INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Reinforcement Learning for Agent Domain Interactions with Non-Expert Humans

Vijaya Kumar Elpula, Research Scholar and Dr. Suribabu Potnuri, Professor
Department of Computer Science and Engineering, J.S. University, Shikohabad, U.P., India email: vijayakumarvarma@gmail.com

ABSTRACT

Robots and agents often struggle to function autonomously in areas of applications where changes occur often and action results are not predictable. In simple domains, human input may assist an agent understand the job and domain well; yet, in complicated domains, people could lack the knowledge or time to provide detailed, precise feedback. Therefore, for intelligent agents to be widely deployed, they must be able to function independently with just sensory inputs and little high-level feedback from humans who aren't domain experts. To achieve this goal, this article lays out the concepts of an enhanced reinforcement learning system that merges bootstrap learning with RL. Without human input, the agent learns via seeing and responding to its surroundings. Assuming that the agent has access to high-level human input, it will incorporate environmental feedback into its action choice policy by gradually adjusting the relative contributions of the feedback systems. We test the framework in two virtual environments: Keepaway Soccer and Tetris.

Keywords—Reinforcement Learning, Human Computer Interaction, Human Robot Interface, Machine Learning, Artificial Intelligence.

I INTRODUCTION

Robots and intelligent agents that engage in dynamic domain interactions with people must be capable of autonomous, efficient, and reliable operation [1, 2]. The majority of current methods for HCI/HRI either rely on sensory inputs to allow the agent to function autonomously [3, 4] or rely on manual training and domain knowledge [5, 6, 7, 8, 9] to teach the agent. The capacity of an agent to function autonomously in non-deterministic, partially observable settings is often limited [10], [11]. But although human input may assist an agent build a detailed picture of the job and domain; people often don't have the knowledge or time to provide detailed, precise, real-time feedback in complicated domains.

Anyone without technical knowledge should be able to use learning agents; all they need is a way to give the agent high-level feedback on how it's doing, like positive or negative reinforcement, or a list of options to choose from. The ability for an agent or robot to gather human input at the right moment and combine it with data retrieved via sensory signals has been the subject of much recent study. Nevertheless, these approaches are only applicable to basic simulated domains or targeted robot activities since they need extensive domain knowledge and fail to imitate the unreliability of human inputs [12], [13], [14]. In this study, we provide an ARL framework that allows agents to combine environmental reinforcement with restricted and inconsistent high-level human input. The agent may continually and progressively adjust the relative contributions of environmental input and human feedback to its action choices via bootstrap learning, which is part of the ARL architecture. Two simulated domains are used to assess the suggested method: (a) a single-agent game called Tetris and (b) a multi-agent game called Keepaway soccer.

Here is how the rest of the paper is structured. The suggested scheme and test domains are detailed in Section III, while Section II covers related work. Section IV presents the experimental data, whereas Section V draws conclusions.

II LITERATURE REVIEW

Some of the significant Human-Computer and Human-Robot Interface (HCI and HRI) issues that have led to the creation of complex methods include autonomous operation, engagement, safety, acceptability, and interaction protocol design [1, 2]. These issues have been the impetus for the development of these complicated approaches. Several algorithms have been developed specifically for the purpose of enabling autonomous operation in HCI/HRI, which is the subject of this study [15]. These algorithms make use of a wide range of sensory inputs, including as visual, verbal, and range data, to represent social and environmental signals. A significant amount of research has been conducted on the use of virtual agents and embodied relational agents in many applications, such as healthcare [16]. However, the present methods often need a significant amount of subject matter knowledge, which limits the scope of their use.

A significant amount of research has also been conducted on the question of the capacity of a robot or virtual agent to acquire new abilities via the display of such skills by humans [6, 7, 17, 18]. A significant number of these methods are centered on the creation of intricate mathematical models by the incorporation of fresh information from a wide range of relevant fields, such as control theory, biology, and psychology, amongst a great number of other fields. There are a few different approaches that have been developed on the basis of theories about human social interactions and the learning process. The associated feedback and information must be provided by human participants who have comprehensive knowledge of both the domain and the agent capacity. This is a significant drawback of these systems.

The use of sparse, high-level human input in robot and agent domains is getting attention from academics as a realistic alternative that is based on the requirements of certain situations. The CoBot, which was developed by Rosenthal and colleagues [14], is one example. This bot is able to learn from its errors, use success and failure probability functions to direct itself to certain locations, and even seeks help from a human when it becomes disoriented. TAMER, which stands for "Training an Agent Manually via Evaluative Reinforcement," is a system that Knox and Stone [13] developed in order to make it possible for humans to teach learning agents. There are a number of linear functions that are used by this framework in order to integrate human and environmental data, optimize a reward function in simulated domains, and perform other activities. The work that is described in this article also includes both reinforcement learning and a system that integrates environmental and human input. Both of these aspects are very essential. The primary difference is that in order to use all of the information that is available to the agent, the two feedback mechanisms continuously alter their separate contributions to the agent's action choice policy by bootstrapping off of each other. This allows the agent to make the most of all of the information that is available to it.

III PROBLEM FORMULATION

This section describes the framework that combines boot- strap learning with reinforcement learning, followed by a description of the specific test domains.

A. The RL Framework and Bootstrap Learning

Reinforcement Learning (RL) is a computational goal- oriented approach, where an agent repeatedly performs actions on the environment and receives a state estimate and a reward signal [19]. It is common to model an RL task as a Markov decision process (MDP). In this paper, the standard formulation is augmented to include the human feedback signal, resulting in the tuple $\langle S, A, T, R, H \rangle$:

- S is the set of states.
- A is the set of actions.
- $T : S \times A \times S' \rightarrow [0, 1]$, is the state transition function.
- $R : S \times A \rightarrow \Re$ is the environmental reward function.
- H is the human reward signal.

At each step, the agent uses a policy to probabilistically select an action $a \in A$ in state $s \in S$:

$$\pi : S \times A \rightarrow [0, 1] \qquad (1)$$

Finding the course of action that will have the most positive impact on future outcomes over a given planning horizon is the goal. Policy gradient algorithms, policy iteration, and value iteration are just a few of the many approaches that may be used to compute this policy. The integration of high-level human input is the most notable departure from the conventional MDP formulation; this information, like environmental feedback, is not always reliable. In addition, whereas environmental feedback is immediate for a given condition and action, human feedback may be a complex function of not just the current state and action but also of previous and future states and actions.

Using a bootstrap learning approach, the action choice policy is shown in Figure 1 as a function of the feedback signals:

$$a = \underset{a \in A}{\operatorname{argmax}} \, f(R,H) \qquad (2)$$

where R is the environmental feedback, H is the human feedback and a is the action choice that maximizes the function of R and H. In the experimental domains described below, the following functions were evaluated:

$$a = \underset{a \in A}{\operatorname{argmax}} \{w_r \cdot R + w_h \cdot H\} \qquad (3)$$

$$a = \underset{a \in A}{\operatorname{argmax}} \{w_r \cdot R(1 + H^{w_h})\}$$

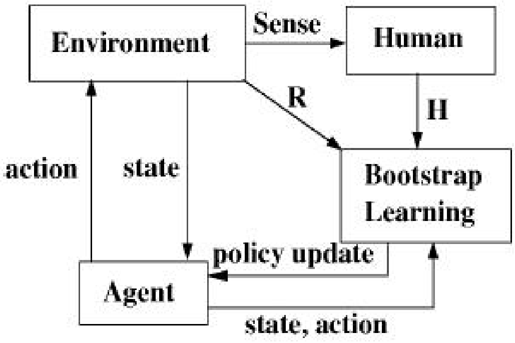


Fig. 1: Augmented RL framework with bootstrap learning.

In where w_r stands for the weight that is assigned to environmental input and w_h stands for the "weight" that is assigned to human feedback. Therefore, it is feasible to set w_r equal to one and use w_h as the relative value of human input since the weights show the relative significance of the information. A similar function was able to attain the maximum performance in the Mountain Car domain when the weight that was ascribed to human input was randomly annealed at the end of each iteration [13]. As a result, a linear function was taken into consideration. One of the distinguishing features of our approach is that it employs a bootstrapping strategy for both of the feedbacks as well. Performance in more complex domains is improved by continuing weight adjustments that are based on the relative usefulness of the feedback signals in improving global performance measures (for example, the amount of time it takes to complete a job) (Section IV). The other combination strategy, which is referred to as the exponential scheme owing to the fact that it employs weight as an exponent, yields the most favourable outcomes among a collection of functions that are equivalent. This technique was used in order to investigate a more robust connection between the reinforcement signals. When employing the bootstrap learning technique, the following is the method that should be followed in order to update the weights:

Without the need for human intervention, the agent is able to experiment with different policies by modifying the parameters of the underlying RL algorithms (for example, policy gradient), and then evaluate the effectiveness of these policies by using an appropriate "performance measure" (for examples, see Sections III-B and III-C). During each and every second, the agent is responsible for keeping a record of the top N policies, which is represented by the symbol πi. The number *i* ranges from 1 to N, and it is arranged in descending order of the performance measures, pmi, which are similarly indexed from 1 to N. What is need to be done in accordance with one of these policies (where $w_r = 1$, H = 0 in Equation 3), where the probability of selecting a policy is inversely proportional to the value of its performance metric in relation to other policies, and where the value of the performance metric is a measure of how well the policy performs.

It is the responsibility of the agent to maintain a separate policy in response to both positive and negative stimuli that are provided by humans. In addition to this, the agent evaluates the degree to which the reaction chosen by the current environmental feedback-based policy is congruent with the response chosen by the human feedback-based policy. One method for determining the degree of congruence between the two policies is to maintain a record of the frequency with which they both result in the same course of action. It is possible to represent the estimated weight that is associated with human feedback as the degree of match between the human feedback-based policy and the best environmental feedback-based policies, where mi and i are values that fall within the range [1, N].

$$w_h = \frac{\sum_i pm_i \times m_i}{\sum_i pm_i} \qquad (4)$$

$$w_h^k = \frac{pm^{k-1}w_h^{k-1} + pm^k m^k}{pm^{k-1} + pm^k} \tag{5}$$

Where a high number indicates that the human comment is highly believed. If significant human input is available, a comparable technique may be used to weight environmental feedback relative to one or more sources of human feedback. It is also feasible to slowly adjust the weights between episodes. Take, for example, the weight that may be changed after episode k:

taking into consideration the performance measure from both this episode and the one before it (*k* minus 1) as a foundation.

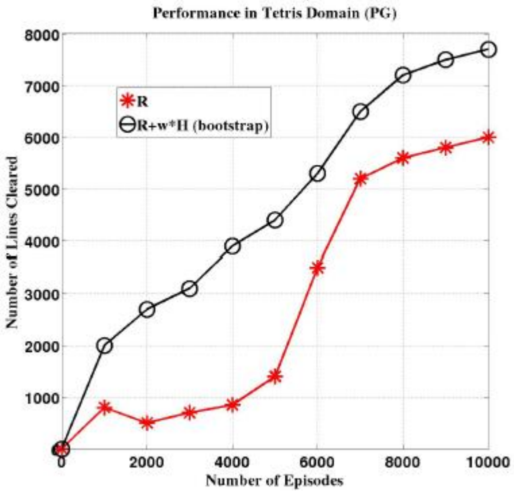
When it comes down to it, it is an online process that is continually changing the action policy. Depending on each feedback mechanism, the agent takes it in turns assuming a policy or policies be ground truth, and then adjusting the policy's weight in accordance with the assumption. Because action choices are based on the integrated policy (Equation 3), the agent is able to quickly react to a wide variety of individuals, unreliable environmental input, and dynamic changes (such as the human observer getting bored or exhausted). Specifically, we will describe two simulated domains that make use of this learning technique in the following paragraphs.

IV Experimental Setup and Results

The results of testing that were conducted in the Keepaway and Tetris domains are detailed in this section. According to one hypothesis, the performance of the bootstrap learning scheme is much superior to that of the underlying reinforcement learning algorithms. On the other hand, the performance of the combination of human and environmental feedbacks is superior than the performance of each respective feedback mechanism on its own. Throughout the remainder of this study, the term "ARL" will be used to designate the upgraded RL framework's utilization of bootstrap learning. A degree of significance of 99% is applied to each and every conclusion, unless it is specifically stated differently. People on their own As a Participant: Four non-expert human participants in the trials were given a high-level explanation of the test domains, available action alternatives, and performance metrics that were to be maximized by the agent or agents. This information was presented to them. The participants had a limited understanding of the domains and states prior to the initiation of these investigations, and they were also unaware of the algorithms that were being used throughout the process. The behaviour of the agent was either favourably or adversely rewarded as a consequence of the input that was provided by humans.

Experiments were conducted in the Tetris environment, with cross-entropy serving as the starting point for the reinforcement learning algorithm. The ARL approach and the linear combination function included in Equation 3 were used in order to integrate the information from both humans and the environment. The strategy that anneals the weight allotted to human feedback at the end of each episode—which integrated the signals—provided the maximum performance in the basic Mountain Car domain [13]. A comparison of performance was carried out using this approach. All of the results of the experiment are shown in Figure 2, with each data point representing the average of twenty different experiments. There were a maximum of five instances of human input included inside each episode, with an average of two occurrences being included in each episode. In the intervals between each trial, participants were provided with a break.

The ARL technique beats both the default CE method and the combination scheme that anneals the weight allotted to human input between episodes in terms of performance, as assessed by the number of lines cleared [13]. This is evident from the fact that the ARL method is superior to both of these aforementioned methods. Because the ARL approach is able to adapt to the unreliability of feedback signals, it is able to make advantage of the complementary



aspects of feedback signals, which results in improved performance. Despite the fact that Figure 2 does not display the outcomes that were obtained only via the use of human input, the ARL technique is superior than it. This is owing to the fact that it is not feasible to provide human input across a variety of states and activities over a significant number of episodes.

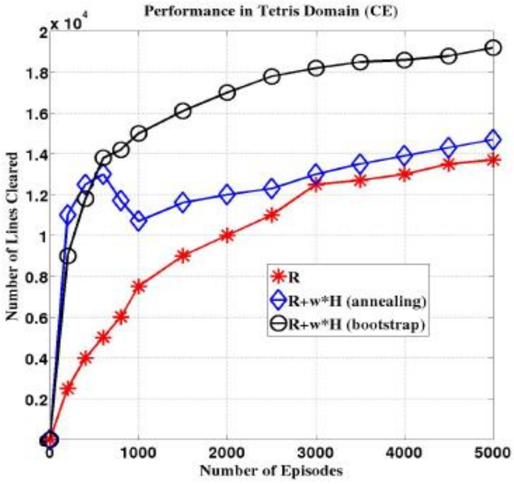


Fig. 2: Performance in the Tetris domain using cross entropy as the default RL algorithm. The ARL approach performs significantly better than CE and the scheme that anneals the weight factor for human feedback.

Fig. 3: Performance in the Tetris domain using policy gradient as the default algorithm. The ARL approach significantly improves the performance of the PG algorithm.

The next thing that needed to be done was to evaluate how well the ARL approach functioned by using policy gradient (PG) as the localization procedure. The results of these tests are shown in Figure 6, which you may see. In comparison to the CE technique, the default PG approach is supposed to reduce the number of lines that are cleaned. This is the desired effect. Although the ARL approach performs much better when the feedback systems are combined, the default PG methodology clears a significantly less number of lines than the ARL method does.

In conclusion, tests were conducted in the Keepaway social domain, where the performance of goalkeepers was rated based on their ability to hold control of the ball for the maximum period of time while maintaining optimum performance.

The method for RL that was used was the SMDP variation of the Sarsa(λ) algorithm. A step-by-step application of the ARL technique was carried out in order to determine the best combination of human and environmental input for the action choice policy. We examined the exponential combination function (Equation 3) as well as the linear combination function (Equation 2) in this section. However, as indicated in Section III-C, it is not possible for humans to offer quick feedback in this area. The performance was examined both with and without the use of the gamma PDF (Figure 2) in order to establish whether or not it is acceptable for credit assignment when human input is provided. As was the case in the Tetris domain, each data point in Figure 4 reflects the average of the twenty trials, which provides a summary of the results. Figure 4 illustrates that the length of each episode may vary; hence, the human participants provided feedback infrequently, no more than twice each episode. Additionally, the human participants provided input that was intentionally incorrect; around twenty percent of the time, individuals get it incorrectly.

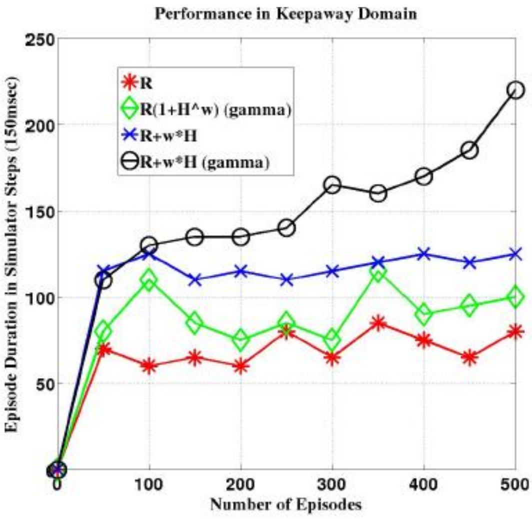


Fig. 4: Performance in the Keepaway soccer domain using policy gradient. The ARL approach performs better than the default *Sarsa* (λ) algorithm. Using the learned gamma distribution for credit assignment significantly boosts performance

Figure 4's graphs show that, despite human input's unreliability, all schemes that integrate feedback mechanisms utilizing the ARL technique outperform the default RL algorithm without any feedback. Bootstrap learning swiftly adjusts the weights when a human participant gives inaccurate input. Also, compared to relying only on human feedback—not shown in Figure 4 due to the difficulty of providing inputs across a large number of episodes—the performance is superior. Compared to the default RL method, the performance is much improved by combining the two feedback systems in a weighted linear combination. When the linear combination function and the gamma PDF are utilized for credit assignment, the best result is attained. It seems that the exponential combination function does not accurately represent the connection between the feedback signals, as it does not significantly enhance performance—even when using the gamma PDF. Using the bootstrap learning scheme and gamma PDF-based credit assignment inside the enhanced reinforcement learning framework shows promise for combining human and environmental feedbacks in additional domains, according to the experimental findings in the two test domains [25].

Challenges to Reliability: The study's results might be influenced by the human participants' skills when human input is used in domains with intelligent agents or autonomous robots. Four human volunteers supplied inputs at irregular intervals for the research detailed in this article. To further support the findings given here, further trials may be necessary in other areas, even if the participants' performance (when looked at individually) was constant across the two domains. More human subjects, more episodes, other levels of additional noise, different combination functions, and different relevant application areas will all be considered in future trials, along with a comprehensive analysis of the accompanying experimental data.

V CONCLUSIONS AND FUTURE WORK

There is a possibility that agents or robots may develop a comprehensive grasp of the task and domain by working alongside people. This would enable them to perform successfully and consistently in contexts that are constantly changing. However, it may be unreasonable to anticipate that a person would possess the expertise and experience necessary to offer agents in complex areas with feedback that is accurate, complete, and delivered in a timely manner. The purpose of this study was to establish a strategy that utilizes bootstrap learning inside an improved reinforcement learning framework. The goal of this approach was to assist agents in making the most of both human-provided, limited-scope, high-level feedback as well as reward signals obtained from environmental interactions. Weights are used to establish the relative contribution that each feedback method makes to the action choice policy of the agent. These weights are altered in a progressive and ongoing manner. As a result of this, the agent is able to make the most efficient use of the data that is available to them. The results of the studies indicate that the approach described in this paper is superior to both the individual feedback systems and the methods that are already in use that combine the two types of feedback systems. The objective of the research that is being addressed in this paper is to construct a robust combination of human inputs and sensory cues. In the future, research may investigate the possibility of including an underlying probabilistic belief representation in order to make it possible for agents (or robots) to operate in partially visible settings and automatically get relevant human input when it is necessary. The focus of research in the future will also be on real-world scenarios in which several individuals or robots collaborate to accomplish a shared objective.

REFERENCES

[1] M. A. Goodrich and A. C. Schultz, "Human-Robot Interaction: A Sur- vey," *Foundations and Trends in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007.

[2] A. Tapus, M. Mataric, and B. Scassellati, "The Grand Challenges in Socially Assistive Robotics," *Robotics and Automation Magazine, Special Issue on Grand Challenges in Robotics*, vol. 14, no. 1, pp. 35–42, March 2007.

[3] M. Cakmak, N. DePalma, R. Arriaga, and A. Thomaz, "Exploiting Social Partners in Robot Learning," *Autonomous Robots*, vol. 29, pp. 309–329, 2010.

[4] J. Forlizzi and C. DiSalvo, "Service Robots in the Domestic Environ- ment: A Study of the Roomba Vacuum in the Home," in *International Conference on Human-Robot Interaction, HRI-06*, Salt Lake City, USA, March 2-4 2006, pp. 258–266.

[5] B. Argall, S. Chernova, M. Veloso, and B. Browning, "A Survey of Robot Learning from Demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009.

[6] C. Breazeal and A. Thomaz, "Learning from Human Teachers with Socially Guided Exploration," in *International Conference on Robotics and Automation*, 2008, pp. 3539–3544.

[7] D. Grollman, "Teaching Old Dogs New Tricks: Incremental Multimap Regression for Interactive Robot Learning from Demonstration," Ph.D. dissertation, Department of Computer Science, Brown University, 2010.

[8] D. Perzanowski, A. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a Multimodal Human-Robot Interface," *IEEE Intelligent Sys- tems*, vol. 16, no. 1, pp. 16–21, January-February 2001.

[9] P. Zang, A. Irani, P. Zhou, C.

Isbell, and A. Thomaz, “Using Training regimens to Teach Expanding Function Approximators,” in *International Joint Conference on Autonomous Agents and Multiagent Systems (AA- MAS)*, 2010, pp. 341–348.

[10] T. Fong, I. Nourbakhsh, and K. Dautenhahn, “A Survey of Socially Interactive Robots,” *Robotics and Autonomous Systems*, vol. 42, no. 3- 4, pp. 143–166, 2003.

[11] S. Thrun, “Toward a Framework for Human-robot Interaction,” *Human- Computer Interaction*, vol. 19, p. 2004, 2004.

[12] J. Biswas and M. Veloso, “WiFi Localization and Navigation for Autonomous Indoor Mobile Robots,” in *International Conference on Robotics and Automation (ICRA)*, Anchorage, USA, May 3-8 2010.

[13] W. Knox and P. Stone, “Combining Manual Feedback with Subsequent MDP Reward Signals for Reinforcement Learning,” in *International Conference on Autonomous Agents and Multiagent Systems*, May 2010.

[14] S. Rosenthal, J. Biswas, and M. Veloso, “An Effective Personal Mobile Robot Agent Through Symbiotic Human-Robot Interaction,” in *Interna- tional Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Toronto, Canada, May 2010.

[15] J. Fasola and M. Mataric, “Robot Motivator: Increasing User Enjoyment and Performance on a Physical/Cognitive Task,” in *International Conference on Development and Learning*, Ann Arbor, USA, August 2010.

[16] A. Rizzo, T. Parsons, G. Buckwalter, and P. Kenny, “A New Generation of Intelligent Virtual Patients for Clinical Training,” in *The IEEE Virtual Reality Conference*, Waltham, USA, March 21 2010.

[17] A. Billard and K. Dautenhahn, “Experiments in Social Robotics: Grounding and Use of Communication in Autonomous Agents,” *Adap- tive Behavior*, vol. 7, no. 3-4, pp. 415–438, 1999.

[18] D. W. Franklin, T. W. Milner, and M. Kawato, “Single Trial Learning of External Dynamics: What can the Brain Teach Us about Learning Mechanisms in Brain Inspired IT,” *International Conference on Brain- Inspired Information Technology*, pp. 67–70, 2007.

[19] R. L. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1998.

[20] O. Buffet and D. Aberdeen, “The Factored Policy-Gradient Planner,” *Artificial Intelligence*, vol. 173, no. 5-6, pp. 722–747, 2009.

[21] I. Szita and A. Lorincz, “Learning Tetris using Noisy Cross Entropy Method,” *Neural Computation*, vol. 18, pp. 2936–2941, 2006.

[22] P. Stone, R. Sutton, and G. Kuhlmann, “Reinforcement learning for robocup soccer keepaway,” *Adaptive Behavior*, vol. 13, pp. 165–188, 2005.

[23] W. E. Hockley, “Analysis of Response Time Distributions in the Study of Cognitive Processes,” *Journal of Experimental Psychology. Learning, Memory*

and Cognition, vol. 10, 1984.

[24] W. Knox, I. Fasel, and P. Stone, "Design Principles for Creating Human- Shapable Agents," in *AAAI Spring Symposium on Agents that Learn from Human Teachers*, 2009.

[25] M. Aerolla, "Incorporating Human and Environmental Feedback for Robust Performance in Agent Domains," Master's thesis, Computer Science, Texas Tech University, May 2011.