-ras international conference on humanoid robots, karlsruhe, germany, sept.29-30.*

Quartermain, D., & Scott, T. H. (1960). Incidental learning in a simple task. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *14*(3), 175–182.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In P. W. F. Black AH (Ed.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Russell, S. J., & Norvig, P. (1995). *Artificial Intelligence: A Modern Approach*. Prentice Hall.

Shen, J., Reingold, E. M., & Pomplun, M. (2000). Soar-RL: Integrating reinforcement learning with Soar. *Perception*, *29*, 241–250.

Siegler, R. S., & Stern, E. (1998). Conscious and unconscious strategy discoveries: A microgenetic analysis. , *127*(4), 377–397.

Stevenson, H. W. (1954). Latent Learning in Children. *Journal of Experimental Psychology*, *47*(1), 17–21.

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

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Veksler, V. D., Gray, W. D., & Schoelles, M. J. (2009). Goal-Proximity Decision Making: Who needs reward anyway? In *31st annual conference of the cognitive science society.*

Voicu, H., & Schmajuk, N. (2002). Latent learning, shortcuts and detours: a computational model. *Behavioural Processes*, *59*(2), 67–86.

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. In *1960 ire wescon convention record* (pp. 96–104). New York: Institute of Radio Engineers.