iCORE ANNUAL REPORT 2010

iCORE CPE GRANT CPE45

DR. RICHARD SUTTON

REINFORCEMENT LEARNING AND ARTIFICIAL INTELLIGENCE

ICORE ANNUAL REPORT MARCH 31, 2010

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 50 members, 33 of whom are graduate students and, of those, 17 of which are recipients of major scholarships. The output of the research program has remained strong, with 38 papers published or accepted for publication in highly-refereed archival venues during the reporting period. One PhD and four 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) the development of the first temporal-difference algorithms for learning with nonlinear function approximation (solving a major open problem in reinforcement learning), 2) our experimental sensor-rich robot, the *Critterbot*, becoming operational, beginning to drive the research program, and running for two days without human intervention, and 3) organizing the Multi-disciplinary Symposium on Reinforcement Learning, held June 18-19 in Montreal, which brought together top scientists from around the world in artificial intelligence, control theory, neuroscience, operations research, and industry.

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 with which we are 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 were limited in key ways, particularity when combined with the ability to form approximate solutions, as is essential in large-scale applications. Last year we introduced a new family of TD algorithms that enable reinforcement learning to be applied efficiently to large-scale applications in which the policy being learned is not followed during learning. This kind of learning, called *off-policy learning*, is important in many applications for safety (to avoid poor performance during learning) and for reaching toward the ambitious goals of AI. Our approach, based on gradient descent in a new objective function, solved the problem of off-policy for the case of linear function approximators. It seemed then that our approach was the beginning of a breakthrough in reinforcement learning.

A year later this still seems to be the case. Over this year we have obtained much greater experience with the new family of algorithms and extended their theory in several ways. Much of this work is very technical, but the overall picture of a significant breakthrough remains. The most important additional advance this year was an extension from linear to nonlinear function approximators, such as neural networks. Prior to our work this year, there were no proven-stable TD methods for nonlinear function approximation—not for off-policy learning nor for the easier case of on-policy learning. Our proof of convergence in both cases solves a major problem in reinforcement learning that has been open since the field began. With our results, combining TD learning with function approximation becomes nearly as straightforward as combining conventional supervised learning with function approximation.

These new results have implications for approximate solutions not only to learning problems, but also to all problems where conventional dynamic programming is applied. A good way to understand the potential impact is to see our work as closing in on lifting what is widely known in control theory and related fields as *the curse of dimensionality*. "The curse of dimensionality" is a reference to the fact that the size of a state space grows exponentially in the number of state variables. This is widely seen as a major problem for

all approaches to optimal control. The curse is so intractable that many consider it to be an inevitable fact of nature, something that does not admit a solution. This may be so, but maybe not; the field of machine learning has made great progress on high-dimensional problems in conventional supervised settings. These advances have not extended to control and TD-prediction settings because of the problems off-policy learning to which we alluded earlier. Our results over the last few years are beginning to make approximation in these settings straightforward for the first time. If this can be done we may eventually be able to lift the curse of dimensionality in a general sense.

An important remaining step is to extend our results to *control* algorithms. Our results so far have been limited to *prediction*. That was appropriate, as the prediction case is the simplest in which previous TD methods have failed; the prediction case had to be solved before addressing control. In principle, now that we have stable TD prediction methods with function approximation, extensions to control should be easier. So far we have found that this is indeed the case, although the extensions are not easy or immediate. As this year ends we have obtained the first preliminary results for an off-policy control method with (linear) function approximation. We call this algorithm linear GQ, for a linear, gradient version of Q-learning, the most popular existing TD control method (which is not stable with linear function approximation). Further work will be needed on gradient-based TD control methods in the years to come.

As the new gradient TD algorithms prove effective, they become key to our 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 do not have this problem. This year we have begun to explicitly design a very general algorithm for universal prediction learning, called GQ(λ), and begun to apply it as a knowledge representation algorithm.

In the area of applications, we have continued to make good progress with Computer Go, with standardized open-source software, and, particularly, with the *Critterbot*. The Critterbot is a sensor-rich robotic platform designed and built by the project in the previous two years (see Figure 1). Applying our methods to a physically realized robot is a useful complement to our studies in computational worlds where everything can be completely controlled and understood. A real robot forces the project 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. This year the Critterbot has provided a useful focal point for the group activity, a single system that can be addressed in multiple ways while reflecting a consensus about objectives. The Critterbot is regularly used by half a dozen project members, and has run for two days at a time without human intervention.

3. RESEARCH PROJECTS

This section describes in more detail some of the research projects making up the research program

The Critterbot and Zephyr

As mentioned above, the *Critterbot* is a robot designed and built by the project that has become a focal point of project research, particularly with regard to the goal of relating sensori-motor data to higher-level knowledge of the world. The Critterbot is shown in Figure 1. This small mobile robot is 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 temperature, current, and velocity of the motors driving the three wheels. The Critterbot receives digital sensory data from over 50 sources every 10 milliseconds, and generates digital motor commands for the three wheels and the 12 display LEDs at the same rate. The Critterbot's mind, the focus of project research, also has many state variables updated up to 100 times a second. To monitor what is going on in real time we have built an flexible display tool, called *Zephyr*, shown in Figure 2. Zephyr enables us to easily pick out any set of Critterbot signals and display them graphically at any temporal resolution. For example, the upper-right portion of Figure 2 shows the temporal evolution of 5800 weights as they are being learned by one of our algorithms.

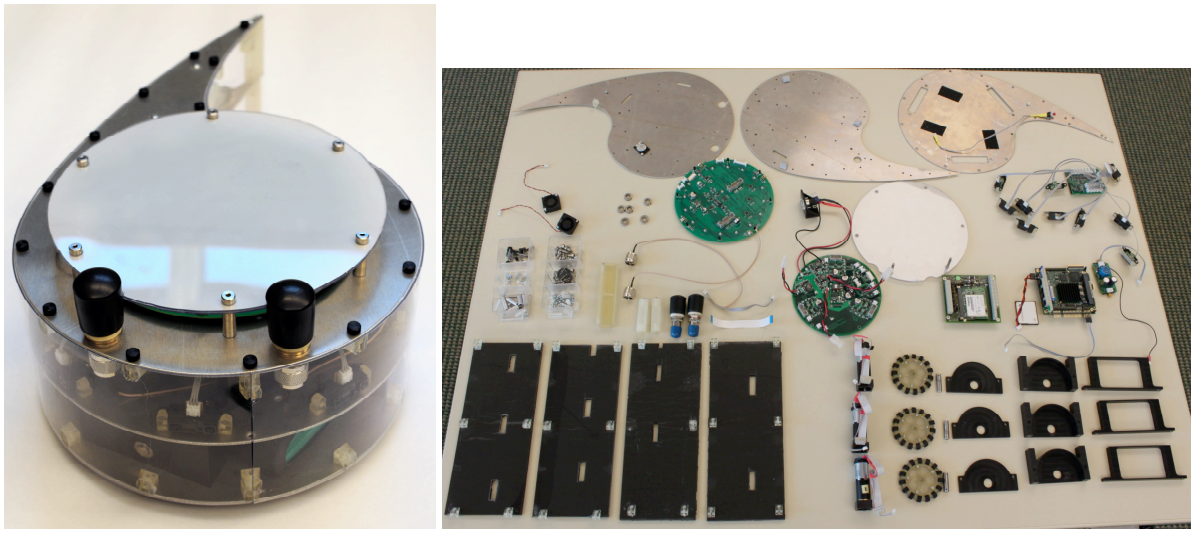

Figure 1. The Critterbot assembled (left) and disassembled (right).

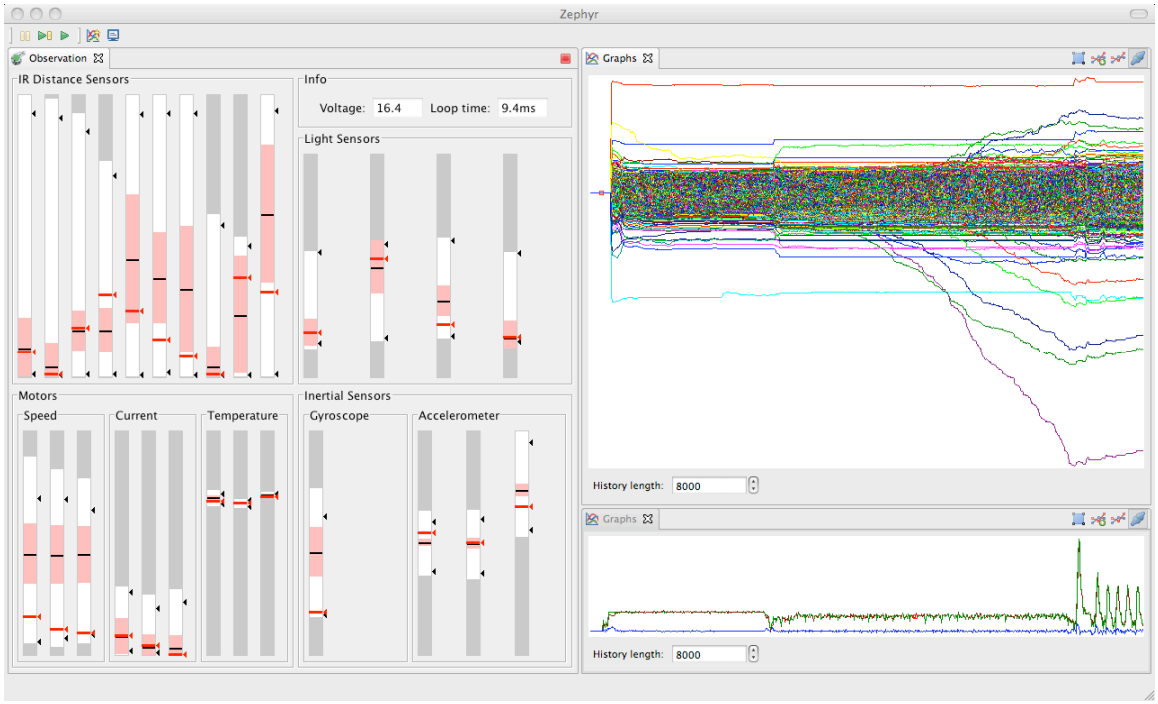

Figure 2. Screenshot of the Zephyr display tool showing Critterbot sensori-motor data (left) and the progress of learning in 5800 weights (right).

Parameter-free Step-size Adaptation

Many of the learning algorithms used in the project have parameters that must be tuned manually for good performance; we are always looking for ways that they can be set automatically. Foremost among these parameters is the step-size parameter of stochastic gradient-descent algorithms. Several methods have been proposed for automatically setting step-size parameters, but unfortunately all of them have at least one parameter of their own, and this meta-parameter must generally be tuned manually to the particular problem, thereby limiting the benefit of these methods. This year we have begun a subproject to develop a superior step-size algorithm with no parameters or meta-parameters, that is, that can be applied with no domain knowledge other than what is needed to formulate a problem as gradient descent.

So far we have shown that all previous step-size methods involve at least one meta-parameter and that there is no single setting of the meta-parameters that produces acceptable performance on all tasks. We are focusing in particular on a family of methods related to an algorithm known as K1, previously developed by Sutton, which performs best of the existing methods but is still sensitive to the meta-parameter. We have developed a new algorithm that we call Normalized K1 and which performs well, without tuning, over a much wider range of problems. So far we have tested Normalized K1 on a range of artificial problems. Next we will stress it more severely by applying it to data from the Critterbot.

Monte-Carlo Simulation Balancing (Computer Go)

Monte-Carlo search methods have proven successful in automating high-performance play in many two-player games, including Poker, Scrabble, and the ancient oriental game of Go which we have used as a test platform for our learning algorithms. In Monte-Carlo search, many games of self-play are simulated, using a *simulation policy* to select moves for both players. A record of wins and loses in the simulated games is kept for each possible move, and move selections are gradually changed to favor the moves that lead to the most wins. As you might imagine, the quality of the moves eventually found in this way are highly dependent on the simulation policy. However, as we have shown in past work, the relationship is not a direct one: simulation policies that are stronger in terms of winning more games are not necessarily the best to use in self-play. More important than the absolute strength of the simulation policy is that outcomes with it be representative of the outcomes of a strong policy. A good simulation policy can make many errors as long as they are *balanced*, favoring neither side of most positions. Thus, in learning a good simulation policy, one could take as an objective not strong play, but balanced play in simulation, or *simulation balancing*.

This year we have come to better understand the phenomena of simulation balancing, and explored new algorithms for learning a balanced simulation policy. One algorithm uses policy-gradient reinforcement methods and a linear function approximator with weights for one hundred simple patterns. Such relatively simple approximators are preferred in the simulation policy so that many games can be simulated quickly in self-play, thus obtaining better Monte-Carlo estimates. Using a simulation policy learned with the objective of optimizing balance, and a simple Monte-Carlo search, we were able to obtain significantly improved simulation policies and similar overall performance to that of a sophisticated Go engine.

Reinforcement Learning in the Brain

In the last 15 years, there has been a flourishing of research into the neural basis of reinforcement learning, drawing together insights and findings from psychology, computer science, and neuroscience. This remarkable confluence of three fields has yielded a growing framework that begins to explain how animals and humans learn to make decisions in real time. The confluence was catalyzed by the discovery of a close correspondence between the behavior of dopamine neurons in classical conditioning tasks and the prediction error in the temporal-difference (TD) algorithm from reinforcement learning.

The RLAI project is focused primarily on the artificial-intelligence aspects of reinforcement learning, but we have participated in some of the related fields as well. One way we have done this in past years is by developing models of classical conditioning that match the neuroscientific and behavioral data better, and by generating additional targeted behavioral data in collaboration with Prof. Kehoe of the University of New South Wales in Australia. Last June we organized (with Prof. Precup of McGill) a special *Multidisciplinary Symposium on Reinforcement Learning*, held in conjunction

with several international computer science conferences in Montreal. We were able to attract top scientists from neuroscience, operations research, and industry as invited speakers and about 150 scientists as participants. The program, videos, and information on this event are available at http://msrl09.rl-community.org. Finally, this year we have written an introduction to the psychology and neuroscience of reinforcement learning targeted at computer scientists and other people without a prior background in these fields.

4. OBJECTIVES FOR NEXT YEAR

Next year we plan to extend our work with gradient TD learning algorithms into forms suitable for control, including both value-function forms, such as GQ, and actor-critic (policy gradient) forms. We will be implementing these algorithms for use in off-policy learning on the Critterbot, towards our goal of learning large amounts of detailed sensori-motor knowledge about the Critterbot's interaction with its body and environment.

A new focus area for our research in the coming year will be in learning state representations from sensori-motor interaction. The problem of discovering useful representations for learning systems has been recognized as crucial for half a century. Progress has been slow, but we hope to be more successful by leveraging off of the rich stream of data provided by the Critterbot's sensori-motor 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 this 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 robot, we believe, will make it easy to automatically create large numbers of naturally interrelated prediction and control tasks.

Finally, we are planning two new application-oriented subprojects in collaboration with the Glenrose Rehabilitation Hospital and the Alberta Innovation Center for Machine Learning. The goals of the first subproject are to conduct research in adaptive game-play using machine learning methods and to develop a computer game to be used by patients at the Glenrose Rehabilitation Hospital that adapts itself to the abilities of the patients. Computer games are commonly used in physiotherapy sessions to motivate patients and make their therapy more interesting. A limitation of current commercial games is that the level of difficulty is often poorly matched to the abilities of the patient. If the game is too hard, then the patient may be discouraged and not complete their therapy, particularly if a therapist is not present. We plan to develop an implementation of a computer game that adapts its difficulty to the patient, to the therapist's specifications, or both, so as to maintain the interest and participation of the patient.

The goal of the second subproject is to develop reinforcement learning algorithms that can help increase limb-deficient patients' ability to control prosthetic devices, while at the same time removing the need for frequent manual adjustments by patients and physiotherapists. Multi-function myoelectric prosthetics are devices that monitor

electrical signals produced by muscle tissue in limb-deficient patients and use these signals to control the movement of a multiple-actuator robotic appendage (Fig. 1a). However, one problem for recent amputees is the transition to their new prosthetic device—myoelectric control can be challenging, and often requires repeated calibration of the artificial limb by patients and physiotherapists; we hope to automate this with software learning algorithms. This project will be in collaboration with researchers at the University of Alberta in the department of mechanical engineering, who have been pursuing the mechanical and electrical design of an inexpensive robotic arm and Myoelectric Training Tool (Fig. 1b) to help prepare patients for powered prosthetics. We plan to work with them to develop the software and learning algorithms for this tool.

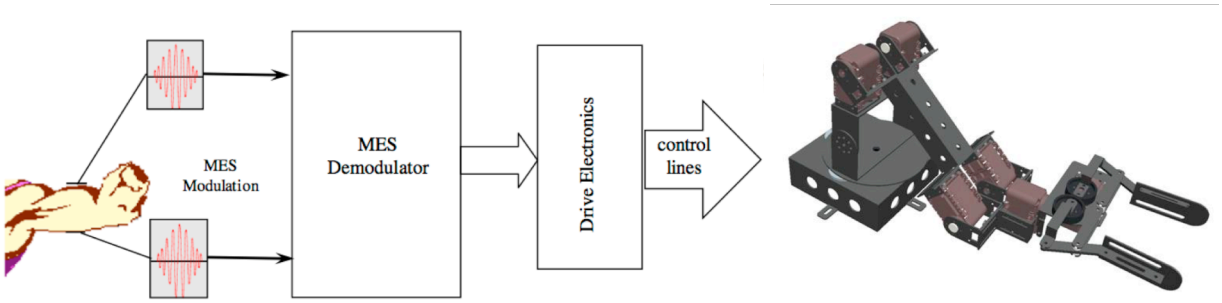

Fig. 1: a) Example of a traditional myoelectric processing pathway, where myoelectric signals are recorded, interpreted, and used to generate control commands for a prosthetic device. (Image: P. Parker et al. / Journal of Electromyography and Kinesiology 16 (2006) 541–548.). b) The AX-12 Smart Arm—mechanical component of the Myoelectric Training Tool. (Image: http://www.ualberta.ca/~mrd1/training_arm/)

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, MITACS, NSERC
Csaba Szepesvári	Faculty team member	AICML, AIF New Faculty, NSERC

c. Post Doctoral Fellows & Research Associates

Name	Role	Awards / Special Info
Elliot Ludvig	Research associate	
Thomas Degris-Dard	Post doctoral fellow	
Wenye Li	Post doctoral fellow	NSERC until Sept 2009. Now an Assistant Professor at Macao Polytechnic Institute
Yuxi Li	Post doctoral fellow	Associate Professor at University of Electronic Science and Technology China, Feb 2010
Joseph Modayil	Post doctoral fellow	
David Pal	Post doctoral fellow	
Patrick Pilarski	Post doctoral fellow	NSERC
Barnabás Póczos	Post doctoral fellow	
Istvan Szita	Post doctoral fellow	
Eric Wiewiora	Post doctoral fellow	Engineering Scientist at Intellisis, Feb 2010
Linli Xu	Post doctoral fellow	Associate professor at University of Electronic Science and Technology China, Feb 2010

d. PhD Candidates

Name	Role	Scholarships / Awards / Special Info
Yasin Abbasi-Yadkori	PhD candidate	
Arash Afkanpour	PhD candidate	iCORE ICT
Gabor Balazs	PhD candidate	
Nolan Bard	PhD candidate	iCORE ICT, NSERC PGS, President's Doctoral Prize of Distinction
Gabor Bartok	PhD candidate	
Katherine Chen	PhD candidate	FGSR Provost Doctoral Entrance Award
Kenneth Dwyer	PhD candidate	iCORE ICT, NSERC, Queen Elizabeth II, Walter H. Johns
Amir massoud Farahmand	PhD candidate	
Alireza Farhangfar	PhD candidate	iCORE ICT
Marc Gendron-Bellemare	PhD candidate	iCORE ICT

Michael Johanson	PhD candidate	iCORE ICT
Anna Koop	PhD candidate	AIF, iCORE ICT, NSERC PGS
Marc Lanctot	PhD candidate	AIF, iCORE ICT, NSERC PGS
Hamid Reza Maei	PhD candidate	iCORE ICT
James Neufeld	PhD candidate	iCORE ICT, NSERC PGS, Walter H. Johns
Chris Rayner	PhD candidate	iCORE ICT, NSERC PGS
Yi Shi	PhD candidate	
David Silver	PhD candidate	iCORE ICT, Senior Postdoctoral Associate at UCL, January 2010
Brian Tanner	PhD candidate	AIF
Adam White	PhD candidate	NSERC CGS
Min Yang	PhD candidate	
Hengshuai Yao	PhD candidate	
Yaoliang Yu	PhD candidate	

e. MSc Candidates

Name	**Role**	**Scholarships / Awards / Special Info**
Kate Davison	MSc candidate	
Michael Delp	MSc candidate	iCORE
Leah Hackman	MSc candidate	Queen Elizabeth II
Michael Joya	MSc candidate	
Ashique Mahmood	MSc candidate	MSc Academic Achievement Award
Yavar Nadaf	MSc candidate	Graduated March 2010
Bernardo Ávila Pires	MSc candidate	
David Schnizlein	MSc candidate	Graduated July 2009, now working at Sandia National Labs in Albuquerque New Mexico.
Kevin Waugh	MSc candidate	Graduated Sept 2009, now a PhD student at Carnegie Mellon University.
Martha White	MSc candidate	AIF, NSERC PGS. Graduated Jan 2010. Now a PhD student with group.

f. Other Members

Name	Role
Akiko Green	Administrative assistant
Mike Sokolsky	Robot engineer until September 2009
Lori Troop	Program administrator
Jesse Tucker	High school summer student, July-August 2009
David Szepesvari	High school summer student, July-August 2009

g. Visitors

Name	Institution
Shai Ben-David	University of Waterloo
Arnaud Doucet	University of British Colombia
Andras Gyorgy	MTA SZTAKI
Sarah Fillipi	TELECOM Paris Tech
Mohammad Ghavamzadeh	INRIA
Gholamreza Haffari	Simon Fraser University
David Kirckpatrick	University of British Columbia
James Kehoe	University of New South Wales
Alexander Klementiev	UIUC
Alessandro Lazaric	INRIA
Joseph Modayil	University of Rochester
Cosmin Paduraru	McGill University
Ron Parr	Duke University
Erik Talvitie	University of Michigan
Shankar Vembu	Fraunhofer IAIS
Joel William Veness	University of New South Wales
Ivor Wai-Hung Tsang	Nanyang Technological University
Weiwei Wang	Nanjing University
Yang Wang	Simon Fraser University
James Witnauer	State University of New York
Xinhua Zhang	ANU/NICTA
Zhihua Zhou	Nanjing University

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 (PI), 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 31/10).
Shai Ben-David, University of Waterloo Yoshua Bengio, McGill University Hugh Chipman, University of Waterloo Dale Schuurmans, University of Alberta Richard Sutton, University of Alberta Pascal Vincent, Montreal University	MITACS Grant, "Statistical Learning of Complex Data with Complex Distributions" to Dale Schuurmans, and the listed collaborators. $91K total to Dale Schuurmans (Apr 1/05-Mar 31/10).
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
E. James Kehoe, University of New South Wales, Australia	Joint research with Rich Sutton and Elliot Ludvig on the relationship between reinforcement learning and learning in animals.
R. Gottardo (PI), UBC B. Clarke, UBC N. de Freitas, UBC A. Dobra, University of Washington A. Doucet, UBC P. Gill, UBC P. Gustafson, UBC P. Hoff, University of Washington L. Inoue, University of Washington K. Murphy, UBC D. Schuurmans, University of Alberta T. Swartz, Simon Fraser University	Pacific Institute of Mathematical Sciences (PIMS) Collaborative Research Grant, "Bayesian Modeling and Computation for Networks." Joint research with Dale Schuurmans, and the listed collaborators. (Jan 2008 – Jan 2010)
Dekang Lin, Google	Joint research with Dale Schuurmans on generalized regularization for large-scale natural language learning tasks
Alexander Smola, Yahoo Novi Quadrianto, ANU	Joint research with Dale Schuurmans on distributed maximum flow algorithms for large-scale, structured machine learning problems
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.
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.
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
Yavar Naddaf	MSc	Game-Independent AI Agents for Playing Atari 2600 Console Games	unknown
David Schnizlein	MSc	State Translation in No-limit Poker	Sandia National Labs, Albuquerque, New Mexico
Kevin Waugh	MSc	Abstraction in Large Extensive Games	PhD student at Carnegie Mellon University
Martha White	MSc	A General Framework for Reducing Variance in Agent Evaluation	PhD student, Department of Computing Science, UofA
David Silver	PhD	Reinforcement Learning and Simulation-Based Search in Computer Go	Senior Postdoctoral Associate at UCL
Wenye Li	Post Doctoral Fellow	Algorithms in Machine Learning	Assistant Professor at Macao Polytechnic Institute
Eric Wiewiora	Post Doctoral Fellow	Reinforcement Learning	Engineering Scientist at Intellisis

8. INTELLECTUAL PROPERTY

None.

9. PUBLICATIONS

REFEREED JOURNAL PUBLICATIONS

S. Bhatnagar, R. S. Sutton, M. Ghavamzadeh, M. Lee, "Natural Actor-Critic Algorithms," *Automatica*, 2009.

A. Antos, V. Grover and Cs. Szepesvári, "Active Learning with Heteroscedastic Noise, *Theoretical Computer Science* (DOI: 10.1016/j.tcs.2010.04.007), 2010.

J.Y. Audibert, R. Munos, and Cs. Szepesvári, "Exploration-Exploitation Trade-off Using Variance Estimates in Multi-Armed Bandits," *Theoretical Computer Science,* 410:1876-1902, 2009.

F. Balci, E. A. Ludvig, R. Abner, X. Zhuang, P. Poon, and D. Brunner, "Motivational Effects on Interval Timing in Dopamine Transporter (DAT) Knockdown Mice," *Brain Research*, 1325, 89-99, 2010.

G. Bartok, Cs. Szepesvári, and S. Zilles, "Models of Active Learning in Group-Structured State Spaces," *Information and Computation*, pp. 364-384, April 2010.

E. J. Kehoe, E. A. Ludvig, and R. S. Sutton, "Magnitude and Timing of CRs in Delay and Trace Classical Conditioning of the Nictitating Membrane Response of the Rabbit (*Oryctolagus cuniculus*)," *Behavioral Neuroscience 123*(5):1095–1101, 2009.

H.R. Maei, K. Zaslavsky, A.H. Wang, A.P. Yiu, C.M. Teixeira, S.A. Josselyn, and P.W. Frankland, "Development and Validation of a Sensitive Entropy-based Measure for the Water Maze," *Frontiers in Integrative Neuroscie*nce, 3:33, 2009.

G. Neu, and Cs. Szepesvári, "Training Parsers by Inverse Reinforcement Learning," *Machine Learning*, 77:303-337, 2009

L.M. Pilarski, P.M. Pilarski, A. Belch, "Multiple Myeloma May Include Microvessel Endothelial Cells of Malignant Origin," *Leukemia and Lymphoma*, 51(4), pp. 592-597, 2010.

I. Szita, M. Ponsen, and P. Spronck, "Effective and Diverse Adaptive Game AI," *IEEE Transactions on Computational Intelligence and AI in Games*, Vol. 1, Nr. 1, pp. 16-27, 2009.

B. Tanner and A. White, "RL-Glue: Language-Independent Software for Reinforcement-Learning Experiments," *Journal of Machine Learning Research*, 10:2133-2136, 2009.

S. Whiteson, B. Tanner, and A White, "The Reinforcement Learning Competitions," *AI Magazine*, to appear.

HIGHLY REFEREED ARCHIVAL CONFERENCE PROCEEDINGS

S. Ben-David, D. Pal, and S. Shalev-Shwartz, "Agnostic Online Learning," *Proc. 22nd Annual Conf. on Learning Theory,* 2009. 40% acceptance

S. Ben David, T. Lu, T. Luu, and D. Pál, "Impossibility Theorems for Domain Adaptation", *Proc. 13th Int'l Conf. on Artificial Intelligence and Statistics* (AISTATS-10), 9:129-136, 2010.

B. Póczos, R. Greiner, Cs. Szepesvári, and L. Li, "Budgeted Distribution Learning of Belief Net Parameters," *Proc. 27th Int'l Conf. on Machine Learning* (ICML-10), 2010. 26% acceptance.

K. Dwyer and G. Kondrak, "Reducing the Annotation Effort for Letter-to-Phoneme Conversion," *Proc. 47th Annual Meeting of the Association for Computational Linguistics* (ACL-IJCNLP), pp. 127-135, 2009. 21% acceptance.

T. Lu, D. Pál, and M. Pal, "Contextual Multi-Armed Bandits," *Proc.13th Int'l Conf. on Artificial Intelligence and Statistics* (AISTATS-10), 9:485-492, 2010.

A. Farhangfar, R. Greiner, and Cs. Szepesvári, "Learning to Segment From A Few Well-Selected Training Images," *Proc. 27th Int'l Conf. on Machine Learning* (ICML-09), pp. 305-312, 2009. 27% acceptance

P. Hooper, Y. Abbasi-Yadkori, R. Greiner, and B. Hoehn, "Improved Mean and Variance Approximations for Belief Net Responses via Network Doubling," *Proc. 25th Conf. on Uncertainty in Artificial Intelligence* (UAI-09), 2009. 31% acceptance

M. Lanctot, K. Waugh, M. Zinkevich, and M. Bowling, "Monte Carlo Regret Minimization for Extensive Games," *Proc. Advances in Neural Information Processing Systems 23* (NIPS-09), 2009. 24% acceptance

L. Li, B. Poczos, C. Szepesva¡ri, and R. Greiner, "Budgeted Distribution Learning in Parametric Models," *Proc. 27th Int'l Conf. on Machine Learning* (ICML-10), 2010. 26% acceptance

H.R. Maei, Cs. Szepesvari, S. Bhatnagar, and R.S. Sutton, "Toward Off-Policy Learning Control with Function Approximation," *Proc. 27th Int'l Conf. on Machine Learning* (ICML-10), 2010. 26% acceptance

H.R. Maei, Cs. Szepesvari, S. Bhatnagar, D. Precup, D. Silver and R.S. Sutton, "Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approximation," *Proc. Advances in Neural Information Processing Systems 23* (NIPS-09), 2009. 24% acceptance

B. Poczos, Y. Abbasi-Yadkori, Cs. Szepesvári, R. Greiner, and N. Sturtevant, "Learning When to Stop Thinking and Do Something!" *Proc. 27th Int'l Conf. on Machine Learning* (ICML-09). pp. 825-832, 2009. 27% acceptance

B. Póczos, S. Kirshner and Cs. Szepesvári, "REGO: Rank-based Estimation of Renyi Information using Euclidean Graph Optimization," *Proc. 13th Int'l Conf. on Artificial Intelligence and Statistics* (AISTATS-10), 2010.

N. Quadrianto, T. Caetano, J. Lim, and D. Schuurmans, "Convex Relaxation of Mixture Regression with Efficient Algorithms," *Proc. Advances in Neural Information Processing Systems 23* (NIPS-09), 2009. 24% acceptance

M. Ravanbakhsh, R. Greiner, and B. Póczos, "A Cross-Entropy Method for Optimizing Partially Decomposable Problems," *Proc. 24th Conf. on Artificial Intelligence* (AAAI-10), 2010.

A. Shademan, A. M. Farahmand, and M. Jägersand, "Robust Jacobian Estimation for Uncalibrated Visual Servoing" *Proc. of the IEEE Int'l. Conf. on Robotics and Automation* (ICRA-10), 2010.

D. Silver, and G. Tesauro, "Monte-Carlo Simulation Balancing," *Proc. 26th Int'l Conf. on Machine Learning* (ICML-09), 2009. 27% acceptance

R. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, Cs. Szepesvári, and E. Wiewiora, "Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation," *Proc.26th Int'l Conf. on Machine Learning*, (ICML-09), pp. 993–1000, 2009. 27% acceptance

I. Szita, and Cs. Szepesvári, "Model-based Reinforcement Learning with Nearly Tight Exploration Complexity Bounds," *Proc. 27th Int'l Conf. on Machine Learning* (ICML-10), 2010. 26% acceptance.

I. Szita, and A. Lőrincz, "Optimistic Initialization and Greediness Lead to Polynomial Time Learning in Factored MDPs," *Proc. 26th Int'l Conf. on Machine Learning* (ICML-09), 2009. 27% acceptance

J. Veness, D. Silver, A. Blair and W. Uther, "Bootstrapping from Game Tree Search *Proc. Advances in Neural Information Processing Systems 23,* (NIPS-09), 2009. 24% acceptance

T. J. Walsh, I. Szita, C. Diuk, and M. L. Littman, "Exploring Compact Reinforcement-Learning Representations with Linear Regression," *Proc. 25th Conf. on Uncertainty in Artificial Intelligence* (UAI-09), 2009. 31% acceptance

K. Waugh, N. Bard, and M. Bowling, "Strategy Grafting in Extensive Games," *Proc. Advances in Neural Information Processing Systems 23,* (NIPS-09), 2009. 24% acceptance

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

H. Yao, S. Bhatnagar, and D. Diao, "Multi-step Linear Dyna-style Planning," *Proc. Advances in Neural Information Processing Systems 23* (NIPS-09*),* 2009. 24% acceptance

Y. Yu, Y. Li, Cs. Szepesvári, and D. Schuurmans, "A General Projection Property for Distribution Families," *Proc. Advances in Neural Information Processing Systems 23* (NIPS 2009), 2009. 24% acceptance

OTHER REFEREED CONFERENCE PUBLICATIONS

Y. Guo and D. Schuurmans, "A Reformulation of Support Vector Machines for General Confidence Functions," *Proc. 1st Asian Conf. on Machine Learning* (ACML-09), 2009.

A. M. Farahmand, M. Ghavamzadeh, Cs. Szepesvári, and S. Mannor, "Regularized Fitted Q-Iteration for Planning in Continuous-Space Markovian Decision Problems," *Proc. of the American Control Conference* (ACC-09), 2009.

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

H.R. Maei and R.S. Sutton, "GQ(λ): A General Gradient Algorithm For Temporal-Difference Prediction Learning with Eligibility Traces," *Proc. 3rd Conf. on Artificial General Intelligence*, (AGI-10), 2010.

Y. Shi, Z. Cai, G. Lin, and D. Schuurmans, "Linear Coherent Bi-cluster Discovery via Line Detection and Sample Majority Voting," *Proc. 3rd Int'l Conf. on Combinatorial Optimization and Applications* (COCOA-09), 2009. LNCS 5573, pp. 73-84.

K. Waugh, M. Zinkevich, M. Johanson, M. Kan, D. Schnizlein, and M. Bowling "A Practical Use of Imperfect Recall," *Proc. 8th Symposium on Abstraction, Reformulation and Approximation* (SARA), 2009. 62.2% acceptance

H. Yao, S. Bhatnagar, and Cs. Szepesvári, "LMS-2: Towards an Algorithm that is as Cheap as LMS and Almost as Efficient As RLS," *Proc. 48th IEEE Conf. on Decision and Control* (CDC-09), 2009. 52% acceptance

R. S. Sutton, "The Grand Challenge of Predictive Empirical Abstract Knowledge," *Working Notes of the IJCAI-09 Workshop on Grand Challenges for Reasoning from Experiences*, 2009.

BOOKS and CHAPTERS

T. Degris and O. Sigaud, "Factored Markov Decision Processes" in *Markov Decision Processes in Artificial Intelligence,* ch. 4, p. 99-125. 2010.

E.A. Ludvig, M. G. Bellemare, and K.G. Pearson, "A Primer on Reinforcement Learning in the Brain: Psychological, Computational, and Neural Perspectives. To appear in E. Alonso, E. Mondragon (Eds.), *Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications*. Hershey, PA: IGI Global.

SPECIAL/INVITED PRESENTATIONS

Date	Title	Venue
April 6, 2009	Real-time Prediction Machines: How Animals Learn to Anticipate the Future (E. Ludvig)	Concordia University
April 9, 2009	Real-time Prediction Machines: How Animals Learn to Anticipate the Future (E. Ludvig)	Center for Theoretical Neuroscience, University of Waterloo
April 14, 2009	Keynote (R. Sutton)	Fourth Bellairs Workshop on Reinforcement Learning
April 29, 2009	Real-time Prediction Machines: How Animals Learn to Anticipate the Future (E. Ludvig)	Queensland Brain Institute, University of Queensland, Australia
April 30, 2009	Manifold-Adaptive Dimension Estimation (Cs. Szepesvári)	Statistics and Actuarial Science, University of Waterloo
May 1, 2009	Active Learning in Regression over Finite Domains (Cs. Szepesvári)	University of Waterloo
June 18, 2009	Deconstructing Reinforcement Learning (R. Sutton)	Keynote at the Multidisciplinary Symposium on Reinforcement Learning, Montreal
July 7, 2009	Reinforcement Learning and Simulation-Based Search in Computer Go (D. Silver)	MIT
Sept 21, 2009	Real-time Prediction Machines: How Animals Learn to Anticipate the Future (E. Ludvig)	Department of Psychology, University of Alberta
Nov 13, 2009	Gradient Descent Methods for Reinforcement Learning (Cs. Szepesvari)	Washington University
Nov 17, 2009	Core Learning Algorithms for Intrinsically Motivated Agents (R. Sutton)	IM-CLeVeR Workshop held in conjunction with the 9th Int. Conf. on Epigenetic Robotics, Venice, Italy
Nov 20, 2009	Rehabilitation Applications of Reinforcement Learning (R. Sutton)	Glenrose Rehabilitation Hospital, Edmonton
Nov 30, 2009	Gradient Descent Methods for Reinforcement Learning (Cs. Szepesvari)	University of Waterloo
Jan 5, 2010	Robots without Supervision (M. Bellemare)	McGill University
Feb 3, 2010	Rank based Euclidean graph optimization methods for information estimation (B. Poczos)	Department of Mathematical and Statistical Sciences, University of Alberta
Feb 8, 2010	Nonparametric information estimation using rank based Euclidean graph optimization methods (B.Poczos)	School of Computer Science, Carnegie Mellon University

March 5, 2010	Toward Learning Human-Level Predictive Knowledge (R. Sutton)	Keynote at the Third Conference on Artificial General Intelligence, Lugano, Switzerland

AWARDS

Michael Delp won the "Best Short Video" award ("WiiGesture") at the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, July 2009.

Csaba Szepesvari won the "Community Appreciation Award" for the video competition at the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, July 2009.

Michael Johanson and Michael Bowling were part of a team that competed on July 15, 2009 in the 4th Annual AAAI Computer Poker Competition; in 6 events. The team placed 1st on four competitions and placed 2nd on two competitions. The results were presented at IJCAI 2009 in Pasadena California.

Best overall paper honorable mention: L. Xu, M. White, and D. Schuurmans, "Optimal Reverse Prediction: A Unified Perspective on Supervised, Unsupervised And Semi-supervised Learning," ICML-2009.

THESES

David Silver (PhD) "Reinforcement Learning and Simulation-Based Search in Computer Go," August 18, 2009.

Yavar Naddaf, (MSc) "Game-Independent AI Agents for Playing Atari 2600 Console Games," March 2010.

David Schnizlein (MSc) "State Translation in No-limit Poker," July 24, 2009.

Kevin Waugh (MSc) "Abstraction in Large Extensive Games," August 2009.

Martha White, (MSc) "A General Framework for Reducing Variance in Agent Evaluation," January 2010.

10. OUTREACH

Richard Sutton, Elliot Ludvig (with Doina Precup of McGill) organized a special meeting to celebrate the multi-disciplinary aspects of reinforcement learning. This meeting was

held on June 18-19, 2009, in Montreal, in conjunction with ICML, UAI and COLT. The web site is http://msrl09.rl-community.org. AICML and RLAI supported this event with primary funding coming from MITACS. This meeting was attended by hundreds of researchers from around the world.

Patrick Pilarski hosted a robotics presentation and interactive demonstration for several groups from the University of Alberta's *DiscoverE Girls in Engineering Mentorship* (GEM) program. The presentation was made to about 40 female students in grades 6-8 on March 6, 2010.

Csaba Szepesvari supervised 2 high school students as part of an internship program during the summer of 2009.

Richard Sutton was interviewed in a special issue on Reinforcement Learning of the German journal *Künstliche Intelligenz* (Artificial Intelligence), on pages 41-43 of issue No. 3 in 2009.

Richard Sutton served on the "panel of experts" of the 2009 Canadian AI Graduate Student Symposium, held May 24, 2009, in conjunction with the 22nd Canadian Conference on Artificial Intelligence.