AITF ANNUAL REPORT 2012

DR. RICHARD SUTTON

REINFORCEMENT LEARNING AND ARTIFICIAL INTELLIGENCE

AITF ANNUAL REPORT MARCH 31, 2012

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 deployment, 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 50 members, 35 of whom are graduate students and, of those, 14 of which are recipients of major scholarships. The output of the research program was particularly strong this year, with 39 papers published or accepted for publication in highly refereed archival venues during the reporting period. Two PhD and two 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 here 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 machine's 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 for doing this.

Highlights of the research program this year include: 1) learning 1000 off-policy predictions in parallel at 10 updates/second on a mobile robot, 2) new actor-critic algorithms for off-policy and continuous-action problems, and 3) fielding an autonomous mobile robot in the Courage Centre at the Glenrose Rehabilitation Hospital.

2. RESEARCH PROGRAM OVERVIEW

In the project's proposal, research was divided into three main target areas. The first is extensions of existing 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. We discuss highlights of our research towards each of these target areas below.

In the area of extending core reinforcement-learning algorithms, our main innovations this year have been in the area of "actor-critic" algorithms, in which a separate decision-making structure is learned in addition to the value functions ubiquitous in reinforcement learning. Our attention has been drawn to this older and, in recent decades, less popular category of algorithms because of our forays in the last few years into robotics. In this and other real-time applications, the separate decision-making structure enables actions to be made with less computation and thus with a faster response time. In this year we have extended existing actor-critic algorithms to make them more practical and to give them a clearer theoretical foundation, particularly in the case of systems, like robots, which choose their control signals from a continuous space. We have also extended our work with gradient algorithms for temporal-difference learning into this area, to produce and the first actor-critic methods that are stable with function approximation under off-policy training.

Off-policy learning has been key to our efforts over the last few years to scale reinforcement learning methods to address the more ambitious goals of AI. Off-policy learning means learning about a possible way of behaving (a *policy*) from following it for a period of time, perhaps only for a fraction of a second, without following it to completion. Moment to moment, the actions an AI agent selects can be seen as parts of many policies, but at most one will be followed to completion. If the agent can learn in parallel about all of the policies, then it can learn vastly more than if it is restricted, as it is with classical "on-policy" temporal-difference algorithms, to learning about only the one that is followed to completion. Last year we showed that thousands of predictions about a robot's sensory data can be learned in real time, but only under on-policy training—that is, only if all of the predictions were about the one policy being followed. And in the years prior to that we have developed the world's first scalable off-policy algorithms. This year we have begun to bring it all together by showing off-policy learning at scale and in real time on our robots. In our latest experiments we have demonstrated off-policy learning of predictions about one thousand different policies, based on thousands of features, online and in real-time at ten updates per second using a mobile robot and a laptop computer. We see this as one of the essential raw capabilities to enable massive learning about the world by a powerful future AI agent.

3. RESEARCH PROJECTS

This section describes in more detail a few of the research projects making up the research program.

Advances in Actor-Critic Reinforcement Learning

Model-free reinforcement-learning algorithms can be divided into two categories. The first category of methods, called *action-value methods*, including Q-learning, Sarsa, and our recently developed GQ algorithms, all work by estimating the value of state-action pairs, that is, by learning to predict the expected long-term reward obtainable starting from each action in each state. The second category of methods, called *actor-critic methods*, instead learn two objects called an *actor* and a *critic*. The critic's job is to estimate the expected long-term reward starting from each state (but not from each action in each state as in an action-value method). The actor's job is to directly store a decision-making rule, a mapping from states to actions, also known as a *policy*. Because the policy is stored directly (rather than being computed from action values), actor-critic methods be easily initialized with human-designed policies. Actor-critic methods are more natural for large or continuous action spaces because they do not require a search over all actions to find the one with highest value. Other important advantages are that actor-critic methods can take actions stochastically with specific learned probabilities, that they can act with minimal computation (and thus react very quickly), and that they make good models of biological learning systems. The intuitive appeal and other advantages of actor-critic methods make them the methods of choice in a variety of reinforcement learning contexts.

Actor-critic reinforcement learning algorithms were first explored in the earliest years of modern reinforcement learning (i.e., in the 1980s) but were then largely overshadowed by the simpler and better understood action-value methods. In 2000, the theory of actor-critic methods received a boost when they came to be understood as approximately following the gradient of the average reward-per-step with respect to the policy (actor) parameters. Several new algorithms were then proposed and analyzed for their convergence properties (including as part of RLAI research) but were rarely used in practice because they were seen as too complex and untried.

This year we have begun a deliberate attempt to eliminate all the perceived drawbacks of actor-critic methods. Our first step was to extend the modern theory of actor-critic methods to apply to continuous action spaces. Our second step was to extend actor-critic methods to the case of off-policy training. This has proved more challenging than expected. We were able to take advantage of our previously developed gradient-TD methods for use in an off-policy critic, but new theory was required to adapt the actor to the off-policy case. Although we obtained a convergence result, we continue to explore other theoretical routes toward an off-policy actor critic. Our third step was to apply the tuning-free step-size adaptation methods developed by the RLAI project to automatically set the parameters of actor-critic methods. Actor-critic methods typically have at least one parameter more than action-value methods that must be set by the user. Our

preliminary experiments suggest that this drawback of actor-critic methods can be eliminated by applying our "Autostep" algorithm.

Off-policy Nexting in Robots

The term "nexting" has been used by psychologists to refer to the propensity of people and many other animals to continually predict what will happen next in an immediate, local, and personal sense. When we hear a melody we predict what the next note will be or when the next downbeat will occur, and are surprised and interested when our predictions are disconfirmed. When nexting, an individual may be predicting many or all of their sensory inputs, and at multiple time scales. When we read, for example, it seems likely that we next at the letter, word, and sentence levels, each involving a substantially different time scale. Nexting can be seen as the most basic kind of prediction, preceding and possibly underlying all the others. That people and a wide variety of animals learn and make simple predictions at a range of short time scales was established so long ago in psychology that it is known as "classical conditioning." Animals seem wired to learn the predictive relationships of their world. To be able to next is to have a basic kind of knowledge about how the world works in interaction with your body. To be able to learn to next—to notice any disconfirmed predictions and continually adjust your nexting—is to be aware of your world in a significant way. To build a robot that can do both of these things is a natural goal, which we have pursued in the RLAI project.

Last year we built a robot that learned and made thousands of predictions at four time scales. However, a limitation of that work was that the predictions were all conditional on the robol continuing to behave exactly as it had during training. Predictions are much more powerful if they can be contingent on multiple, hypothetical ways of behaving. Learning these requires more powerful, off-policy learning algorithms such as the gradient-TD methods we have developed in previous years. These algorithms are more complicated to use, and off-policy learning is more complicated to measure, so it was appropriate that we explored on-policy nexting first.

This year we have finally brought nexting and off-policy learning together in a big way. We have conducted extensive experiments with learning to predict the consequences of many different ways of behaving from a single stream of experience (i.e., with off-policy nexting). All of our predictions are learned in parallel, in real time, and make full use of the experience stream. We have explored off-policy nexting on three robotic platforms and for predicting myoelectric signals from a human patient. Our most extensive experiments have been on the Critterbot, a custom robot designed and built from scratch as part of RLAI research. In our largest experiment on the Critterbot, predictions for one thousand different policies were learned and made online at ten updates per second using a standard Macbook Pro laptop computer.

One of the biggest challenges in off-policy learning is that the learning cannot easily be seen or measured. The robot is following at most one of the policies; predictions about that policy can be assessed for accuracy by comparing them to what happens, but all the other predictions have nothing to be compared with. One approach to addressing this is to

interrupt the normal behavior of the robot and run all of the policies in sequence with learning turned off, resetting the robots state after each such test trial. Of course, this explicit-test-trial approach is not feasible in practice because it negates all the advantages of learning about the policies in parallel if you have to test them in sequence. To address this, we have developed an alternate approach in which a measure of learning progress, based on the theory of gradient-TD methods, is computed *online*, without interrupting normal behavior. We have validated the online measure empirically by comparing it to the results of explicit test trials. The online measure introduces a slight delay, but otherwise performs similarly to the test trials and, unlike them, is feasible and cheap to use in practice.

Fielding a Mobile Robot at the Glenrose Hospital

In January 2011, the RLAI project began a new initiative, funded by MITACS, the Alberta Ingenuity Centre for Machine Learning, and the Glenrose Rehabilitation Hospital, to field an interactive mobile robot at the new Building Trades of Alberta Courage Centre. The Courage Center is a new facility at the Glenrose intended to showcase the use of new technologies in rehabilitative medicine, such as robotics, virtual reality, simulation, and bionics. Robotic devices are becoming more affordable and more widely used in physiotherapy. Robots can significantly enhance conventional therapy by compensating for physical disabilities of the patient, by providing additional feedback to the patient and therapist, and by motivating the patient to continue the therapy by themselves (e.g., at home). Robotic technology includes mobility aids for locomotion and navigation as well as prostheses and other manipulation aids. For RLAI research, deploying a robot in the Glenrose Hospital provides a challenging testbed and an opportunity to investigate technological, social, and medical challenges in the application of reinforcement learning.

The objective of this project is to explore social interaction between robots and members of the public, including patients. This year, we fielded a robot that is mobile, autonomous, and minimally supervised in a public space at the entrance to the Courage Centre. The physical robot is a research variant of a successful commercial line of robot vacuum cleaners developed by the iRobot company. Using our software, the robot operated autonomously for multiple months and exhibited a multi-faceted behaviour. It darted around, expressed surprise by playing music when exposed to novel sensory experiences, autonomously returned to the dock to recharge, and remained quiet on evenings, nights and weekends to minimize disruption. Over a period of time, several patients in the hospital brought chairs and sat to watch the robot. The robot required minimal oversight from the Glenrose staff. Only a few previous autonomous mobile robots have been deployed long term in public spaces; and of these, none has displayed this range of behaviour with minimal supervision in a hospital. Developing trust and confidence with the patients and staff in the hospital is both an important accomplishment and a prerequisite for deploying more complex robot behaviours in the future. This infrastructure provides the necessary foundation for investigating how people interact with the robot over long durations and how the robot might learn which behaviours are most engaging for people.

Reinforcement Learning Methods in Assistive Robotics

Assistive biomedical devices augment the abilities of amputees and other patients with lost physical or cognitive function. Specifically, these devices replace abilities lost due to traumatic injury, disease, aging, or congenital complications. This project focuses on one representative class of assistive robots: powered artificial limbs. Powered prostheses monitor electromyographic (EMG) signals produced by muscle tissue in a patient's body, and use these signals to control the movement of a robotic appendage with one or more controllable dimensions. Such devices are tightly coupled to a human user, with control processes that operate at high frequency and over extended periods of time. Despite the potential for improved function with myoelectric control, many patients reject the use of powered artificial limbs. Recent needs assessment surveys of prosthesis users point out three key principal reasons for patient rejection: lack of intuitive control, insufficient functionality, and insufficient feedback from the myoelectric device. RLAI research aims to remove these barriers and increase the independence and ability of amputees through the use of adaptive machine learning techniques.

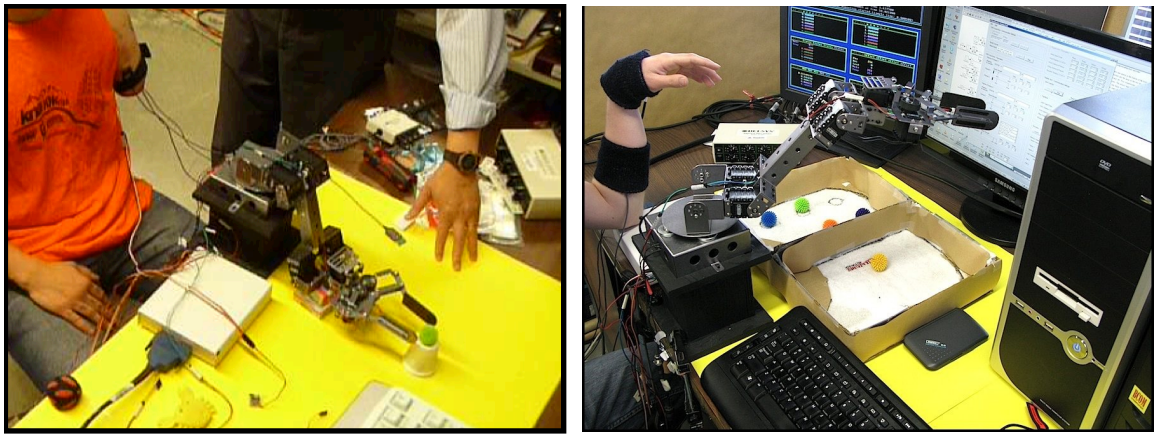

Figure 1: New real-time machine learning techniques are applied to human-robot interactions involving amputee (left) and non-amputee (right) participants; reinforcement learning and adaptive prediction techniques are able to streamline and enhance the human control of a complex robotic system.

RLAI assistive robotics work is a joint venture between the RLAI and the AICML, the Glenrose Rehabilitation Hospital (GRH), the UofA Faculty of Rehabilitation Medicine, and the UofA Dept. of Mechanical Engineering. Over the past two years, the project has explored fundamental and applied methods for real-time prediction, adaptive control, and direct human-machine interaction. Key outcomes to date include a number of international publications and presentations, repeat media coverage, award recognition, and a demonstration of patient-ready machine learning technology. The project has also developed a framework for translating developed technologies into clinical application. Ongoing work addresses three aims: fundamental algorithm development, practical translation of methods to patient use, and convincing demonstrations of clinical benefit.

These are important steps toward intuitive prosthetic control and improved quality of life for amputees.

4. OBJECTIVES FOR NEXT YEAR

With regard to core reinforcement learning algorithms, we plan to combine our new actor-critic algorithms with the Autostep step-size adaptation algorithm, and make other refinements, to produce a completely parameter-free actor-critic algorithm. We will also investigate other ideas for extending the theory of Autostep and of off-policy actor-critic reinforcement learning. Finally, work will continue with novel gradient-TD algorithms to maximize the efficiency of off-policy learning via action-value methods.

With regard to the larger ambitions of AI, work will proceed on several fronts. One of the advantages of off-policy learning is that it frees behavior from having to match the target policies, enabling it to be scupted to other purposes. For example, the psychological notion of curiosity is that we sculpt our behavior so as to maximize our learning. Using off-policy learning it is possible to implement a computational version of curiosity. We can apply our parallel "nexting" architecture to learn a great many predictions and measure the total learning progress over all predictions. This measure can then be used as a reward for the learning agent that controls the robot's behavior. The result, we suspect, will be a robot that behaves so as to maximize the amount of learning it does, in other words, that exhibits a computational form of curiosity.

One of the longest-standing challenges in AI has been the automated construction of appropriate representations from and for sensorimotor interaction. Progress in the field has been slow, but we hope to be more successful by leveraging off of the rich stream of data provided by our robots' sensorimotor experience. This data provides many subtasks in the form of sensory signals to predict and control, all of which are due to the same physical system. A multi-task setting such as our nexting architecture may be key to making progress on representation learning. In past work the construction of tasks and their interrelatedness has been largely artificial and therefore suspect. Our robots, we believe, will make it easy to automatically create large numbers of naturally interrelated prediction and control tasks.

With regard to the mobile robot in the Courage Centre at the Glenrose Hospital, our plan for next year is to add a responsive component to the robot. Currently, people respond to the robot, primarily by watching it, but the robot's behavior is not influenced by anything that people do. For true interaction, of course, the influences must go both ways. Our first step will be to formally quantify the effect of the robot on people. We will add a Kinect sensor to the installation at the Glenrose that enables us to locate people and identify their poses and movements. The second step will be to allow simple forms of interaction, in which people influence the robot by touching it and rewarding it, perhaps analogous to the idea of a "petting zoo." We also seek to improve patient control of prosthetic arms (a form of robot) using user-generated reward signals in a similar way.

5. RESEARCH TEAM MEMBERS AND CONTRIBUTIONS

a. Team Leader

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

b. Faculty Team Members

Michael Bowling	Faculty team member	AICML, NSERC
Dale Schuurmans	Faculty team member	AICML, CRC II Chair, NSERC, MITACS
Csaba Szepesvári	Faculty team member	AICML, NSERC
Martin Mueller	Associated faculty member	DARPA, NSERC

c. Post Doctoral Fellows & Research Associates

Name	Role	Awards / Special Info
András György	Research Associate	Computer and Research Institute of the Hungarian Academy of Sciences
Thomas Degris-Dard	Post doctoral fellow	MITACS
Hamid Reza Maei	Post doctoral fellow	NSERC
Joseph Modayil	Post doctoral fellow	
David Pal	Post doctoral fellow	
Patrick Pilarski	Post doctoral fellow	NSERC until October 2011
Harm van Seijen	Post doctoral fellow	
Joel Veness	Post doctoral fellow	
Xinhua Zhang	Post doctoral fellow	

d. PhD Candidates

Name	Role	Scholarships / Awards / Special Info
Yasin Abbasi-Yadkori	PhD candidate	
Arash Afkanpour	PhD candidate	AICT ICT
Gabor Balazs	PhD candidate	
Nolan Bard	PhD candidate	AITF ICT
Neil Birch	PhD candidate	AIF, NSERC PGS
Gabor Bartok	PhD candidate	
Katherine Chen	PhD candidate	
Kenneth Dwyer	PhD candidate	
Amir massoud Farahmand	PhD candidate	
Alireza Farhangfar	PhD candidate	AICT ICT
Marc Gendron-Bellemare	PhD candidate	AITF ICT
Ruitong Huang	PhD candidate	
Michael Johanson	PhD candidate	AICT ICT
Anna Koop	PhD candidate	NSERC PGS
Marc Lanctot	PhD candidate	AICT ICT
Ashique Mahmood	PhD candidate	Provost Doctoral Entrance Award
Gergely Neu	PhD candidate	
Aslan Ozlem	PhD candidate	Provost Doctoral Entrance Award
Bernardo Ávila Pires	PhD candidate	University of Alberta Doctoral Recruitment Scholarship
James Neufeld	PhD candidate	AITF ICT
Chris Rayner	PhD candidate	AITF ICT, NSERC PGS
Yi Shi	PhD candidate	
Adam White	PhD candidate	
Martha White	PhD candidate	AIF
Min Yang	PhD candidate	
Hengshuai Yao	PhD candidate	
Yaoliang Yu	PhD candidate	

e. MSc Candidates

Name	Role	Scholarships / Awards / Special Info
Mohammad Ajallooeian	MSc candidate	
Hao Cheng	MSc candidate	
Christian Denk	MSc candidate	Technical University of Munich
Leah Hackman	MSc candidate	
Pooria Joulani	MSc candidate	
Michael Joya	MSc candidate	
Parisa Mazrooei	MSc candidate	
Roshan Shariff	MSc candidate	
Mostafa Vafadost	MSc candidate	

f. Other Members

Name	Role
Beverly Balaski	Administrative assistant
Lori Troop	Program administrator
Alexandra Kearney	High school summer student, July-August 2011
Deon Nicholas	High school summer student, July-August 2011 High school summer student, July-August 2011
Allan Wu	High school summer student, July-August 2011
Kaiven Zhou	High school summer student, July-August 2011

g. Visitors

Name	Institution
Chris Archibald	Stanford University
András Antos	Computer and Research Institute of the Hungarian Academy of Sciences
Nicholas Butko	University California San Diego
Éva Czabarka	University of South Carolina
András György	Computer and Research Institute of the Hungarian Academy of Sciences
Simon Haykin	McMaster University
Lihong Li	Yahoo! Research

Dominik Meyer	Technische Universität München
Harm van Seijen	University of Amsterdam
Yevgeny Seldin	Max Planck Institute
Ted Selker	CyLab Mobility Research Center
Hao Shen	Technische Universität München
László Székely	

6. GRADUATES

Name	Degree	Research topic	Current Position
Hamid Reza Maei	PhD	Gradient Temporal-Difference Learning Algorithms	Postdoctoral Fellow Stanford University
Amir massoud Farahmand	PhD	Regularization in Reinforcement Learning	Postdoctoral Fellow, McGill University
Bernardo Ãvila Pires	MSc	Statistical Analysis of L1-regularized Linear Estimation with Applications	PhD candidate in RLAI team
MahdiehSadat "Neda" Mirian	MSc	A Computational Model of Learning from Replayed Experience in Spatial Navigation	Software engineer, Microsoft

7. COLLABORATIONS

Provincial	
Participants	Nature of Collaboration
Alberta Ingenuity Center for Machine Learning	R. Sutton, D. Schuurmans, Cs. and Szepesvári M. Bowling are among the ten principle investigators for this center at the University of Alberta. Total annual funding for AICML is $2M.
Patrick M. Pilarski and Richard Sutton (RLAI), Linda Pilarski (U of A, Faculty of Medicine), and Carina Debes Marun (Cross Cancer Institute).	Collaboration with researchers from the AHFMR Team Microfluidics, Alberta Innovates Health Solutions, University of Alberta, and Cross Cancer Institute to explore intelligent biomedical image analysis methods to facilitate rapid lab-on-a-chip diagnostics. Part of the technology transfer focused "nanoBridge Research Grant RES-NAN-07-G10: Development of Fluorescence In Situ Hybridization (FISH) Platform, Chip, and Analysis Software" ($145K)

Patrick Pilarski and Wilsun Xu, Department of Electrical and Computer Engineering, University of Alberta	Preliminary collaboration in the application of reinforcement learning to specific challenges in power systems engineering, both on a household scale and on the scale of provincial power generation and distribution.
Patrick Pilarski and Richard Sutton (RLAI), and Wilsun Xu (Department of Electrical and Computer Engineering, University of Alberta).	Preliminary collaboration in the application of reinforcement learning to specific challenges in power systems engineering, both on a household scale and on the scale of provincial power generation and distribution.
Patrick Pilarski and Richard Sutton (RLAI), Martin Ferguson-Pell (Dean, Faculty of Rehabilitation Medicine, University of Alberta), Liping Qi (University of Alberta, Rehabilitation Medicine), Simon Grange (University of Calgary / University of Alberta / Alberta Health Services).	Collaboration between the RLAI and the Rehabilitation Robotics Sandbox Laboratory (Rehabilitation Medicine, U. Alberta); this collaboration focuses on using new machine learning methods to predict fatigue in wheelchair users and enable novel muscle stimulation paradigms. Collaboration between the RLAI and the Rehabilitation Robotics Sandbox Laboratory (Rehabilitation Medicine, U. Alberta); this collaboration focuses on using new machine learning methods to predict fatigue in wheelchair users and enable novel muscle stimulation paradigms.
Thomas Degris-Dard, MITACS, Glenrose Rehabilitation Hospital	A collaboration to introduce an interactive mobile robot component to the new Building Trades of Alberta Courage Center in the Glenrose Rehabilitation Hospital

National	
Participants	Nature of Collaboration
D. Precup, McGill University	Richard SuttonCo-organized the 7th Bellairs Workshop on Reinforcement Learning
Joelle Pineau, McGill University	Joint research with Csaba Szepesvári on the theory of reinforcement learning.
Daniel Lizotte, University of Waterloo	Joint research with Michael Bowling on reinforcement learning for medical applications

International	
Participants	Nature of Collaboration
Tiberio Caetano, NICTA	Joint research with Dale Schuurmans on machine learning in structured-output prediction problems.
Yuhong Guo, Temple University	Joint research with Dale Schuurmans on machine learning in bioinformatics.
Andras Gyorgy, MTA SZTAKE, Hungary	Joint research with Csaba Szepesvári on the theory of online learning in MDPs
Andras Antos, MTA SZTAKI, Hungary	Joint research with Csaba Szepesvári on active learning.
Olivier Cappé, CNRS	Joint research with Csaba Szepesvári on scaling up online learning algorithms and applications to telecommunication
Rong Zheng, University of Houston	Joint research with Csaba Szepesvári on application of online learning techniques in networking
Gilles Stoltz , Ecole normale supérieure, France	Joint research with Csaba Szepesvári on bandit problems with large action sets.
Susan Murphy, University of Michigan	Joint research with Michael Bowling on reinforcement learning for medical applications
Martin Zinkevich, Google Research	Joint research with Michael Bowling on online learning methods
Kevin Waugh, Carnegie Mellon University	Joint research with Michael Bowling on massive-scale computational game theory
Nathan Sturtevant, University of Denver	Joint research with Michael Bowling on multiplayer game theory and automating heuristic construction for search
Patrick M. Pilarski (RLAI) and Sophia Adamia (Dana Farber Cancer Institute, Harvard Medical School, USA).	Investigation of cancer-related genetic markers in patients. Collaboration involving machine learning and biomedical data mining methods.
Patrick Pilarski and Richard Sutton (RLAI), Jason P. Carey (Department of Mechanical Engineering, University of Alberta), Farbod Fahimi (University of Alabma at Huntsville, USA), Michael R. Dawson, Jacqueline S. Hebert, K. Ming	"Reinforcement Learning for Adaptive Prosthetics;" this collaboration investigates the use of reinforcement learning and real-time machine learning to enable adaptive, intuitive control of myoelectric prostheses and other assistive robotic devices.

Chan (Glenrose Rehabilitation Hospital).	

8. INTELLECTUAL PROPERTY

Intellectual Property	Status	Short Description
PATENTS	none	
LICENSES		
Rebel Entertainment		Michael Bowling entered into an option agreement to license the media rights to Polaris, our computer program for playing heads-up limit Texas hold'em
Spinoff Companies	none	

9. PUBLICATIONS

REFEREED JOURNAL PUBLICATIONS

A. Antos, G. Bartok, D. Pal, and Cs. Szepesvari, "Toward a Classification of Finite Partial-Monitoring Games," *Theoretical Computer Science*, to appear.

A. Afkanpour, Cs. Szepesvari, and M. Bowling, "Alignment Based Kernel Learning with a Continuous Set of Base Kernels," *Machine Learning Journal (MLJ)*, to appear.

D. Lizotte, R. Greiner, and D. Schuurmans, "An Experimental Methodology for Response Surface Optimization Methods," *Journal of Global Optimization (JOGO)*, Jun. 2011.

A. M. Farahmand, and Cs. Szepesvári, "Regularized Least-Squares Regression with Beta-mixing Input Processes," *Journal of Statistical Planning and Inference* (JSPI), 2011.

A. M. Farahmand and Cs. Szepesvári, "Model Selection in Reinforcement Learning," *Machine Learning Journal* (MLJ), 2011.

M. Ponsen, S. de Jong, and M. Lanctot, "Computing Approximate Nash Equilibria and Robust Best-Responses Using Sampling," *Journal of Artificial Intelligence Research (JAIR)*, Dec. 2011, pp. 562-605.

Y. Shi, M. Hasan, Z. Cai, G. Lin, and D. Schuurmans, "Linear Coherent Bi-clustering via Beam Searching and Sample Set Clustering, " *Discrete Mathematics, Algorithms and Applications*, Dec. 2011.

D. Silver, R. S. Sutton, M. Müller, "Temporal-difference Search in Computer Go," *Machine Learning 87*(2):183-219, 2012.

Cs. Szepesvari, and A. Farahmand, "Model Selection in Reinforcement Learning," *Machine Learning Journal (MLJ)*, Jun. 2011.

Cs. Szepesvari, and A. Farahmand, "Regularized Least-Squares Regression with beta-mixing Input Processes," *Journal of Statistical Planning and Inference (JSPI)*, Feb. 2012.

S. Wang, D. Schuurmans, Y. Zhao, "The Latent Maximum Entropy Principle," *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*.

S. Wang, S. Wang, Li Cheng, R. Greiner, D. Schuurmans, "Exploiting Syntactic, Semantic and Lexical Regularities in Language Modeling via Directed Markov Random Fields, " *Computational Intelligence*, Feb. 2012.

X. Zhang, S. V. N. Vishwanathan, and A. Saha, "Smoothing Multivariate Performance Measure," *Journal of Machine Learning Research (JMLR)*.

HIGHLY REFEREED ARCHIVAL CONFERENCE PROCEEDINGS

Y. Abbasi-Yadkori, and Cs. Szepesvari, "Regret Bounds for the Adaptive Control of Linear Quadratic Systems," *Proc. 24th Annual Conf. on Learning Theory* (COLT 2011), Jun. 2011, pp. 1-26.

Y.Abbasi-Yadkori, D. Pal, and Cs. Szepesvari, "Improved Algorithms for Linear Stochastic Bandits," *Proc. 25th Annual Conf. Neural Information Processing Systems* (NIPS 2011), Dec. 2011.

Y. Abbasi-Yadkori, D. Pal, and Cs Szepesvari, "Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits." *Proc. 15th Int'l Conf. on Artificial Intelligence and Statistics* (AISTATS 2012), Apr. 2012, to appear.

G. Bartok, D. Pal, and Cs. Szepesvari, "Minimax Regret of Finite Partial-Monitoring Games in Stochastic Environments," *Proc. 24th Annual Conf. on Learning* Theory (COLT 2011), Jun. 2011, Journal of Machine Learning Research - Proceedings Track 19: 133-154.

T. Degris, M. White, and R. S. Sutton, "Linear Off-Policy Actor Critic," *Proc. 29th Int'l Conf. on Machine Learning* (ICML 2012), July 2012.

T. Degris, P. M. Pilarski, R. S. Sutton, and A. Mahmood, "Tuning-Free Step-Size Adaptation," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2012), Mar. 2012.

T. Degris, P.M. Pilarski, and R. S. Sutton, "Model-Free Reinforcement Learning with Continuous Action in Practice," *2012 American Control Conference* (ACC), to appear. A version in french will also appear in the *Proceedings of the 7th Journées Francophones Planification, Décision, et Apprentissage pour la Conduite de Systèmes*, 2012.

M. Gendron-Bellemare, J. Veness, and M. Bowling, "Investigating Contingency Awareness using Atari 2600 Games," *Proc. 26th Conf. on Artificial Intelligence* (AAAI 2012), July 2012, 25% acceptance.

R. Gibson, M. Lanctot, and N. Burch, "Generalized Sampling and Variance in Counterfactual Regret Minimization," *Proc. 26th Conf. on Artificial Intelligence* (AAAI 2012), July 2012, 25% acceptance.

Y. Guo, and D. Schuurmans, "Adaptive Large Margin Training for Multilabel Classification," *Proc. 25th Conf. on Artificial Intelligence* (AAAI 2011), Aug. 2011, 25% acceptance.

K. Hajebi, Y. Abbasi-Yadkori, H. Shahbazi, and H. Zhang, "Fast Approximate Nearest-Neighbor Search with k-Nearest Neighbor Graph," *Proc. 20th Int'l Joint Conf. on Artificial Intelligence* (IJCAI 2011).

M. Johanson, K. Waugh, M. Bowling, and M. Zinkevich, "Accelerating Best Response Calculation in Large Extensive Games," *Proc. 20th Int'l Joint Conf. on Artificial Intelligence* (IJCAI 2011).

M. Johanson, N. Bard, M. Lanctot, and R. Gibson, "Efficient Nash Equilibrium Approximation through Monte Carlo Counterfactual Regret Minimization," *Proc. 11th Int'l Conf. on Autonomous Agents and Multi-Agent Systems* (AAMAS 2012), Jun. 2012, to appear, 20% acceptance.

W. Li, and D. Schuurmans, "Modular Community Detection in Networks," *Proc. 20th Int'l Joint Conf. on Artificial Intelligence* (IJCAI 2011), Jul. 2011.

D. Pal, B. Poczos, and Cs. Szepesvari, "Estimation of Renyi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs," *Proc. 25th Annual Conf. Neural Information Processing Systems* (NIPS 2011), Dec. 2011, to appear, 24% acceptance.

P. M. Pilarski, M. R. Dawson, T. Degris, J. P. Carey, and R. S. Sutton, "Dynamic Switching and Real-time Machine Learning for Improved Human Control of Assistive Biomedical Robots," *Proc. 4th IEEE Int'l Conf. on Biomedical Robotics and Biomechatronics (*BioRob*)*, to appear.

P.M. Pilarski, M. R. Dawson, T. Degris, F. Fahimi, J. P. Carey, and R. S. Sutton, "Online Human Training of a Myoelectric Prosthesis Controller via Actor-Critic Reinforcement Learning," *Proc. IEEE International Conference on Rehabilitation Robotics (ICORR)*, Jun. 2011, pp. 134.

C. Rayner, M. Bowling, and N. Sturtevant, "Euclidean Heuristic Optimization," *Proc. 25th Conf. on Artificial Intelligence*, (AAAI 2011), Aug. 2011, pp. 81-86, 25% acceptance.

C. Szepesvari, S. Filipi, A. Garivier, and O. Cappé, "Parametric Bandits: The Generalized Linear Case," *Proc. 25th Annual Conf. Neural Information Processing Systems* (NIPS 2011), Dec. 2011, to appear, 24% acceptance.

J.Veness, M. Lanctot, and M. Bowling, "Variance Reduction in Monte Carlo Tree Search," *Proc. 25th Annual Conf. Neural Information Processing Systems* (NIPS 2011), Dec. 2011, to appear, 24% acceptance.

M. White, Y. Yu, X. Zhang, and D. Schuurmans, "Multi-view Subspace Learning: An Efficient Convex Approach," *Proc. 26th Conference on Artificial* Intelligence (AAAI 2012), July 2012, to appear.

M. White, and D. Schuurmans, "Generalized Optimal Reverse Prediction," *Proc. 15th Int'l Conf. on Artificial Intelligence and Statistics* (AISTATS 2012), to appear, 33% acceptance.

Y. Yu, and Cs. Szepesvari, "Analysis of Kernel Mean Matching under Covariate Shift," *Proc. 29th Int'l Conf. on Machine Learning* (ICML 2012), Feb. 2012.

Y. Yu, and D. Schuurmans, "Rank/Norm Regularization with Closed-Form Solutions: Application to Subspace Clustering," *Proc. 27th Conf. on Uncertainty in Artificial Intelligence (UAI 2011)*, Jul. 2011, 34% acceptance.

X. Zhang, Y. Yu, M. White, R. Huang, and D. Schuurmans, "Convex Sparse Coding, Subspace Learning, and Semi-Supervised Extensions," *Proc. 25th Conf. on Artificial Intelligence* (AAAI 2011), Aug. 2011, 25% acceptance.

X. Zhang, A. Saha, and S. V. N. Vishwanathan, "Smoothing Multivariate Performance Measures," *Proc. 27th Conference on Uncertainty in Artificial Intelligence* (UAI 2011), Jul. 2011, 34% acceptance.

X. Zhang, Ankan Saha, and S. V. N. Vishwanathan, "Accelerated Training of Max-Margin Markov Networks with Kernels," *Proc. 18th Int'l Conf. on Algorithmic Learning Theory (ALT)*, Oct. 2011.

BOOKS and CHAPTERS

Y. Shi, J. Zhou, D. Wishart, G. Lin, *Protein Contact Order Prediction: The Most Recent Update*, Dec. 2011, edited by Yi Pan, Wiley.

L. Cheng, M. Gong, D. Schuurmans, and T. Caelli, "Real-time Discriminative Background Subtraction," in *IEEE Transactions on Image Processing,* http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=83 (digital publication)

OTHER CONFERENCE AND WORKSHOP PROCEEDINGS

T. Degris, J. Modayil, "Scaling-up Knowledge for a Cognizant Robot." In notes of the *AAAI Spring Symposium on Designing Intelligent Robots: Reintegrating AI*, 2012.

S. Legg, and J. Veness, "An Approximation of the Universal Intelligence Measure," *Solomonoff 85th Memorial Conf.* Nov. 2011, 50% acceptance.

J. Modayil, A. White, R. S. Sutton, "Multi-timescale Nexting in a Reinforcement Learning Robot." *Proceedings of the 2012 Conference on Simulation of Adaptive Behavior*, to appear.

R. S. Sutton, "Beyond reward: The problem of Knowledge and Data" (extended abstract). *Proceedings of the 21st International Conference on Inductive Logic Programming*. Windsor Great Park, United Kingdom, 2012.

G. Velikic, J. Modayil, M.Thomsen, M. Bocko, and A. Pentland, "Predicting the Near-future Impact of Daily Activities on Heart Rate for at-risk Populations," *Proc. 13th Int'l Conf. On E-Health Networking, Application & Services* (IEEE Healthcom 2011), Jun. 2011.

MAGAZINE ARTICLES

P.M. Pilarski, "Intellient Artificial Limbs," *Alberta ICT Magazine*, 2nd Edition (April), pp. 28–30, 2012.

P.M. Pilarski, "The Nuts and Bolts of Learning," *Alberta ICT Magazine*, 2nd Edition (April), pp. 31, 2012.

SPECIAL/INVITED PRESENTATIONS

Person	Title	Venue
R. Sutton	Beyond Reward: The Problem of Knowledge and Data	21st International Conference on Inductive Logic Programming, Windsor Great Park, United Kingdom
R. Sutton	Learning About Sensorimotor Data	25th Annual Conference on Neural Information Processing Systems, Granada, Spain
R. Sutton	Learning About Sensorimotor Data	21st Annual Conference of the Japanese Neural Network Society, Okinawa, Japan.
R. Sutton	The Imperative of Learning from Interaction	AAAI Workshop on Lifelong learning from Sensorimotor Experience
D. Schuurmans	Convex Sparse Coding, Subspace Learning, and Semi-supervised Extensions	University College London
D. Schuurmans	Convex Feature Discovery	ICDM 2011 Workshop: Optimization for Emerging Data Mining Problems
M. Bowling	Abstraction with an Adversary – workshop talk	25th Conference on Artificial Intelligence, San Francisco, California
M. Bowling	AI After Dark: Computers Playing Poker	Blizzard Entertainment
T. Degris	Reinforcement Learning in the Courage Centre	Glenrose Rehabilitation Hospital, Edmonton, Alberta
T. Degris	Real-time Learning of Abstract Knowledge from Low-level Signals	Flowers Group, INRIA Bordeaux
T. Degris	An Encouraging Mobile Robot in the Glenrose Rehabilitation Hospital	7th International Congress on Industrial and Applied Mathematics
J. Modayil	Scaling Up Knowledge for a Cognizant Robot	AAAI Spring Symposium on Designing Intelligent Robots
J. Veness	Variance Reduction in Monte-Carlo Tree Search	University College London
J. Veness	The Switch Distribution	Guest Lecture for Advanced AI class at University of New South Wales
P. M. Pilarski	Reinforcement Learning in the Courage Centre & Reinforcement Learning for Adaptive Prosthetics	Prosthetics, Orthotics, and Seating Department, Glenrose Rehabilitation Hospital, Edmonton, Alberta
P. M. Pilarski	New Methods for Real-time Machine Learning and their Application to Assistive Medical Robotics	International Seminar Series on Biomedical Engineering, Department of Medical System Engineering, Chiba University, Japan

SPECIAL/INVITED PRESENTATIONS (continued)

Person	Title	Venue
X. Zhang	Accelerated Training of Max-Margin Markov Networks with Kernels	International Conference on Algorithmic Learning Theory
X. Zhang	Accelerated Optimization for Machine Learning: A Smoothing Approach	Aalto University, Finland

AWARDS

Michael Bowling: Honourable Mention for the Alan Blizzard Award.

Patrick Pilarksi: First Prize (Oral Presentation Award) at the 2011 University of Alberta PDF Research Day.

Patrick Pilarski: Best Conference Paper Award Finalist at the 2011 IEEE International Conference on Rehabilitation Robotics.

Patrick Pilarski: Best Poster Award (Postdoctoral Category) at the Alberta Innovates Technology Futures Summit 2011.

THESES

Hamid Reza Maei, PhD, "Gradient Temporal-Difference Learning Algorithms," August 29, 2011.

Amir Massoud Farahmand, PhD, "Regularization in Reinforcement Learning," September 14, 2011.

Bernardo Ãvila Pires, MSc, "Statistical Analysis of L1-regularized Linear Estimation with Applications," September 30, 2011.

Neda Mirian HosseinAbadi, MSc, "A Computational Model of Learning from Replayed Experience in Spatial Navigation," October 30, 2011

10. OUTREACH

Csaba Szepesvari and Richard Sutton each supervised two high school students in July and August as part of the High School Internship Program at the Department of Computing Science.

Patrick Pilarski presented a hands-on robotics and artificial intelligence demonstration and lecture for students in the Department of Computing Science Summer Camps (Riveting Robotics) for two groups of students, ages 11-13 and ages 14-16.

Patrick Pilarski gave a tour and demonstration of the AICML Adaptive Prosthetics Project to senior government ministers and deputy ministers, university administration, and members of the media. Part of the Edmonton Health Clinic Academy grand opening event, and Rehabilitation Robotics Laboratory opening tour.

Patrick Pilarski and Adam White hosted a group of Grade 4 students visiting the RLAI lab as a special field trip for their schools new Robotics Club. Students were from Windsor park Elementary School Robotics Club. Students taught a robot to move through reinforcement learning, and explored hands-on ideas in intelligent artificial limbs.

Patrick Pilarski hosted a demonstration and interactive session for critically or terminally ill youth in connection with Make a Wish Foundation, Canada.

Martha White hosted a panel for grade ten students from WP Wagner school to discuss studying and research at the Department of Computing Science.

Katherine Chen hosted an object-oriented programming exercise using Alice for grade nine girls at the Women in Technology program.