Report from Dagstuhl Seminar 11131

Exploration and Curiosity in Robot Learning and Inference

Edited by

Jeremy L. Wyatt[1], Peter Dayan[2], Ales Leonardis[3], and Jan Peters[4]

1 **University of Birmingham, GB,** `j.l.wyatt@cs.bham.ac.uk`
2 **University College London, GB,** `dayan@gatsby.ucl.ac.uk`
3 **University of Ljubljana, SI,** `alesl@fri.uni-lj.si`
4 **MPI für biologische Kybernetik – Tübingen, DE,** `jan.peters@tuebingen.mpg.de`

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 11131 "Exploration and Curiosity in Robot Learning and Inference". This seminar was concerned with answering the question: *how should a robot choose its actions and experiences so as to maximise the effectiveness of its learning?*. The seminar brought together workers from three fields: machine learning, robotics and computational neuroscience. The seminar gave an overview of active research, and identified open research problems. In particular the seminar identified the difficulties in moving from theoretically well grounded notions of curiosity to practical robot implementations.

Seminar 27. March – 1. April, 2011 – www.dagstuhl.de/11131
1998 ACM Subject Classification I.2.9 Robotics
Keywords and phrases Artificial Intelligence, Robotics, Learning, Exploration, Curiosity
Digital Object Identifier 10.4230/DagRep.1.3.67

1 Executive Summary

Jeremy L. Wyatt
Peter Dayan
Ales Leonardis
Jan Peters

Aims and background to the seminar

This seminar was concerned with answering the question: *how should a robot choose its actions and experiences so as to maximise the effectiveness of its learning?*.

This seminar was predicated on the assumption that to make significant progress in autonomous robotics, systems level theories of how robots should operate will be required. In recent years methods from machine learning have been employed with great success in robotics, in fields as diverse as visual processing, map building, motor control and manipulation. The machine learning algorithms applied to these problems have included statistical machine learning approaches, such as EM algorithms and density estimation, as well as dimensionality reduction, reinforcement learning, inductive logic programming, and other supervised learning approaches such as locally weighted regression. Most of these robot learning solutions currently require a good deal of supervised learning, or structuring of the learning data for a specific task. As robots become more autonomous these learning algorithms will have to be embedded in algorithms which choose the robot's next learning

Exploration and Curiosity in Robot Learning and Inference, *Dagstuhl Reports*, Vol. 1, Issue 3, pp. 67–95
Editors: Peter Dayan, Ales Leonardis, Jan Peters, and Jeremy L. Wyatt
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

experience. The problems become particularly challenging in the context of robotics, and even worse for a robot that is faced with many learning opportunities. A robot that can perform both manipulation, language use, visual learning, and mapping may have several quite different learning opportunities facing it at any one time. How should a robot control its curiosity in a principled way in the face of such a variety of choices? How should it choose data on the basis of how much it knows, or on how surprising it finds certain observations? What rational basis should robot designers choose to guide the robot's choice of experiences to learn from? There has been initial progress in several fields, including *machine learning*, *robotics* and also in *computational neuroscience*. In this seminar we brought together these three communities to shed light on the problem of how a robot should select data to learn from, how it should explore its environment, and how it should control curiosity when faced with many learning opportunities.

Machine Learning

This problem of how to explore is one that has been studied both in the context of reinforcement learning (exploration vs. exploitation) and supervised learning (active learning). Within the former the language of the sequential decision making community is that of MDPs and POMDPs. These are related to Bayesian perspectives on supervised Machine Learning in that we can think of a posterior over hypotheses that results from data that may be seen. In statistics this is related to the idea of conducting pre-posterior analysis. However, the theories from sequential decision making and active learning are currently unintegrated and partial, and it is not clear how they should apply in robotics. The current machine learning preference for Bayesian methods suggests that the ways that model uncertainty can be captured and exploited will be critical. During the seminar we looked for suggestions from this community as to how problems of exploration and curiosity in robotics can be formalised, especially at a systems level.

Robotics

Within areas like SLAM (Simultaneous Localisation and Mapping) the problem of how to select data has been addressed, but heuristic measures of exploratory worth are typically employed. Again, the principal formalism is that of Bayesian filtering, within which a POMDP is posed, but typically only the belief filter part is used. Rather than look ahead at all possibilities, heuristics such as information gain are used. There are also other approaches necessary in areas where for example robot learners are learning associations between data from multiple modalities, from time series, and where there maybe limited intervention from humans. Again, learning approaches in contemporary robotics are typically statistical, but there are other approaches. Techniques are also adapted to the domain, such as in the community working on the use of robotics in scientific discovery in the laboratory, where the robot has the ability to determine which experiments to perform, and the methods used employ a great deal of structure and prior knowledge about the domain. There are also challenges for exploration and curiosity from the use of robots for scientific exploration, such as in planetary missions, and in subsea exploration. We looked at how these problems are currently posed, and the challenges they pose for machine learning approaches to data selection.

Animal Cognition and Computational Neuroscience

The field of computational neuroscience has many insights to offer roboticists. Animals are forced to consider at each moment how they select data. There are examples of this in studies of motor learning, animal foraging, studies of neurotransmitters, and particular learning circuits, as well as in the study of areas of the brain concerned with action selection. Computational neuroscience has also been strongly influenced by statistical, particularly Bayesian approaches to inference and learning. For example much recent work strongly suggests a Bayesian underpinning to learning in the motor system, and other work has investigated possible neural bases for learning in the face of suprise and uncertainty. Work on reinforcement learning has linked with studies of brain areas such as the Basal ganglia, and there is debate as to whether or not the purpose of certain neurons is to provide a cue for learning in the face of novelty. This is related to the idea of infotaxis as a general mechanism for exploration control in some animal. The connection to the statistical theories from machine learning and optimal control are intriguing. This gives us a strong basis for the hope that a common framework for exploration and curiosity might emerge as a consequence of this seminar.

Summary of objectives

In summary the objectives of this seminar were to:

- Identify the different formulations of exploration and curiosity control, and to categorise robot problems into appropriate classes.
- Share statistical and non-statistical representations suitable for control of curiosity and exploration across communities.
- Identify the links between studies of learning control and motivation in computational neuroscience and formalisations from robotics and machine learning.
- Discuss possible formalisations of the problem of learning one of many possible tasks.
- Identify whether solution classes are heuristic or optimal.

Summary of the seminar program

The seminar was grouped into three themes, roughly according to Marr's levels of description: computational, algorithmic and implementational. Many talks crossed more than one level, but within these themes we were able to organize talks around more specific research areas. These areas were:

1. Ideas from neuroscience about the implementation of exploration and action in the brain.
2. Evidence from the ethology and psychology about the requirements for exploration, and algorithmic frameworks that fit the data on human behaviour.
3. Computational frameworks for intrinsic motivation and the evolution of extrinsic reward functions.
4. Algorithms and properties for specific sub-problems within curiosity and exploration: such as visual object search or the behaviour of greedy algorithms for solving sub-modular problems.
5. Robot implementations of algorithms for control of exploration and curiosity in real tasks.

Summary of the fundamental results

The main findings presented can be grouped into four parts. It is worth stating from the outset that a very large number of the talks, though by no means all, employed a reward based framework. It is not possible in this summary to mention all of the thirty talks given, instead we mention talks that illustrate the common themes of the seminar.

First a tutorial on the Basal Ganglia and it's role in action selection, including for exploration was given by Humphries. In this field there are now a range of computational models that simulate some of the internal workings of the Basal Ganglia. It was clear, however, that there are numerous structures about which little is known, and that many details of the models remain to filled in, or to be tested. Dayan provided evidence that was broadly negative with respect to a Bayesian view of exploration in humans. Dayan showed that human behaviour in a non-stationary bandit task is not better explained by a Bayesian view than by a simple soft-max reinforcement learning model. Sloman argued that the requirements to support exploration include the need to decompose domains into reuseable patterns. Contrary to a large number of the speakers Sloman argued against a statistical approach to exploration control. In the workshop as a whole statistical methods, with rewards, and often Bayesian inference were dominant, but these talks from biology present evidence that was not always supportive of this dominant approach.

Several different frameworks for intrinsic motivation were given. In several of these (Schmidhuber, Polani, Auer) the idea that exploration is driven by curiosity to enable greater understanding and ability to exploit the environment was central. These approaches can be contrasted with those that are ultimately driven by the need to maximise extrinsic rewards (Tishby, Starzyk). There seemed little question that all of these frameworks are quite general, but no clear unifying account is available. Others (Uchibe, Barto, Elfwing) showed ways to evolve extrinsic reward functions that ultimately contribute to overall agent fitness. Overall the division seems to be between information seeking and value seeking frameworks for self-motivation.

In algorithms the important findings concerned cases where problems that in the general case are intractable can be tackled much more effectively in special cases. Tsotsos showed as part of his talk that some visual search problems are tractable even though in the general case they are not. Krause showed that where problems have a sub-modular property that greedy algorithms can be close to optimal. Dearden showed how for a particular search task that entropic heuristics perform close to the level of more computational expensive information lookahead methods. While the most general algorithmic frameworks to exploration are based on the solution of POMDPs, each of these talks showed that a solution to a simpler problem can often provide very good performance.

In robotic tasks some approaches were necessarily more pragmatic, and this meant that many moved away from a purely reward based framework. Several showed ways of approximating solutions to POMDPs in real robot systems. These included using hierarchical approaches, sampling methods, limited horizon lookahead, or methods that split the problem into parts with, and without state uncertainty (Wyatt, Peters, Martinez-Cantin). While some advocated implementation of the principled frameworks for intrinsic motivation (Pape), problems were often moved away from the common statistical, reward based framework to enable solutions. The benefit of heuristic goal selection methods on top of precise planning approaches to achieving selected goals was demonstrated (Hanheide, Skocaj). A variety of robotic tasks were shown to be tackled with active learning, including motor control and social learning (Peters, Lopes).

The main themes that emerged were that while the dominant paradigm was one that was statistical and reward based, there were alternatives. While there were theoretically rigorous frameworks based on rewards, these were actually not much used by roboticists, who preferred pragmatic approaches. In the middle sit those exploring algorithms that while still approximate, offer some performance bounds relative to that which is optimal, howsoever defined.

2 Table of Contents

Executive Summary

Overview of Talks

Working Groups and Open Problems

3 Overview of Talks

3.1 A Model for Evaluating Autonomous Exploration

Peter Auer (Montan-Universität Leoben, AT)

I consider the problem of evaluating exploration strategies without predefined goals or rewards. Instead of evaluating only the behavioural complexity of the system after exploration, I propose to measure what the system can accomplish.

In a rather simplistic scenario this could be measured by the number of states the system can get into efficiently.

3.2 Learning Qualitative Spatio-temporal models of Activities and Objects from Video

Anthony G. Cohn (University of Leeds, GB)

In this talk I will present ongoing work at Leeds on building models of the world from observation, concentrating on unsupervised learning. The representations exploit qualitative spatio-temporal relations. A novel method for robustly transforming video data to qualitative relations will be presented.

I will present results from several domains including a kitchen scenario and an aircraft apron.

3.3 Polar Exploration

Peter Dayan (University College London, GB)

I will discuss the exploratory and exploitative behavior of human subjects in a four-armed restless bandit task. Despite our best analytical efforts, we could find no evidence that subjects awarded exploration bonuses to options they hadn't tried for a while. Instead, their exploratory behaviour was well-captured by a form of softmax choice in a conventional reinforcement learning model.

Fronto-polar cortex, a large and poorly understood area of the human brain, was specifically activated on trials that this model classified as exploratory, to a degree that depended on the requirement for cognitive control associated with those trials.

3.4 Robot Inference in Ocean Exploration

Richard W. Dearden (University of Birmingham, GB)

We look at the problem of finding hydrothermal vents using an autonomous underwater vehicle. This is an exploration problem where reward is only associated with finding a few specific states, rather than the quality of the map discovered. The problem can be formulated as a partially observable Markov decision process, but is far too large to be solved exactly. We examine two approaches, one based on approximating the solution to the POMDP using forward search in belief space to take account the information gained through each observation made, and the other based on treating the problem as a mapping problem and using entropy reduction. We show that both approaches perform much better than the state of the art, and that statistically the two approaches perform very similarly, suggesting that entropy reduction is a useful heuristic even for problems like this where it is clearly optimising the wrong criterion.

3.5 Supporting exploration in children with disabilities through lifelong robotic assistants

Yiannis Demiris (Imperial College London, GB)

Children and adults with sensorimotor disabilities can significantly increase their autonomy through the use of assistive robots. As the field progresses from short-term, task- specific solutions to long-term, adaptive ones, new challenges are emerging. In this talk a lifelong methodological approach is presented, that attempts to balance the immediate context-specific needs of the user, with the long-term effects that the robot's assistance can potentially have on the user's developmental trajectory. I will use examples from adaptive robotic wheelchairs assisting young children and adults to illustrate the methodology, and I will discuss the underlying computational learning mechanisms and robotic infrastructure.

3.6 Embodied Evolution of Learning Ability and the Emergence of Different Mating Strategies in a Small Robot Colony

Stefan Elfwing (Okinawa Institute of Science and Technology, JP)

In this study, we use a framework for performing embodied evolution with a limited number of robots, by utilizing time-sharing in subpopulations of virtual agents hosted in each robot. Within this framework, we explore the combination of within-generation learning of basic survival behaviors by reinforcement learning, and evolutionary adaptations over the generations of the basic behavior selection policy, the reward functions, and meta-parameters for reinforcement learning. We apply a biologically inspired selection scheme, in which there is no explicit communication of the individuals' fitness information. The individuals can only reproduce offspring by mating-a pair-wise exchange of genotypes-and the probability that an individual reproduces offspring in its own subpopulation is dependent on the individual's "health," i.e., energy level, at the mating occasion. In addition, we investigate the emergence of different mating strategies, i.e., different basic behavior selection policies.

In the experiments, we observed two individual mating strategies: 1) Roamer strategy, where an agent never waits for potential mating partners and 2) Stayer strategy, where an agent waits for potential mating partners depending on the agent's current state. The most interesting finding was that in some simulations the evolution produced a mixture of mating strategies within a population, typically with a narrow roamer subpopulation and a broader stayer subpopulation with distinct differences in genotype, phenotype, behavior, and performance between the subpopulations.

3.7 Hierarchical object inference and exploration in mobile whiskered robots

Charles Fox (University of Sheffield, GB)

I will describe work on object recognition with autonomous mobile whiskered robots. Whiskers are highly localised sensors and the question of "where to look next" to gather useful information is important. I will describe a Bayesian Blackboard approach to recognising hierarchical objects, using Gibbs sampling and annealing together with blackboard system heuristics. Entropy in the Gibbs sampler provides a natural measure of saliency which may be used as a basis to guide exploration.

3.8 Learning-based modeling and stability design of interactive human locomotion patterns

Martin A. Giese (Universitätsklinikum Tübingen, DE)

Joint work of M.A. Giese, A. Mukovskiy, J.J. Slotine L. Omlor

The online synthesis of interactive human body movements is a key problem for different technical applications, such as robotics and computer graphics. We present a biologically inspired algorithm for the modeling of complex body movements based on dynamic primitives by a combination of unsupervised learning and methods from nonlinear dynamics. We illustrate how this method can be applied for the synthesis of collective behaviors of groups of locomoting agents, taking the full nonlinearity of the kinematics into account. For selected cases we show how Contraction Theory can be applied to analyze and design the stability properties of the emerging collective behavioral patterns.

Supported by the DFG, EU FP6 Project COBOL, and FP7 Projects TANGO and AMARSI.

3.9 Dora: explore and exploit probabilistic knowledge for object search

Marc Hanheide (University of Birmingham, GB)

I am presenting Dora, our robot systems that explores its environment in a task-driven way to find an object. Besides taking topological and spatial knowledge pre-acquired as part of a mapping process into account it is also equipped with a probabilistic ontology that enables it to exploit common-sense knowledge. This probabilistic model comprises information about the likelihood of finding certain objects in particular categories of rooms (e.g. that cornflakes are usually found in kitchens) and observation models of successfully detecting an object if it is present. By reasoning within this probabilistic spatial representation that is incorporates perceived instance knowledge a probabilistic belief state of the world is maintained. Planning within this probabilistic representation using a switching planner, switching between sequential and contingent planning sessions, we can exploit the common-sense knowledge to find objects more efficiently in many cases. However, the system can still cope with violated assumptions, i.e. when common-sense knowledge is not applicable, and sensing errors. I discuss the limitations and caveats of our approach so far and briefly talk about the challenge to combine task-driven with curiosity driven exploration in our world. The accompanying video can be viewed at
http://www.youtube.com/watch?v=0QcmSDZR-c4.

3.10 Neural substrates for action selection: the basal ganglia

Mark D. Humphries (ENS – Paris, FR)

All animals must continuously sequence and co-ordinate behaviors appropriate to both their context and internal milieu if they are to survive. It is natural to wonder what parts of the nervous system - the neural substrate - evolved to carry out this action selection process. I will discuss the proposal that the vertebrate brain has co-evolved (or co-opted) specialised and centralised neural systems for action selection, handling both the competition between systems accessing the final motor pathway and the open-ended nature of a flexible, extensible behavioural repertoire. In particular, I will focus on the set of fore- and mid-brain nuclei called the basal ganglia that seem to implement a repeated circuit design ideal for computing input selection. Of particular note for understanding spatial exploration are the circuits implemented by the ventral basal ganglia, which integrate information on current position, overall strategy, and reward. The talk will first lay out this circuit, sketching the unique features of both its components and overall architecture that have given clues to its function. I will then address the breadth of computational modelling of the generic basal ganglia circuit, considering both mechanism-mining (build biologically accurate model, search for insights) and mechanism-mapping (take algorithm, fit to extant neural circuit) approaches to understanding neural computation.

3.11 Exploration in Relational Worlds

Kristian Kersting (Fraunhofer IAIS – St. Augustin, DE)

One of the key problems in model-based reinforcement learning is balancing exploration and exploitation. Another is learning and acting in large relational domains, in which there is a varying number of objects and relations between them. We provide a solution to exploring large relational Markov decision processes by developing relational extensions of the concepts of the Explicit Explore or Exploit (E3) algorithm. A key insight is that the inherent generalization of learnt knowledge in the relational representation has profound implications also on the exploration strategy: what in a propositional setting would be considered a novel situation and worth exploration may in the relational setting be an instance of a well-known context in which exploitation is promising. Our experimental evaluation shows the effectiveness and benefit of relational exploration over several propositional benchmark approaches on noisy 3D simulated robot manipulation problems.

3.12 Adaptive Submodularity: A New Approach towards Active Learning and Stochastic Optimization

Andreas Krause (ETH Zürich, CH)

Joint work of Krause, Andreas; Golovin, Daniel; Ray, Debaivoti

Many information gathering problems require us to adaptively select observations to obtain the most useful information. These problems involve sequential stochastic optimization under partial observability – a fundamental but notoriously difficult challenge. Fortunately, often such problems have a structural property that makes them easier than general sequential stochastic optimization. In this talk, I will introduce this structural property – a new concept that we call adaptive submodularity – which generalizes submodular set functions to adaptive policies.

In many respects adaptive submodularity plays the same role for adaptive problems such as sequential experimental design as submodularity plays for nonadaptive problems (such as placing a fixed set of sensors). Specifically, just as many nonadaptive problems with submodular objectives have efficient algorithms with good approximation guarantees, so too do adaptive problems with adaptive submodular objectives. I will illustrate the usefulness of the concept by giving several examples of adaptive submodular objectives arising in diverse applications including sensor selection, viral marketing and active learning.

Proving adaptive submodularity for these problems allows us to recover existing results in these applications as special cases and handle natural generalizations. In an application to Bayesian experimental design, we show how greedy optimization of a novel adaptive submodular criterion outperforms standard myopic heuristics such as information gain and value of information.

3.13 Active Exploration in Social Learning

Manuel Lopes (INRIA – Bordeaux, FR)

In the work we present a system to learn task representations from ambiguous feedback. We consider an inverse reinforcement learner that receives feedback from a user with an unknown and noisy protocol. The system needs to estimate simultaneously what the task is, and how the user is providing the feedback. We further explore the problem of ambiguous protocols by considering that the words used by the teacher have an unknown relation with the action and meaning expected by the robot. We present computational results in a large scale problem and on a simplified robotic one. We show that it is possible to learn the task under a noisy and ambiguous feedback. Using an active learning approach, the system is able to reduce the length of the training period.

3.14 Non-convex optimization for active robot learning

Ruben Martinez-Cantin (CUD – Zaragoza, ES)

Active learning provides the optimal trade-off between statistical learning and decision making. Traditionally, both fields have been characterized for relying in convex functions to solve high dimensional problems. However, many robotics problems can be formulated as a low dimensional problem in continuos state and action spaces. Thus, non-convex optimization methods such as EI can be applied efficiently.

3.15 Towards practical implementations of curious robots

Leo Pape (IDSIA – Lugano, CH)

Schmidhuber's theory of compression-driven progress considers limited computational agents that try to represent their observations in an efficient manner. Finding efficient representations entails identifying regularities that allow the observer to compress the original observations and predict future observations. Compression progress is achieved when the agent discovers previously unknown regularities that allow for increased compression of observations. The compression progress of the compression mechanism is monitored by an action generation method that generates actions for which it expects further improvement in the compressor. This allows the agent to focus on collecting interesting observations, that is, observations that are novel, yet learnable by the compressor. In my talk I will discuss how this general principle of compression progress is used for the implementation of curious robots.

3.16 Exploration in Learning of Motor Skills for Robotics

Jan Peters (MPI für biologische Kybernetik – Tübingen, DE)

Intelligent autonomous robots that can assist humans in situations of daily life have been a long standing vision of robotics, artificial intelligence, and cognitive sciences. A elementary step towards this goal is to create robots that can learn tasks triggered by environmental context or higher level instruction. However, learning techniques have yet to live up to this promise as only few methods manage to scale to high-dimensional manipulator or humanoid robots. In this talk, we investigate a general framework suitable for learning motor skills in robotics which is based on the principles behind many analytical robotics approaches. It involves generating a representation of motor skills by parameterized motor primitive policies acting as building blocks of movement generation, and a learned task execution module that transforms these movements into motor commands. We discuss learning on three different levels of abstraction, i.e., learning for accurate control is needed to execute, learning of motor primitives is needed to acquire simple movements, and learning of the task-dependent "hyperparameters" of these motor primitives allows learning complex tasks. We discuss task-appropriate learning approaches for imitation learning, model learning and reinforcement learning for robots with many degrees of freedom.

Empirical evaluations on a several robot systems illustrate the effectiveness and applicability to learning control on an anthropomorphic robot arm. A large number of real-robot examples will be demonstrated ranging from Learning of Ball-Paddling, Ball-In-A-Cup, Darts, Table Tennis to Grasping.

3.17 Potential Information Flows as Driving Principle for Agents

Daniel Polani (University of Hertfordshire, GB)

In recent years, models for self-motivation in robots and agents have attracted significant interest in the research community. One challenge is to formulate such intrinsic drives with as little "engineer's bias" as possible. Many important approaches have been suggested in the last decades, which one can roughly classify e as process-oriented and structure-oriented. We consider the structure-oriented quantity of "empowerment", which is the potential Shannon information flow in the external perception-action loop of agents. Embedded in the framework of information theory, it displays a number of properties that make it promising for the use as an intrinsic motivation drive in a structured environment. Empowerment probes salient aspects of the structure of an agent's environment, and displays intuitively plausible solutions. The talk will introduce the concept, demonstrate a number of examples and demonstrate several properties of empowerment which may make it relevant not only for constructing artificial drives but also for speculations on biological ones.

3.18 Formal Theory of Fun and Creativity for Curious Agents

Juergen Schmidhuber (IDSIA – Lugano, CH)

Main reference J. Schmidhuber, "Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010)." IEEE Transactions on Autonomous Mental Development, 2(3):230-247,2010.
URL http://www.idsia.ch/~juergen/ieeecreative.pdf

Our fast deep / recurrent neural nets recently achieved numerous 1st ranks in many pattern recognition competitions and benchmarks, without any unsupervised pre-training. The future, however, will belong to active systems learning to sequentially shift attention towards informative inputs, not only solving externally posed tasks, but also their own self-generated tasks designed to improve their understanding of the world according to our Formal Theory of Fun and Creativity, which requires two interacting modules: (1) an adaptive (possibly neural) predictor or compressor or model of the growing data history as the agent is interacting with its environment, and (2) a (possibly neural) reinforcement learner. The learning progress of (1) is the FUN or intrinsic reward of (2). That is, (2) is motivated to invent skills leading to interesting or surprising novel patterns that (1) does not yet know but can easily learn (until they become boring). We discuss how this principle explains science and art and music and humor.

Further details are available at

- Formal Theory of Fun & Creativity: www.idsia.ch/~juergen/creativity.html
- Artificial Curiosity: www.idsia.ch/~juergen/interest.html

3.19 Exploration in parameter space - a practical evaluation

Frank Sehnke (ZSW – Stuttgart, DE)

Parameter Exploring Policy Gradients like NES, SDE-PG, PGPE and NPGPE have drawn some attention lately. We will summarise why this is the case, and focus on the question why exploring in parameter space is so special. Furthermore, learning of real-world complex RL tasks with these methods has shown that subtle deviations from the mathematically correct implementation of the exploration part lead to significant improvements to both robustness of the algorithm and quality of the solution found. The changes introduced are strikingly similar to the underlying principles of exploration in evolutionary methods, and as a whole give rise to a provocative question: Are we optimising the wrong parameters for exploration in PG?

3.20 Interactive learning in dialogue with a tutor

Danijel Skocaj (University of Ljubljana, SI)

Joint work of Danijel Skocaj, Matej Kristan, Alen Vrecko, Marko Mahnic, Barry Ridge, Ales Leonardis, Pierre Lison, Miroslav Janicek, Geert-Jan M. Kruijff , Ivana Kruijff-Korbayova, Raveesh Meena, Michael Zillich, Kai Zhou, Moritz Göbelbecker, Thomas Keller, Marc Hanheide, Nick Hawes
URL http://cogx.eu/results/george/

In this talk we will present representations and mechanisms that facilitate continuous learning of visual concepts in dialogue with a tutor and show the implemented robot system. We will present how the beliefs about the world are created by processing visual and linguistic information and how they are used for planning the system behaviour aimin at satisfying its internal drive - to extend its knowledge. The system facilitates different kinds of learning initiated by the human tutor or by the system itself. We demonstrate these principles in the case of learning about object colours and basic shapes and present the experimental results that justify such mixed-initiative learning process.

3.21 How to combine AI and Piaget's genetic epistemology (understanding possibility and necessity in multiple exploration domains)

Aaron Sloman (University of Birmingham, GB)

It is not widely known that shortly before he died Jean Piaget and his collaborators produced a pair of books on Possibility and Necessity, exploring questions about how two linked sets of abilities develop:

(a) The ability to think about how things might be, or might have been, different from the way they are.
(b) The ability to notice limitations on possibilities, i.e. what is necessary or impossible.

I believe Piaget had deep insights into important problems for cognitive science that have largely gone unnoticed, and are also important for research on intelligent robotics, or more generally Artificial Intelligence (AI), as well as for studies of animal cognition and how various animal competences evolved and develop.

The topics are also relevant to understanding biological precursors to human mathematical competences and to resolving debates in philosophy of mathematics, e.g. between those who regard mathematical knowledge as purely analytic, or logical, and those who, like Immanuel Kant, regard it as being synthetic, i.e. saying something about reality, despite expressing necessary truths that cannot be established purely empirically, even though they may be initially discovered empirically (as happens in children).

It is not possible in one seminar to summarise either book, but I shall try to present an overview of some of the key themes and will discuss some of the experiments intended to probe concepts and competences relevant to understanding necessary connections.

In particular, I hope to explain: (a) The relevance of Piaget's work to the problems of designing intelligent machines that learn the things humans learn.

(Most researchers in both Developmental Psychology and AI/Robotics have failed to notice or have ignored most of the problems Piaget identified.) (b) How a deep understanding of AI, and especially the variety of problems and techniques involved in producing machines that can learn and think about the problems Piaget explored, could have helped Piaget describe and study those problems with more clarity and depth, especially regarding the forms of representation required, the ontologies required, the information processing mechanisms required and the information processing architectures that can combine those mechanisms in a working system – especially architectures that grow themselves.

That kind of computational or "design-based" understanding of the problems can lead to deeper clearer specifications of what it is that children are failing to grasp at various stages in the first decade of life, and what sorts of transitions can occur during the learning. I believe the problems, and the explanations, are far more complex than even Piaget thought. The potential connection between his work and AI was appreciated by Piaget himself only very shortly before he died.

One of the key ideas implicit in Piaget's work (and perhaps explicit in something I have not read) is that the learnable environment can be decomposed into explorable domains of competence that are first investigated by finding useful, reusable patterns, describing various fragments. Then eventually a large scale reorganisation is triggered (per domain) which turns the information about the domain into a more economical and more powerful generative system that subsumes most of the learnt patterns and, through use of compositional semantics in the internal representation, allows coping with much novelty – going far beyond what was learnt.

(I think this is the original source of human mathematical competences.)

Language learning seems to use a modified, specialised, version of this more general (but not totally general) mechanism, but the linguistic mechanisms were both a later product of evolution and also get turned on later in young humans than the more general domain learning mechanisms. The linguistic mechanisms also require (at a later stage) specialised mechanisms for learning, storing and using lots of exceptions to the general rules induced (the syntactic and semantic rules).

The language learning builds on prior learning of a variety of explorable domains, providing semantic content to be expressed in language. Without that prior development, language learning must be very shallow and fragmentary – almost useless.

When two or more domains of exploration have been learnt they may be combinable, if their contents both refer to things and processes in space time. E.g. a domain of actions on sand and a domain of actions on water could be combined, producing things like sandcastles with moats, mud, etc.

I think Piaget was trying to say something like this but did not have the right concepts, though his experiments remain instructive.

Space-time is the the great bed in which many things can lie together and produce novelty. One of the features is a kind of continuity of interaction of structures (e.g. levers, gear wheels, pulleys, string, etc.) that I don't think current forms of representation in AI are well suited to.

Moreover, I have the impression that attempts to deal with uncertainty by using probabilities are completely mistaken, and that biological evolution produced something far more powerful, mainly concerned with use of non-metrical or semi-metrical (partially ordered) forms of representation.

Producing working demonstrations of these ideas in a functional robot able to manipulate things as a child does will require major advances in AI, though there may already be more work of this type than I am aware of.

A version of this talk presented at Birmingham on 21st Feb 2011 is available here (PDF): http://www.cs.bham.ac.uk/research/projects/cogaff/misc/talks/#talk90

(Comments and criticisms welcome, before, during or after Dagstuhl!)

Since I originally gave the talk I have discovered that some of the ideas are also in Annette Karmiloff-Smith's 1992 book, Beyond Modularity, though she does not stress, as Piaget does, the importance of being able to think about possibility and necessity.

3.22 A Mechanism for Learning, Attention Switching, and Cognition

Janusz Starzyk (Ohio University, US)

This talk would directly relate to questions of how should a robot control its curiosity when faced with many learning opportunities and a variety of choices, how should it choose data on the basis of how much it knows and what are his objectives, or on how surprising it finds certain observations? In this talk a new machine learning method called motivated learning (ML) will be presented.

ML applies to autonomous embodied intelligence agents who build perceptions and learn their actions so as to maximize the effectiveness of their mental development through learning. Motivated learning drives a machine to develop abstract motivations and choose its own goals. ML also provides a self-organizing system that controls a machine's behavior based on competition between dynamically-changing motivations, perceptions and internal thoughts.

This provides interplay of externally driven and internally generated signals that control a machine's behavior.

ML method can be combined with artificial curiosity and reinforcement learning.

It enhances their versatility and learning efficiency, particularly in changing environments with complex dependencies between environment parameters. It has been demonstrated that ML not only yields a more sophisticated learning mechanism and system of values than reinforcement learning (RL), but is also more efficient in learning complex relations and delivers better performance than RL in dynamically changing environments.

In addition, ML provides a much needed mechanism for switching a machine's attention to new motivations and implementation of internal goals. A motivated learning machine develops and manages its own motivations and selects goals using continuous competition between various levels of abstract pain signals (including curiosity and possible attention switching signals). This form of distributed goal management and competing motivations is a core element of central executive control that may govern the cognitive operation of intelligent machines.

This talk will present basic properties and underlying philosophy of ML. In addition, it will present the basic neural network structures used to create abstract motivations, higher level goals, and subgoals. It will show some simulation results to compare ML and RL in environments of gradually increasing sophistication and levels of difficulty.

Subsequently, an organization of central executive unit for a cognitive system will be described. This unit uses distributed and competing signals that represent goals, motivations, emotions, and attention. Only after the winner of such competition is established, it drives a focus point of a conscious experience. This computational model of consciousness mimics the biological systems functionally and retains a well-defined architecture necessary for implementing consciousness in machines. In addition a new concept of mental saccades will be presented to explain the attention switching and focusing in this computational model of consciousness.

The model uses competition among three different types of signals in a cognitive cycle of the agent. Mental saccades may not relate to a similar mechanism in human mind, however, such mechanism is useful for the computational implementation of machine consciousness.

3.23 Information flows and discounting in planning and learning

Naftali Tishby (The Hebrew University of Jerusalem, IL)

We argue that information seeking and reward seeking behavior should be optimized together in a "free-energy" minimization, that combines physical rewards with information gains and costs. The local information gain includes both the control/decision complexity and information provided by the environment responses (Tishby & Polani, 2009). One interesting interpretation of this framework is that information terms can be considered as internal rewards while physical (standard) rewards are external to the organism. Perfect adaptation to the environment happens when these two reward systems are proportional to each other. This can happen either through adaptation of the (subjective, belief states) transitions probabilities or of the external rewards (or both).

In this talk I will discuss another new application of this framework – information discounting. The Bellman equations for information have no explicit natural discounting. But we know that information about future event is sub-extensive for any stationary environment (Bialek, Nemenman, Tishby 2002).

This suggests that information gains must also be discounted, either exponentially (for finite dimensional environment, when predictive information grows logarithmically) or hyperbolically for infinite dimensional environments (when predictive information grows like a power-law) to be consistent with the predictive information decay. This can provide a new principled approach for general reward discounting in exploration/exploitation models.

3.24 How should Robbie search for my lost cup?

John K. Tsotsos (York University – Toronto, CA)

I begin with the workshop themes of curiosity, exploration and learning. I present a view of curiosity as attention to the relevant and ignoring the irrelevant. The model of Selective Turning is briefly presented. For exploration, I show an algorithm to search for an object in an unknown room.

Examples of the performance of the algorithm are included. A comparative experiment testing performance characteristics given different search policies, different starting points and different object placement, all with or without prior knowledge is presented. In all cases, the policy of maximising the probability of finding the object while minimizing distance travelled is superior. The presentation ends by pointing to a number of areas where learning procedures would be helpful and lead to a more robust process.

3.25 Finding Intrinsic Rewards by Embodied Evolution and Constrained Reinforcement Learning

Eiji Uchibe (Okinawa Institute of Science and Technology, JP)

Finding the design principle of reward functions is a big challenge both in artificial intelligence and neuroscience. Successful acquisition of a task usually requires not only the rewards for goals, but also for intermediate states to promote effective exploration. We propose a method to design 'intrinsic' rewards of autonomous robots by combining constrained policy gradient reinforcement learning and embodied evolution. To validate the method, we use the Cyber Rodent robots, in which collision avoidance, recharging from battery pack, and 'mating' by software reproduction are three major 'extrinsic' rewards. We show in hardware experiments that the robots can find appropriate 'intrinsic' rewards for the vision of battery packs and potential mating partners to promote approach behaviors.

3.26 Goal-Directed Action Generalization

Ales Ude (Jozef Stefan Institute – Ljubljana, SI)

Acquisition of new sensorimotor knowledge by imitation is a promising paradigm for robot learning. To be effective, action learning should not be limited to direct replication of movements obtained during training but must also enable the generation of actions in situations a robot has never encountered before. In this talk I will present a methodology that enables the generalization of the available sensorimotor knowledge. New actions are synthesized by the application of statistical methods, where the goal and other characteristics of an action are utilized as queries to create a suitable control policy, taking into account the current state of the world. Nonlinear dynamic systems are employed as a motor representation. The proposed approach enables the generation of a wide range of policies without requiring an expert to modify the underlying representations to account for different task-specific features and perceptual feedback.

3.27 Robot Visual Learning

Markus Vincze (TU Wien, AT)

Allowing a robot to acquire 3D object models autonomously not only requires robust feature detection and learning methods but also mechanisms for guiding learning and assessing learning progress. In this talk we presented probabilistic measures for observed detection success, predicted detection success and the completeness of learned models, where learning is incremental and online.

This allows the robot to decide when to add a new keyframe to its view-based object model, where to look next in order to complete the model, predicting the probability of successful object detection given the model trained so far as well as knowing when to stop learning.

3.28 Active information gathering in vision and manipulation

Jeremy L. Wyatt (University of Birmingham, GB)

We can think of vision as a process by which an agent gathers information which is relevant to a task. In this approach the value of perceptual information is anchored in the additional reward the information allows you to get in a task. I describe two pieces of work. The first is the use of POMDPs to plan task specific visual processing. In this we show that by planning to reduce uncertainty we get visual routines for specific tasks that are more reliable and faster than simply running all visual operators. In the second piece of work we pose gaze allocation in a similar reward based framework. In this work camera movements generate images that reduce uncertainty about the locations of objects and places involved in a manipulation task. In a reward-based framework the robot picks the camera movements that generate the most additional expected reward for the motor behaviours (here grasps) by reducing the relevant uncertainty. Thus both these pieces of work show how perception can be viewed as an information gathering process that can be grounded via the notion of task rewards.

3.29 Embodied Learning of Qualitative Models

Jure Zabkar (University of Ljubljana, SI)

People most often reason qualitatively and if the goal of AI is to mimic human intelligence, the robots should also be equiped with qualitative reasoning (QR) capabilities. QR is based on approximate understanding of functional relationships between the quantities. Qualitative models are less precise than numerical ones but more robust (noise resistant) and high-level (therefore simpler and easier to understand). In this talk, I will present some algorithms for learning qualitative models and their applications in embodied systems. Most of this work was carried out in collaboration with Janez Demsar, Martin Mozina and Ivan Bratko and supported by FP6 project XPERO.

4 Working Groups and Open Problems

4.1 Group 1: Self-motivation

The membership of this group was: Andy Barto, Daniel Polani, Tali Tishby, Aaron Sloman, and Jürgen Schmidhuber. This group was concerned with frameworks for understanding intrinsic motivation. The key questions addressed included:

- What is necessary for computational accounts of intrinsic motivation?
- How do different computational theories overlap and how do they differ?
- Are there universal principles?
- How do computational theories relate to what psychologists have said about intrinsic motivation and what behavioral data show?

The core discussion centered around low-level (and essentially information-theoretic) concepts to characterize intrinsically motivated behaviour - we did not discuss higher-level concepts. Sloman wondered if key issues were not addressable using information theory due to its apparent lack of semantics. Tishby argued that this view arose from a widespread misunderstanding of the concept of the concept of information. In fact, semantics, so the argument goes, can be put in via the concept of "relevant" and "valuable" information which carries information about the task to do. Additional structure can be added via the embodiment and the ensuing preferred information channels by which an agent interacts with the environment. This physical structure imposes constraints on the form of the information processing/propagation – thus, "not all information is created equal".

Particular topics and questions that were identified as being of future interest to the community included:

1. Empowerment: projecting potential action sequence into states that can be reached in the near future. Andy suggested a match with the ''drive for mastery" hypothesized in psychology.
2. Predictive Information: Entropy is related to predictability. Der and Ay use this for generating sensitive and predictive behaviour online. Complexity arises because predictive information makes the future predictive from the past, but the past needs to be rich at the same time. A 'poor' past will have no information to predict about the future.
3. Other intrinsic motivation measures: there are many more heuristic measures such as learning progress and the autotelic principle. These capture aspects of what one expects from self-motivated behaviour (including Csíkszentmihályi's notion of 'flow'). They still, however, contain still some arbitrary aspects which one may want to avoid.
4. Compression: in Schmidhuber's compression scheme the machine attempts to compress the entire past. The expected compression progress is used as a driver for 'artificial curiosity' and 'satisfaction' of the success criterion. One open question is how does the machine provide any externally visible abstraction? A second issue is whether the machine is to be limited - this is plausible in view of realistic machines and it is a Kolmogorov analogue of Shannon-type constraints and thus fits the general philosophy well: abstraction as a response to constraints: "making a virtue out of necessity".
5. Valuable information: Tishby has proposed the idea of valuable information. This brings us back to the Shannon picture of information. In this framework the machine needs to remember or compress only what is valuable: this is a very key point. There is no need to take in all information, only what will contribute to future reward. This is bounded by what can at all be predicted in the future, i.e. predictive information.

There was a more extensive discussion around the idea of valuable information, and it's relation to other concepts. If we consider the entropy of a series of a system through time, there is an extensive part (that grows linearly with time and is "uninteresting") and a subextensive part (convex, less-than-linear) component. Predictive information, the information that the past at some point has about the future, happens to be the same as the subextensive part of the entropy of a system. However, compressed info may not be unpackable, i.e. the information may be compressed in such a way that some external agent may unpack it (this must be possible, or it could not be called compression), but it may no longer be accessible to the agent itself which may be not complex enough to unpack the message (at least that's how Polani understands it). Barto thought that this is the same idea that comes up in the deterministic case as the idea of a homomorphic image of a dynamical system: Given some variables that are of interest, what is the minimal state representation that can provide enough information to 1) determine itself at the next time step, and 2) determine the values of the variables of interest. On that level, this seems to largely agree with Tishby's relevant/valuable information (via information bottleneck), in the deterministic case. Again, however, it is important to note that in Tishby's framework only part of the past is important. Barto wondered if indefinite memory systems might violate this: e.g., setting a bit in the arbitrary past can influence the present. Tishby and Polani argued that that is still just a part of the past. Polani says it is either known to the agent (then just part of the memory) or will play a role in the future, and then discovered to be an "inconsistency" that has not yet been observed. The concept of natural cut-offs was then discussed. These are the finite life-spans of organisms that force exploitation and they create natural time-scales and cut-offs for learning phases. Finally it was asked what would an agent do using Tishby's scheme if it had no value: i.e. we only use information gain as a source of reward? After some discussion, Tali concluded that it would act to efficiently gain information about state transitions.

Other areas of related work were also identified. Mike Duff's work on Bayesian optimal exploration and Bellman's "curse of the expanding grid" was identified as relevant. In this case exploration is posed as a reasoning in a meta-MDP whose states consist of physical states composed with beliefs states about model parameters (e.g., transition probabilities). Duff used the decreases in the variance of belief states as a reward function for doing RL in the meta state. How is this related to Tishby's ideas? It was agreed that while there were similarities that Tishby's approach is distinguished from Bayesian approaches.

Summary: In conclusion much of the discussion boiled down to the issue of whether we desire information-seeking behaviour or value-seeking behaviour. Two important open questions concern whether these are really similar? If not then when not?

4.2 Group 2: Scaling methods for exploration and curiosity

This break out group (Jeremy Wyatt, Eiji Uchibe, Kristian Kersting, Marc Hanheide, Stefan Elfwing, Frank Sehnke, Leo Pape, Jure Zakbar) was concerned with the issue of how to scale existing methods. It seemed to us that there were several challenges to scaling techniques for exploration and uncertainty to larger problem domains and more complex robot systems. Each of these defines a open challenge for the research community. Each is detailed in turn.

A. How to integrate common sense and structured knowledge with uncertainty?

It is an old AI chestnut that common sense knowledge can speed up reasoning and make action more efficient by restricting our space of reasonable actions. However, since many of the approaches for exploration control are statistical, capturing common sense knowledge means we need to be able to express it in some similar manner. In the past five years ontology based reasoning has been applied in robots (e.g. the work of Saffioti (Orebro), Beetz et al. (TU Munich)). This work has mostly been applied in making inferences about objects and rooms in topological and semantic SLAM. Work by Kuipers et al uses a map with several levels of abstraction (such as metric, topological, and semantic, sometimes referred to as the spatial semantic hierarchy). The top level typically represents the properties of objects and places and the relations between them as a labeled graph, where the labels can be inferred partly based on an ontology capturing common sense knowledge by using is-a and has-a relationships. Jensfelt et al have now shown how to attach probabilities to such a map. The representation then becomes a chain graph. Hanheide et al showed in the seminar how to reason with such a map, both to infer the probabilities of particular labels and also to plan courses of exploratory action. An alternative approach by Zender et al is to use non-monotonic reasoning to draw and withdraw inferences in a logical framework.

Open questions include:

1. What are the advantages and disadvantages of the logical and probabilistic approaches, particularly concerning the time complexity of inference and the reliability of the results? Does probabilistic inference scale badly with the size of the graph representing default knowledge?
2. How should probabilities be incorporated into such representations of categorical knowledge in a principled manner? Is it feasible to expect us to be able to populate such complex structures with probabilities grounded in real data prior to operation? What do google – vision/machine learning type approaches have to offer us here?

B. Are Partially Observable models unnecessarily hard?

Partial observability as modelled by the stochastic process community, e.g. POMDPs is both a blessing and a curse. It provides us with a rather general framework for reasoning under uncertainty about the value of information. But even the most benign classes of POMDPs are at best in NP, and even practical approximate algorithms for them scale only to hundreds of states. In addition learning in POMDPs is hard. This begs the question are POMDPs too general a formulation, Here we felt there were two issues. What proofs can we provide about existing simple approaches such as greedy approaches to information gathering? Second what alternatives are there to POMDPs as a formalism, including special cases that are tractable. We deal with each in turn:

Open questions include:

1. Performance bounds for simple methods: Work on topics like sub-modularity and adaptive sub-modularity shows that it is sometimes possible to prove that we can do well with simple algorithms. There are two main open questions with sub-modularity for exploration:
   a. Does some form of sub-modularity apply where the experiments depend on the state, and where the state can be changed by the robot's actions?
   b. Does some form of sub-modularity apply where the robot physically alters the external world state, e.g. by moving objects, or alternatively alters the belief states of other agents (e.g. via dialogue)?
2. More restrictive models: Some workers (e.g. Kaelbling) have proposed dividing their problems up into a probabilistic bit and a non-probabilistic bit, e.g. just using POMDPs to model small patches of state uncertainty in an otherwise observable world. Is this an approach in which it is possible to derive bounds on how sub-optimal performance is? Work by Hanheide et al presented at the seminar also touched on algorithms that split

the exploration problem into a probabilistic and a non-probabilistic part. What kinds of algorithms are there that take broadly this approach?

C. Will data based machine learning make the problem go away?

One major motivation for autonomous exploration is because of the limited experience of the robots that we use in practice. However, there is a lot of work on learning for vision from the vast amounts of data available on the internet. Is it actually possible that very large available data sets, e.g. google for vision or very large scale SLAM will make much of the exploration problem go away? In other words we asked the question whether in the long run there are solutions other than autonomous exploration for filling out the knowledge bases of our robots sufficiently well to enable them to perform a wide variety of tasks.

D. How hard can open worldness be in practice?

Modern planning approaches are very efficient for closed worlds. One difficulty when applying such planning approaches to exploration problems is that for many exploration problems the robot has to reason about open worldness. This could be reasoning about the addition of new objects or places as they are found, or reasoning about changes to the types of entities it may have in its knowledge base. There are two ways to deal with this:

1. Soft open worldness: We try to retain the benefits of efficient algorithms for reasoning in closed worlds by allowing only a finite amount of open worldness, e.g. a finite number of additional objects or categories in the planning domain, each of which have some unknown properties. These are essentially slots that we can fill via planning actions.
2. Hard open worldness: we essentially abandon current efficient methods for reasoning in closed worlds and allow the planning domain to be modified on the fly during the planning process. This would require a whole new branch of planning research.

E. How do we fairly compare apples and oranges as the varieties of exploration opportunities rise in number?

One of the major themes emerging from the seminar is the issue of how intrinsic and extrinsic rewards are fairly compared. Tishby presented one way to approach this problem at the seminar. However, there is also the issue of how various sources of intrinsic reward should be fairly compared while leading to efficient learning. The frameworks of Polani and Schmidhuber address these problems, but since the only extant empirical results are for grid worlds or similar, we are not yet clear about how the data efficiency of these as the number of possible learning activities increases in real robot worlds. In robot implementations at the moment the comparison is done heuristically using fudge factors. This is unsatisfactory. This lead us to the general open question:

1. As possibilities for learning activities increase do existing ways of comparing value break. How can we fairly, and yet tractably compare very different learning goals as the number of different types of goal rises? It seemed to us there were two possible answers:
   a. Ground rewards for information types in a known distribution over future tasks.
   b. The notions of empowerment and compression are perfectly sufficient and efficient.

4.3 Group 3: Optimality and sub-optimality in exploration

Group members were Peter Auer, Peter Dayan, Richard Dearden, Charles Fox, Mark Humphries, Andreas Krause, Manuel Lopes, and Ruben Martinez-Cantin.

Solved problems: Although more can always be done, there are a number of fairly strong results for bandit problems and MDPs:

1. Bandits: there are Gittins indices in a (brittle) Bayesian case; regret bounds with stochastic arms (at log t cost), or with adversarial arms ($\sqrt{t}$ cost); good ideas for continuous bandit problems with smoothness assumptions (e.g. Gaussian process payoffs) or contextual bandits (with useful 'input' information; the regret bounds being norm dependent). Most methods are based on forms of optimism. Of course, not everything about bandits is solved: particularly for complex action spaces.
2. Reinforcement learning in MDPs: the two main successful models are regret-like bounds, which achieve regret of $\sqrt{t}$ against the optimal policy in the non-adversarial setting, and phasic (E^3) or continuously exploring R_{max} , which provide guarantees about how long it takes to provide a near-optimal policy, without worrying about the regret along the way. There are also many results on MDPs with function approximation – key here is whether the optimal value function is in the approximation class; there are then results on the loss compared to an optimal policy. Most methods require knowledge of the maximum reward or policy payoff (or tricks to get around this).
3. Continuous time/space linear quadratic stochastic control: these problems are also easy.
4. Information gathering POMDPs: in which actions do not affect the state at all, may have special solutions (like the SPRT and cases of submodularity); extensions to planning informative sampling trajectories might be possible.

Unsolved problems: These fall into different categories. First are problems that are well-accepted by the community as being hard, and as having attracted substantial thought, if not results:

1. POMDPs: we couldn't think of any generally successful approach to POMDPs. Indeed, it was reported that not only are POMDPs NP-hard, but also approximations to POMDPs are NP-hard; there is no simple, but non-trivial, subclass of POMDP that is easy, apart perhaps for information gathering POMDPs; and methods such as factored representations that work well for MDPs, are much less effective for POMDPs. Note that even the planning problem is hard for general POMDPs, given knowledge of the model.
2. Continuous time problems: most work has been done in discrete time problems. There is much less known about exploration in continuous time problems, notably semi-Markov decision-problems (SMDPs), for instance when agents can choose when to act. Problems that haven't been fully explored or perhaps completely precisely formulated:
3. Problem classes: we agreed that there is a need for a suite of benchmark problems that probe different aspects of the difficulty of exploration in various ways.
4. Multi-task learning: many issues surrounding curiosity or the early acquisition of competence could be illuminated in a multi-task setting, with the statistical structure of the commonality between tasks determining what the benefits could be of the competence. We therefore need models for this statistical structure and generalization.
5. Satisficing: we had some discussion about the effect of resource limitations in space and/or time on finding and executing optimal policies. There are some models for speed-accuracy tradeoffs in simple information-gathering tasks (as in the SPRT), but they don't seem to have been generalized to MDPs. Actual bounds on resources (e.g. different polynomial orders) can depend sensitively on the model of computation supported by the hardware, and so are hard to assess. There is work in general on parallelizability (circuit complexity), but we didn't know of work on MDPs.

6. Risk-sensitivity: there is some work on risk-sensitive solutions – using different norms, or penalizing variance and higher order moments as well as means (Singh; Kappen). This could also be formulated in terms of multiple objectives ('maximize reward, but don't get near to the following states'). Observations
7. Computationally simple heuristics like entropy reduction can work very well, at least as well as much more complex methods such as depth-limited model-search. Characterizing when simple heuristics work well is a key, if somewhat amorphous, research question.
8. For problems satisfying submodularity, being greedy about choice can be near-optimal. In some applications, the main objective function of interest may not be submodular. In these applications, it may be of interest to identify surrogate objectives, which are submodular and thus can be greedily optimized, while still approximating the true objective. The relevant submodular functions may not be as simple as entropy reduction.
9. There was some debate about the evolutionary importance of certain sorts of learning and planning. Animals are endowed with extremely strong heuristics (rats act as if things above them are threats; things below them are edible; we have complex ways of assigning salience to sensory inputs) – perhaps these, and other sources of priors such as environmental affordances of object classes, could provide a powerful framework for avoiding doing too much learning.
10. Nobody knew of quantum algorithms for fast (PO)MDP solving . . .

4.4 Group 4: Unifying themes

The working group consisted of John Tsotsos, Tony Cohn, Janusz Starzyk, Danijel Skočaj, and Aleš Leonardis.

Summary: The working group discussed more general issues such as:

1. What do the approaches described in the talks have in common?
2. Is there a unifying framework or topology for the approaches described in the talks?
3. What are the open questions and solution classes?

A lively exchange of diverse opinions converged on a consensus that two themes have been prevalent at the seminar, namely, a rather theoretical treatment of reinforcement learning and, on the other hand, presentations of real application scenarios. While the former excel in abstract models and rigorous treatment of computational issues, the examples shown to support the theoretical findings are often based on unrealistic assumptions and limited to graphical animations and simulations. On the other hand, real applications demonstrated how noise and various constraints, stemming from the real-world, can be tackled, but then most of the approaches relied on heuristics and provided almost no theoretical bounds or proofs on the performance. To bring closer together the theory and the practice should be one of the main goals of the field over the next years.

In a similar vein the working group also discussed a dichotomy between general theoretical approaches on the one hand, and specific (cased-based) applications on the other. A translation of general theoretical principles (such as information gain as the driving force for award driving curiosity and exploration) into specific application domains (such as manipulation, navigation, spatial mapping) appears to be hard and significant efforts towards more generalized solutions should be sought.

Finally, the working group also acknowledged a positive influence that interdisciplinary research and scientific insights in neuroscience and psychology can have on the field. In particular, developmental psychology could provide some indications towards the answers to the questions such as “how / when / under which circumstances does the curiosity arise?”, “what are the requirements?”, “what is the role of scaffolding?”.

In summary, the following open questions have been identified:

1. How do we bridge the gap between general theory and specific real-world applications?
2. How do we bridge the gap between low-level (signal, sensor driven approaches) methods and high-level (symbolic, semantic based methods)?
3. How do we achieve generalization and scalability (e.g. shareability among tasks)?
4. What is the relationship between the theories in engineering science and the neuroscience?

Participants

- Peter Auer
Montan-Universität Leoben, AT
- Andrew Barto
University of Massachusets – Amherst, US
- Anthony G. Cohn
University of Leeds, GB
- Peter Dayan
University College London, GB
- Richard W. Dearden
University of Birmingham, GB
- Yiannis Demiris
Imperial College London, GB
- Stefan Elfwing
Okinawa Institute of Science and Technology, JP
- Charles Fox
University of Sheffield, GB
- Martin A. Giese
Univ.klinikum Tübingen, DE
- Marc Hanheide
University of Birmingham, GB
- Mark D. Humphries
ENS – Paris, FR
- Kristian Kersting
Fraunhofer IAIS – St. Augustin, DE
- Andreas Krause
ETH Zürich, CH
- Ales Leonardis
University of Ljubljana, SI
- Manuel Lopes
INRIA - Bordeaux, FR
- Ruben Martinez-Cantin
CUD – Zaragoza, ES
- Leo Pape
IDSIA - Lugano, CH
- Jan Peters
MPI für biologische Kybernetik – Tübingen, DE
- Daniel Polani
University of Hertfordshire, GB
- Jürgen Schmidhuber
IDSIA – Lugano, CH
- Frank Sehnke
ZSW – Stuttgart, DE
- Danijel Skocaj
University of Ljubljana, SI
- Aaron Sloman
University of Birmingham, GB
- Janusz Starzyk
Ohio University, US
- Naftali Tishby
The Hebrew University of Jerusalem, IL
- John K. Tsotsos
York University – Toronto, CA
- Eiji Uchibe
Okinawa Institute of Science and Technology, JP
- Ales Ude
Jozef Stefan Institute – Ljubljana, SI
- Patrick van der Smagt
DLR Oberpfaffenhofen, DE
- Markus Vincze
TU Wien, AT
- Jeremy L. Wyatt
University of Birmingham, GB
- Jure Zabkar
University of Ljubljana, SI