iCORE ANNUAL REPORT 2009

iCORE CPE GRANT CPE45

DR. RICHARD SUTTON

REINFORCEMENT LEARNING AND ARTIFICIAL INTELLIGENCE

ICORE ANNUAL REPORT MARCH 31, 2009

1. EXECUTIVE SUMMARY

The RLAI research program pursues an approach to artificial intelligence and engineering problems in which they are formulated as large optimal-control problems and approximately solved using reinforcement-learning methods. Reinforcement learning is a new body of theory and techniques for optimal control that has been developed in the last twenty years primarily within the machine learning and operations research communities, and which have separately become important in psychology and neuroscience. Reinforcement learning researchers have developed novel methods to approximate solutions to optimal control problems that are too large or too ill-defined for classical solution methods such as dynamic programming. For example, reinforcement-learning methods have obtained the best known solutions in such diverse automation applications as helicopter flying, elevator scheduling, playing backgammon, and resource-constrained scheduling.

The objectives of the RLAI research program are to create new methods for reinforcement learning that remove some of the limitations on its widespread application and to develop reinforcement learning as a model of intelligence that could approach human abilities. These objectives are pursued through mathematics, through computational experiments, through the development of robotic systems, and through the development and testing of computational models of natural learning processes.

The research team consists of about 45 members, 30 of whom are graduate students and, of those, 21 of which were recipients of major scholarships. The output of the research program has remained strong, with 23 papers published or accepted for publication in highly-refereed archival venues during the reporting period. Two PhD and six MSc students were graduated.

The primary focus of the research program has been on how intelligent machines represent their knowledge of the world. The key question is how to organize the knowledge such that it can be verified, learned, and used autonomously without continual tending by human experts. This project has pursued an unusual approach in which knowledge is expressed in terms of the machines' sensors and actuators, thereby enabling it to be compared directly to experiential data. Substantial further progress was made this year toward formalizing the core learning algorithms and developing planning algorithms.

Highlights of the research program this year include 1) the development of new gradient-based temporal-difference learning methods for off-policy learning, apparently solving a key long-standing problem, 2) the construction of the Critterbot, a new robotic platform that acts as a common focus for the project's research into experience-grounded knowledge, and 3) a major new release of RL-Glue, our research and teaching software.

2. RESEARCH PROGRAM OVERVIEW

This is the first year of the renewal of the RLAI project. In the renewal proposal, the project's research was divided into three main target areas. The first is extensions of conventional reinforcement learning algorithms; there are many open problems in reinforcement learning, and we seek to solve them as opportunities arise. The second area is the extension of reinforcement learning ideas to address the more ambitious goals of artificial intelligence (AI). There is a natural transition from the more advanced reinforcement learning methods to mechanisms for knowledge representation, search, and human-level reasoning. A major goal for the project is to explore, implement, and illustrate these relationships. The third main area of RLAI research is a focus on applications—on designing algorithms and software that are well-suited for applied research, and on several specific applications with which we are currently working. We discuss highlights of our research towards each of these target areas next.

In the area of core reinforcement-learning algorithms, there has been some exciting progress. There can be no more core reinforcement-learning algorithm than temporal-difference (TD) learning. A variant of TD learning is used in almost all large-scale applications of reinforcement learning. However, existing TD algorithms are limited in key ways that prevent realization of the full range of potential applications, including applications toward the ambitious goals of AI. On specific problems and training regimes, TD algorithms are known to be unstable and can only be made stable by severely restricting the power of the approximations that they can consider, or by greatly increasing the complexity of the algorithms. For the past fifteen years, finding a way to remove these limitations while retaining TD learning's key strengths has been one of the most important, if not *the* most important, open problem in the field of reinforcement learning. We believe that we may have solved this problem this year.

Our solution to this open problem is an extension of the gradient-based temporal-difference (GTD) algorithms we introduced last year, but found to be too slow for practical use. This year we developed similar ideas but based on a more thorough theoretical foundation, and proposed two new algorithms, linear GTD2 and linear TDC, which appear to retain all of the advantages of conventional linear TD learning, including its fast rate of learning and computational simplicity, while being stable on all problems. We have shown this on a number of small test problems, on test problems for which conventional TD learning is unstable, and on a larger 5x5 Computer Go. The new algorithms were stable on all problems and, on those on which conventional TD learning was also stable, the new algorithms converged at a similar rate to TD learning. More experiments are needed before we can declare the open problem fully closed, but it already seems clear that our new algorithms are a significant breakthrough.

If the new gradient TD algorithms continue to prove effective, they will become key to all our future efforts in the area of extending reinforcement learning to the more ambitious goals of AI. The simplest way to understand this is to consider the notion of a universal prediction-learning algorithm. Brain scientists and computational scientists have proposed that a single, sufficiently powerful algorithm for prediction learning might

suffice to explain most of the decision-making and organizational powers of the human mind. Conventional TD learning algorithms such as TD(λ) have been candidates for the single algorithm, but poor ones because of their instability. The new gradient TD algorithms are much more likely to eventually play this role.

In particular, they solve the problem of off-policy learning. Off-policy learning refers to learning about one way of behaving while actually behaving in another way. It arises most commonly in reinforcement learning algorithms that learn about an optimal way of behaving while actually behaving in a more exploratory way. Off-policy learning also arises when we try to simultaneously learn predictions about many different ways of behaving, a process known as *intra-option learning*. Although it has not yet been done as of this writing, it should be possible to straightforwardly extend the new gradient TD methods to intra-option learning methods that can be used off-policy, and thus learned in parallel from a single stream of experience. This would be a significant step toward the ability to ground abstract knowledge in predictions of sensori-motor experience.

Finally, in the area of applications, we have continued to make good progress with Computer Go and with standardized software, as discussed in subsequent sections, but we have made the most progress this year in designing and constructing a robotic platform for application of research ideas. The *Critterbot* (Figure 1, left) is a small mobile robot outfitted with an unusually rich set of redundant sensors, including infrared proximity sensors, directional light sensors, binaural microphones, a three-axis accelerometer, a gyroscope, a radio-spectrum sensor, a compass, and sensors for the battery level and for the current, temperature, and velocity of the three motor driving the wheels. A real-time display of the instantaneous sensor values and various statistics of their recent values is shown in the second panel of Figure 1. We have deliberately avoided laser range-finders and other sensors that would tempt us to think in Cartesian rather than experience-centered terms. The robot can translate and rotate independently and simultaneously in any direction and can also express itself through a speaker and a circle of bright color LEDs around its upper surface.

Applying our methods to a physically realized robot such as the Critterbot is a useful complement to our studies in computational worlds where everything can be completely controlled and understood. A real robot forces one to come to grips with temporal issues such as sensing and acting delays, asynchrony of perception and action, and the need for real-time responses. It can also provide a focus for group activity, a single system that can be addressed in multiple ways while reflecting a consensus about objectives. We have also built a 2D Critterbot simulator (see right panel of Figure 1) that supports the same sensori-motor interface as the physical Critterbot. A standard RL-Glue interface is available to both the robot and the simulator.

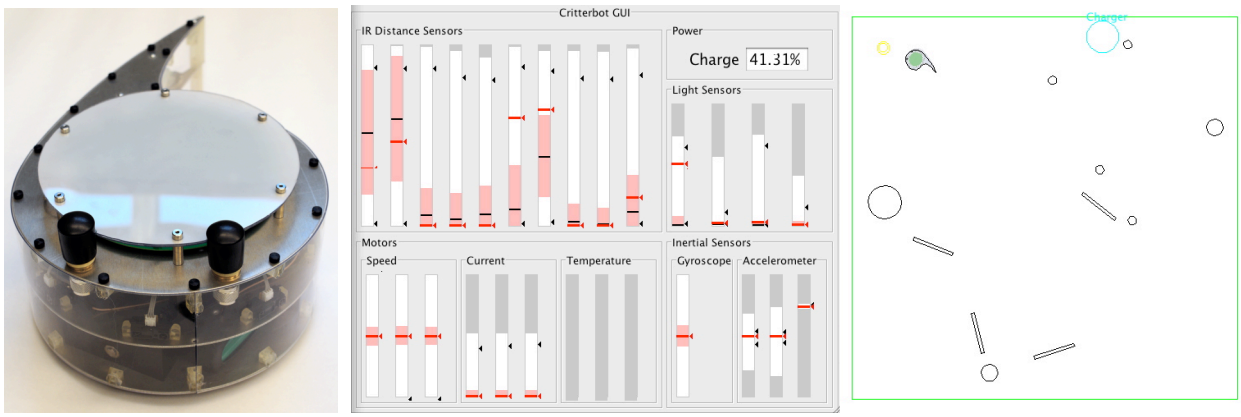

Figure 1. Left) the Critterbot, a sensor-rich mobile robot approximately one foot in diameter; middle) a GUI visualization of its sensor values; right) a 2D Critterbot simulator, with the simulated Critterbot in the upper left of a bounded field with various objects. Agent programs and the GUI can be used interchangeably with the physical Critterbot and the simulator.

3. RESEARCH PROJECTS

This section describes in more detail a few of the other research projects (not mentioned above) making up the research program.

Computer Go

The ancient oriental game of Go has long been a challenge to artificial intelligence. The techniques that have worked so well in chess and in so many real-world applications seemed to have no traction in Go because of its large branching factor, which makes traditional search impossible. After decades of research, the best Computer Go programs were still no challenge to weak amateur players. Three years ago it looked likely that computers might never play a strong game of Go. In these last three years, however, the world of Computer Go has undergone a revolution, in large part because of the incorporation of sample-based search (planning) and learning methods influenced by reinforcement learning research. RLAI project members have played key roles in these developments. Principal investigator Csaba Szepesvari developed the key Monte-Carlo tree search method known as UCT, now used in all the top programs. PhD student David Silver collaborated with Sylvain Gelly at the University of South Paris, to help develop *MoGo*, a powerful Computer Go program based on UCT. Articles on *MoGo* have been written in *The Economist*, *Scientific American* and *Pour La Science*.

This year the revolution in Computer Go has continued apace. *MoGo* is now the world's first master-level 9x9 Computer Go program; the winner of the gold medal at the 19x19 Computer Go Olympiad; and the first program to beat a human professional in even games on 9x9 boards and in handicap games on 19x19 boards. Silver's research contributions include the heuristic UCT algorithm and the UCT-RAVE algorithm; these are now used in all the top Go programs. This year he also worked with Gerald Tesauro at IBM's Watson Research Center, as part of a DARPA-funded joint research project, to develop the first algorithms for automatically learning effective rollout policies for

Monte-Carlo simulation. He has also collaborated with Joel Veness at the University of New South Wales, to develop the strongest chess program with an evaluation function learnt entirely from self-play.

Computational Models of Animal Learning

Reinforcement learning is studied in psychology and neuroscience as well as in engineering and artificial intelligence. In psychology, reinforcement-learning methods are important models of elemental learning processes in animals, such as classical conditioning. In neuroscience, reinforcement-learning methods are the dominant models of reward systems in the brain, in particular of the dopamine system. Working with Professor Jim Kehoe of the University of New South Wales, one of the world's foremost experts in animal learning psychology, we have been exploring ways in which natural learning systems might provide insights into computational reinforcement learning algorithms, as has happened in the past, leading to some of the most effective modern algorithms such as TD(λ) and Q-learning.

Last year we developed a computational model of classical conditioning and related it to both behavioral and neurophysiological data. The key innovation of that model was the use of simple stimulus events such as onsets and offsets to trigger a temporally extended internal representation—modeled as a sequence of temporally extended internal "micro-stimuli" with a range of delays and proportional dispersions. We are now exploring the use of similar ideas to represent state in artificial systems such as the Critterbot.

This year we have explored the implications of our micro-stimulus model as a theory of the role of the hippocampus—a key brain structure in learning and memory—in simple conditioning experiments. We hypothesized that hippocampus affects the speed at which micro-stimuli are recruited and decay. This together with some other natural assumptions about stimulus representations was sufficient to explain the differential effect of hippocampal damage on trace and delay conditioning, and to make several novel predictions for new experiments.

Autonomous Outdoor Navigation and Goal Finding

In the last few years we have conducted research on the problem of autonomous robotic geocaching, which involves locating a goal object in an unstructured outdoor environment given only its rough GPS position. The potential benefits of this research include applications to agriculture, search and rescue, surveying, transportation, and interstellar exploration. There are a number of attributes of the autonomous geocaching problem, which encompasses the problems of unstructured outdoor navigation and goal finding, that make it both interesting and challenging. These include the countless outdoor hazards which make obstacle detection difficult; large and interconnected obstacles inherent in outdoor environments (e.g. ditches, dense forest, buildings) which necessitate frequent backtracking and sophisticated path-planning; and the lack of additional information, such as satellite maps, GPS waypoints, or obstacle descriptions, available to the robot. These difficulties have limited previous work in outdoor

navigation to the development of systems that rely on structural cues (e.g. road, paths, manually colored obstacles) to aid both obstacle detection and goal finding. This reliance limits these systems to a narrow range of environments.

This year we have produced and validated the final version of a robotic system, named *Kato*, capable of solving the task of fully autonomous geocaching. Kato was constructed by outfitting a self-balancing Segway robotics platform with sensors and computing resources. These sensors included a GPS receiver and inertial sensor for position estimation, a laser rangefinder for obstacle detection, and a camera for goal identification. The key challenges in solving the geocaching task include: constructing a map of the terrain as the robot navigates, using this map for planning and following a path to the GPS coordinate, and finally searching the local area for the goal object. The effectiveness of the system was tested in two different outdoor environments, and its performance navigating over several kilometers was compared to that of a human expert teleoperating the robot. Results indicated that, in the test environments, the robotic system was able to navigate to, and detect, the goal object with a high level of dependability. Moreover, path planning, navigational efficiency (maximum speed), and obstacle avoidance was similar when the system was operated autonomously to when it was teleoperated by a human expert. Overall, this work makes strides towards the development of a cost-effective robotic system that can effectively operate in challenging real-world environments.

4. OBJECTIVES FOR NEXT YEAR

Among our immediate objectives for next year are the further testing and development of the new gradient TD learning methods. We will expand the empirical tests into a larger 9x9 Computer Go application with over one million features and to a control context based on learning from self-play. Because these are gradient methods, it is natural to seek also their extension to the case of nonlinear function approximation, finally putting the theory of TD learning on a par with the comparable conventional theory of supervised learning. Finally, we seek to extend the gradient TD methods to an intra-option form for learning general option-models with recognizers. This will take us close to the goal of a single, universal prediction-learning algorithm. Over the next year we will be seeking to make this extension theoretically and to test its effectiveness experimentally.

One of the testbeds that we will use to test this and other learning algorithms is the Critterbot. Experimental work with physical robots is always challenging and often takes longer than planned. Nevertheless, we expect to conduct a number of learning experiments next year. The first milestone will be to demonstrate a reinforcement-learning algorithm running and learning on the robot in real-time. Another milestone that we seek to achieve next year is the demonstration of the robot interacting with a human trainer to learn a task more quickly than could be learned by the robot alone, or by the robot interacting with a human programmer. A third milestone that we plan to achieve next year is the learning of option models using both conventional and gradient-based TD learning algorithms. The plan here is to give the robot a variety of built-in policies, such as wall-following, spinning in place, wandering at random, traveling in straight lines, and

navigation to beacons, and then to task it to learn the likely outcomes of all the corresponding options, and finally to task it to chose among the options to achieve a goal.

A major objective, which we intend to reach toward next year, is the exploration of directed methods for *discovering* representations and other structural elements of world knowledge. These include state variables, new features (by combining existing features), and ultimately options and subgoals. A natural place to begin is with the discovery of state variables for the Critterbot. The Critterbot processes sensor data and generates actions 100 times a second. At this rate, the effects of actions are not visible in sensor data until several, and perhaps many, time steps later. To make sense of this, the AI agent will need to develop memories of recent actions, that is, state variables. We anticipate that such low-level dependencies of the present on the past can be used as a major driving force for the discovery of state variables that will then prove useful in higher-level abstractions.

5. RESEARCH TEAM MEMBERS AND CONTRIBUTIONS

a. Team Leader

Name	Role / Topic	Awards / Special Info
Richard Sutton	Principal investigator	NSERC, AICML

b. Faculty Team Members

Name	Role / Topic	Awards / Special Info
Michael Bowling	Principal investigator Robotics, games, reinforcement learning	NSERC, AIF New Faculty, AICML
Dale Schuurmans	Principal investigator Probabilistic methods in artificial intelligence, machine learning	CRC II Chair, NSERC, AICML
Csaba Szepesvári	Principal investigator Nonparametric learning, statistical techniques	NSERC, AIF New Faculty, AICML

c. Postdoctoral Fellows

Name	Topic	Awards / Special Info
Thomas Degris-Dard	Critterbot	
Mohammad Ghavamzadeh	Bayesian reinforcement learning, value function approximation in Reinforcement Learning	Research Scientist at INRIA France, September 2008
Wenye Li	Algorithms in machine learning	NSERC

Yuxi Li	Machine learning, computational finance	
Elliot Ludvig	Computational models of animal learning	
Barnabás Póczos	Machine learning, unsupervised learning, independent component analysis	
Eric Wiewiora	Reinforcement learning	
Linli Xu	Unsupervised support vector machines, Unsupervised and semi-supervised learning	

d. PhD Candidates

Name	Role / Topic	Scholarships / Awards / Special Info
Yasin Abbasi-Yadkori	Continuum-armed bandit problems	
Arash Afkanpour	Multitask learning	iCORE ICT
Nolan Bard	Applying state estimation techniques for modeling dynamic agents in Texas Hold'em poker	NSERC PGS, iCORE ICT, President's Doctoral Prize of Distinction
Gabor Bartok	Representation learning	
Amir massoud Farahmand	Regularities in Reinforcement learning	
Marc Gendron-Bellemare	Model-based Reinforcement learning	NSERC CGS, iCORE ICT, Walter H. Johns
Michael Johanson	Opponent modeling in computer poker	iCORE ICT
Anna Koop	Robot epistemology: knowledge representation for autonomous artificial intelligence	AIF, iCORE ICT, NSERC PGS
Marc Lanctot	Solving large games using sampling I regret minimization	AIF, iCORE ICT, NSERC PGS
Hamid Reza Maei	Reinforcement learning with function approximation from off-policy data	iCORE ICT
James Neufeld	Subjective mapping on mobile robots	NSERC PGS, iCORE ICT, Walter H. Johns
Chris Rayner	Modeling high-dimensional data streams	NSERC PGS, President's Doctoral Prize of Distinction
Yi Shi	Genetic network inference based on gene expression data and prior knowledge	PhD Provost Scholarship

David Silver	Reinforcement learning in computer go	iCORE
Brian Tanner	Parameter free reinforcement learning	AIF, iCORE ICT, Honorary Izaak Walton Killam Memorial Scholarship, Ralph Steinhauer Awards of Distinction
Adam White	Efficient algorithms, interfaces and training techniques to teaching autonomous robots	AIF, iCORE ICT, NSERC, Presidents Doctoral Distinction
Min Yang	Machine learning	China Scholarship Council
Hengshuai Yao	Reinforcement learning	
Yaoliang Yu	Robust approach for uncertain Markov decision processes	

e. MSc Candidates

Name	Role / Topic	Scholarships / Awards / Special Info
Kate Davison	Automatic generation of bidding strategies for bridge	
Leah Hackman	Feature discovery and option learning in Reinforcement learning	
Brad Joyce	Continuous action reinforcement learning	
Yavar Nadaf	General game playing in Atari 2600	Government of Alberta
David Schnizlein	State translation in no-limit poker	M.Sc. Academic Achievement Award
Kevin Waugh	Abstraction in large extensive games	NSERC PGS, iCORE ICT
Martha White	An automated approach to agent evaluation in extensive games	NSERC CGS, iCORE, Isabel Soper Memorial Graduate Scholarship

f. Other Team Members

Name	Role
Akiko Green	Administrative assistant
Mathew Radkie	Programmer
Mike Sokolsky	Robot engineer
Lori Troop	Program administrator
Eric Verbeek	Undergraduate research assistant

g. Visitors

Name	Institution
Andras Antos	MTA SZTAKI, Hungary
Shai Ben-David	University of Waterloo
Shalabh Bhatnagar	India Institute of Science
Tiberio Caetano	NICTA, Australia
Lawrence Cayton	University of California San Diego
Nicolo Cesa-Bianchi	University of Calgary
Wenyuan Dai	Shanghai Jiao Tong University, China
Nando de Freitas	University of British Columbia
Reza Haffari	Simon Fraser University
Alexander Hentschel	University of Calgary
Fancisco Melo	Carnegie Mellon University
Remi Munos	INRIA, France
David Pal	University of Waterloo
Dmitry Pechyony	Technion, Israel
Doina Precup	McGill University
Özgür Simsek	Max Planck Institute, Germany
Kevin Small	University of Illinois at Urbana-Champaign
Istvan Szita	Rutgers University
Russ Tedrake	Massachusetts Institute of Technology
Gerald Tesauro	IBM T. J. Watson Laboratories
Shankar Vembu	Fraunhofer IAIS, Germany
Huizhen Yu	University of Helsinki, Finland

6. COLLABORATIONS

Provincial	
Participants	Nature of Collaboration
Alberta Ingenuity Center for Machine Learning	R. Sutton, D. Schuurmans, Cs. Szespesvari and M. Bowling are among the ten principle investigators for this center at the University of Alberta. Total annual funding for AICML is $2M.

National	
Participants	Nature of Collaboration
Doina Precup, McGill University Prakash Panangaden, McGill University Yoshua Bengio, Montreal University Shie Mannor, McGill University Richard Sutton, University of Alberta	NSERC Collaborative Research and Development Grants-Project, "Learning and prediction in high-dimensional stochastic domains," with Nortel Networks and Bell Canada. $186,523 total funding (Sept 1/06 – Aug 31/09).
Yoshua Bengio, McGill University Hugh Chipman, University of Waterloo Shai Ben-David, University of Waterloo Richard Sutton, University of Alberta Pascal Vincent, Montreal University	MITACS grant "Statistical Learning of Complex Data with Complex Distributions" to D. Schuurmans, and the listed collaborators. $75K total to D. Schuurmans (Apr 1/05-Mar 31/09).
Shie Mannor, McGill University	Joint research with Cs. Szepesvari on efficient reinforcement learning.
Russ Greiner, University of Alberta	Joint research with Cs.Szepesvari on active learning.
Douglas Wiens, University of Alberta	Joint research with Cs. Szepesvari on active learning.

International	
Participants	Nature of Collaboration
Shalabh Bhatnagar, Indian Institute of Science, Bangalore	Joint research with Rich Sutton and Csaba Szepesvari on stochastic approximation
Jean-Yves Audibert, CERTIS, France	Joint research with Csaba Szepesvari on variance estimation in bandit problems.
Remi Munos, INRIA, Lille, France	Joint research with Csaba Szepesvari on batch reinforcement learning.
Peter Auer, Ronald Ortner, University of Leoben, Austria,	Joint research with Csaba Szepesvari on efficient exploration.
Andras Antos, MTA SZTAKI, Hungary	Joint research with Csaba Szepesvari on active learning.
Olivier Cappé, LTCI	Joint research with Csaba Szepesvari on restless bandits.
Tiberio Caetano, NICTA	Joint research with Dale Schuurmans on tractable approximation architecture for computer vision.
Li Cheng, TTI-Chicago	Joint research with Dale Schuurmans on discriminative structured learning models for computer vision.
Yuhong Guo, Temple University	Joint research with Dale Schuurmans on scalable optimization algorithms in machine learning.
Matthew Taylor and Fei Sha, University of Southern California	Joint research with Michael Bowling on transfer in reinforcement learning.
E. James Kehoe, University of New South Wales, Australia	Joint research with Rich Sutton, Elliot Ludvig, and Eric Verbeek on the relationship between reinforcement learning and learning in animals.
Gerry Tesauro, T.J Watson Research Centre	Joint research with David Silver on computer go research.
Joel Veness, ANU Sydney	Joint research with David Silver on bootstrapping from game tree search

Industrial	
Participants	Nature of Collaboration

7. GRADUATES

Name	Degree	Research topic	Current Position
Mohammad Ghavamzadeh	Post Doctoral Fellow	Bayesian reinforcement learning, value function approximation in reinforcement learning	Research Scientist at INRIA, France, September 2008
Daniel Lizotte	PhD	Practical bayesian optimization	Postdoc, University of Michigan
Qin Wang	PhD	Learning structured classifiers for statistical dependency parsing	Research Scientist at AT&T Interactive, CA
Yasin Abbasi-Yadkori	MSc	Forced-exploration based algorithms for playing in bandits with large action sets	PhD, University of Alberta
Nolan Bard	MSc	Using state estimation for dynamic agent modeling	PhD, University of Alberta
Varun Grover	MSc	Active learning and its application to heteroscedastic problems	.NET Technical Analyst, Accenture, Edmonton, AB.
Volodymyr Mnih	MSc	Efficient stopping rules	PhD, University of Toronto
James Neufeld	MSc	Autonomous outdoor navigation and goal finding	PhD, University of Alberta
Masoud Shahamiri	MSc	Reinforcement learning in environments with independent delayed-science dynamics	Unknown

8. INTELLECTUAL PROPERTY

None.

9. PUBLICATIONS

REFEREED JOURNAL PUBLICATIONS

F. Balci, E.A. Ludvig, J. Gibson, B.D. Allen, K.M. Frank, B.J. Kapustinski, T.E. Fedolak, and D. Brunner, "Pharmacological Manipulations of Interval Timing on the Peak Procedure in Male C3H Mice," *Psychopharmacology 201*, 67-80, 2008.

T. Degris, O. Sigaud, P. Wuillemin, "Apprentissage par reinforcement factorisè pour le comportment de personages non joueurs," *Revue d'Intelligence Ariticielle 23*, pp. 221-251, 2009.

A. M. Farahmand, M. N. Ahmadabadi, B. N. Araabi, C. Lucas, "Interaction of Culture-based Learning and Cooperative Co-evolution and its Application to Automatic Behavior-based System Design," *IEEE Transactions of Evolutionary Computation*, 2009.

E. J. Kehoe, K. N. Olsen, E. A. Ludvig, R. S. Sutton, "Scalar Timing Varies with Response Magnitude in Classical Conditioning of the Nictitating Membrane Response of the Rabbit (Oryctolagus cuniculus)," *Behavioral Neuroscience 123*, pp. 212-217, 2009.

E. A. Ludvig, R. S. Sutton, E. J. Kehoe, "Stimulus Representation and the Timing of Reward-prediction Errors in Models of the Dopamine System," *Neural computation*, 20, 3034-3054, 2008.

H. R. Maei, K. Zaslavsky, C. M. Teixeira, P. W. Frankland, "What is the Most Sensitive Measure of Water Maze Probe Test Performance?" *Frontiers in Integrative Neuroscience 3*, Mar. 2009.

B. Póczos, A. Lőrincz, "Identification of Recurrent Neural Networks by Bayesian Interrogation Techniques," *Journal of Machine Learning Research 10*, pp. 515-554, 2009.

HIGHLY REFEREED ARCHIVAL CONFERENCE PROCEEDINGS

A. Antos, V. Grover, and Cs. Szepesvári, "Active Learning in Multi-Armed Bandits," *Proc. of the 19th Int'l Conf. on Algorithmic Learning Theory* (ALT 2008), 2008. 67% acceptance

G. Bartok, C. Szepesvari, and S. Zilles, "Active Learning of Group-Structured Environments," *Proc. of the 19th Int'l Conf. on Algorithmic Learning Theory* (ALT 2008), 2008. 67% acceptance

S. Bubeck, R. Munos, G. Stoltz and Cs. Szepesvari, "Online Optimization in X-Armed Bandits," *Proc. Conf. on Neural Information Processing Systems* (NIPS 08), 2009. 24% acceptance

Z. Cai, Y. Shi, M. Song, R. Goebel, and G. Lin, "Smoothing Blemished Gene Expression Microarray Data via Missing Value Imputation," *Proc. of the 30th Annual Int'l Conf. of the IEEE Engineering in Medicine and Biology Society* (IEEE EMBC 08), pp. 5688-5691. 20% acceptance

A. M. Farahmand, M. Ghavamzadeh, C. Szepesvari, and S. Mannor, "Regularized Policy Iteration," *Proc. Conf. on Neural Information Processing Systems* (NIPS 08), 2009. 24% acceptance

A. M. Farahmand, A. Shademan, M. Jagersand, C. Szepesvari, "Model-based and Model-free Reinforcement Learning for Visual Servoing," *Proc. of Int'l Conf. on Robotics and Automation* (ICRA 09), IEEE, 2009. 43% acceptance

M. Johanson, M. Bowling, "Data Biased Robust Counter Strategies," *Proc. of the 12th Int'l Conf. on Artificial Intelligence and Statistics* (AISTATS 09), 2009. 30% acceptance

S. Kirshner, and B. Póczos, "ICA and ISA Using Schweizer-Wolff Measure of Dependence," *Proc. Int'l Conf. on Machine Learning* (ICML 08), 2008, pp. 464-471. 27% acceptance

Y. Li, Cs. Szepesvari, D. Schuurmans, "Learning Exercise Policies for American Options," *Proc. of the 12th Int'l Conf. of Artificial Intelligence and Statistics* (AISTATS 09), 2009. 30% acceptance

E.A. Ludvig, R.S. Sutton, E. Verbeek, and J. Kehoe, "A Computational Model of Hippocampal Function in Trace Conditioning," *Proc. Conf. on Neural Information Processing Systems* (NIPS 08), 2009, 24% acceptance

D. Schnizlein, M. Bowling, D. Szafron, "Probabilistic State Translation in Extensive Games with Large Action Sets," *Proc. of the 21st Int'l Joint Conf. on Artificial Intelligence* (IJCAI 09), 2009. 26% acceptance

R. S. Sutton, Cs. Szepesvari, H. R. Maei, "A Convergent $O(n)$ Algorithm for Off-policy Temporal-difference Learning with Linear Function Approximation," *Proc. Conf. on Neural Information Processing Systems* (NIPS 08), 2009. 24% acceptance

K. Waugh, D. Schnizlein, M. Bowling, D. Szafron, "Abstaction Pathology in Extensive Games," *Proc. of the 8th Int'l Joint Conf. on Autonomous Agents and Multi-Agent Systems* (AAMAS 09), 2009. To appear. 22% acceptance

M. White, M. Bowling, "Learning a Value Analysis Tool For Agent Evaluation," *Proc. of the 21st Int'l Joint Conf. on Artificial Intelligence* (IJCAI 09), 2009. 26% acceptance

L. Xu, W. Li, D. Schuurmans, "Fast Normalized Cut with Linear Constraints," *Proc. of the Int'l Conf. of Computer Vision and Pattern Recognition* (CVPR 09), IEEE Press, 2009. 26% acceptance

M. Yang, Y. Li, D. Schuurmans, "Dual Temporal Difference Learning," *Proc. of the Int'l Conf. on Artificial Intelligence and Statistics* (AISTATS 09), 2009. 30% acceptance

OTHER CONFERENCE PUBLICATIONS

M. Cutumisu, D. Szafron, M.Bowling, R.S. Sutton, "Agent Learning using Action-Dependent Learning Rates in Computer Role-Playing Games," *Proc. of the 4th Conf. on Artificial Intelligence and Interactive Digital Entertainment* (AIIDE 08), pp. 22-29, 2008.

Y. Guo and D. Schuurmans, "Efficient Global Optimization for Exponential Family PCA and Low-rank Matrix Factorization," *Proc. of the Allerton Conf. on Communication, Control, and Computing* (Allerton 08), 2008.

A. M. Farahmand, M. Ghavamzadeh, C. Szepesvari, S. Mannor, "Regularized Fitted Q-iteration for Planning in Continuous-Space Markovian Decision Problems," *Proc. of the American Control Conf.* (ACC 09), 2009.

Y. Li and D. Schuurmans, "Learning and Exercise Policy for American Options on Real Data," *International Symposium on Financial Engineering and Risk Management 2008* (FERM 08).

Y. Li and D. Schuurmans, "Policy Iteration for Learning an Exercise Policy for American Options," *European Workshop on Reinforcement Learning* (EWRL 08).

Y. Li, L. Cheng, and D. Schuurmans, "Inference of the Structural Credit Risk Model using MLE," *Proc. of the IEEE Symposium on Computational Intelligence for Financial Engineering* (CIFEr 09), 2009. 30% acceptance

E.A. Ludvig and A. Koop, "Learning to Generalize Through Predictive Representations: A Computational Model of Mediated Conditioning," *From Animal to Animats 10: Proc. of Simulation of Adaptive Behavior* (SAB 08), 2008, pp. 342-351. 46% acceptance

A. Shademan, A. M. Farahmand, M. Jägersand, "Towards Learning Robotic Reaching and Pointin: An Uncalibrated Visual Servoing Approach," *Proc. of the 6th Canadian Conf. on Computer and Robot Vision* (CRV 09), 2009.

Y. Shi, Z. Cai, G. Lin, D. Schuurmans, "Linear Coherent Bi-cluster Discovery via Line Detection and Sample Majority Voting," *Proc. of the 3rd Annual Int'l Conf. on Combinatorial Optimization and Applications* (COCOA 2009), 2009.

SPECIAL/INVITED PRESENTATIONS

Date	**Title**	**Venue**
May 2008	Active Learning in Multi-Armed Bandits (Cs. Szepesvari)	ParisTech Machine Learning Seminar, ParisTech, Ecole des Mines de Paris
May 2008	Active Learning in Multi-Armed Bandits (Cs. Szepesvari)	Lille Machine Learning Seminar, INRIA Lille
May 2008	Mind and Time: A View of Constructivist Reinforcement Learning (R. Sutton)	Eighth European Workshop on Reinforcement Learning, Lille France
May 2008	Bayesian Reinforcement Learning (M. Ghavamzadeh)	INRIA Lille, France

May 2008	Bayesian Reinforcement Learning (M. Ghavamzadeh)	Systems and Modeling Research Unit, University of Liege, Belgium
May 2008	Stimulus Representation in Reinforcement-Learning Models of Brain and Behavior (E. Ludvig)	Computational Memory Lab at the University of Alberta
June 2008	Learning Models of the World with Temporal-difference Networks (M. Bellemare)	McGill University
July 2008	Policy Iteration for Learning an Exercise Policy for American Options (Y. Li)	Department of Computer Science, Fudan University, Shanghai, China
Sept. 2008	Statistical Learning Theory and Sequential Decision Making, six lectures (Cs. Szepesvari)	Eleventh Machine Learning Summer School, Ile de Re France
Sept. 2008	Reinforcement Learning and Knowledge Representation, six lectures (R. Sutton)	Eleventh Machine Learning Summer School, Ile de Re France
Oct. 2008	AI After Dark: Computers Playing Poker (M. Bowling)	University of Washington
Oct. 2008	Stimulus Representation in Reinforcement-Learning Models of Brain and Behavior (E. A. Ludvig)	Riken Brain Science Institute, Tokyo Japan
Dec. 2008	Real-time Prediction Machines: How Animals Learn to Anticipate the Future (E. A Ludvig)	Reed College, Portland Oregon, USA
Dec. 2008	The RLAI Robotic Simulator (M. Gendron-Bellemare)	McGill University, Montreal Canada
Jan. 2009	Symposium on Autonomous Systems (Cs. Szepesvari)	MPI for Biological Cybernetics, Tübingen Germany
Feb. 2009	Unsupervised and Semi-supervised Machine Learning (D. Schuurmans)	Machine Learning Summer School, Australian National University
Feb. 2009	Greedy Importance Sampling (D. Schuurmans)	Machine Learning Program, NICTA Canberra Australia
Feb. 2009	New Temporal-Difference Methods Based on Gradient Descent (R. Sutton)	University of Southern California
Feb. 2009	Real-time Prediction Machines: How Animals Learn to Anticipate the Future (E. A Ludvig)	Texas Christian University, Fort Worth Texas, USA
Mar. 2009	The Polaris Computer Poker Program (M. Bowling)	Alberta Gaming Research Institute's 8th Annual conference

AWARDS

Dale Schuurmans renewed his Canada Research Chair, Tier 2.

Dale Schuurmans received an NSERC Discovery Accelerator award.

Dale Schuurmans received the 2008 departmental "Celebration of Excellence" award for Faculty Service.

Anna Koop received the 2008 departmental "Celebration of Excellence" award for Outstanding MSc thesis.

Kevin Waugh received the 2008 departmental "Celebration of Excellence" award for Outstanding Research Achievement as an MSc Student.

David Schnizlein received the 2008 departmental "Celebration of Excellence" award for Outstanding Academic Achievement as an MSc student.

THESES

Daniel Lizotte (PhD) "Practical Bayesian Optimization," July 18, 2008.

Qin Iris Wang (PhD) "Learning Structured Classifiers for Statistical Dependency Parsing," May 2008.

Yasin Abbasi-Yadkori (MSc) "Forced-Exploration Based Algorithms for Playing in Bandits with Large Action Sets," January 30, 2009.

Nolan Bard (MSc) "Using State Estimation for Dynamic Agent modeling," June 2008.

Varun Grover (MSc) "Active Learning and its Applications to Heteroscedastic Problems," January 9, 2009.

Volodymyr Mnih (MSc) "Efficient Stopping Rules," August 25, 2008.

James Neufeld (MSc), "Autonomous Outdoor Navigation and Goal Finding," June 2008.

Masoud Shahamiri (MSc) "Reinforcement Learning in Environments with Independent Delayed-Sense Dynamics," July 2008.

10. OUTREACH

Michael Bowling presented for the High School Internship Program (HIP) Lunch & Learn series. "AI After Dark: Computers Playing Poker", July 24, 2008.

Michael Johanson was a volunteer for the W.P. Wagner High School FIRST program, which introduces high school students to science and technology fields.

Richard Sutton served on the scientific advisory board of the USA's NSF Temporal Dynamics of Learning Center (spread across a half-dozen USA universities and the University of Victoria).

Richard Sutton has helped organize a new special meeting to celebrate the multi-disciplinary aspects of reinforcement learning with Doina Precup and Elliot Ludvig. This meeting will be held on June 18-19/09 in Montreal in conjunction with ICML and UAI and COLT. The web site is http://msrl09.rl-community.org. AICML is supporting this event, as is RLAI. Primary funding is coming from MITACS. This meeting will probably be attended by hundreds of researchers from around the world.

Richard Sutton and Doina Precup organized the Fourth Annual Bellairs Workshop on Reinforcement Learning, attended by 14 RLAI project members and 24 other researchers from around the world, in Holetown, Barbados on April 13-17, 2009.

11. FINANCIAL REPORT

Financial Report at March 31, 2009

REVENUES	
iCORE	650,000.00
Interest	4,088.51
Total Revenue	654,088.51
EXPENSES	
Faculty salaries and benefits	186,025.16
Undergraduate salaries and benefits	
Post Docs and Graduate students	102,550.48
Support staff salaries and benefits	65,717.59
Communications, Outreach, Travel	17,369.63
Equipment	1,826.66
General Operations	17,528.78
TOTAL EXPENSES	391,018.30
BALANCE at April 1, 2008	109,455.43
TOTAL REVENUE (Apr 1/08-Mar 31/09)	654,088.51
TOTAL EXPENSES (Apr 1/08 – Mar 31/09)	391,018.30
BALANCE at March 31, 2009	372,525.64