Novelty Detection and Learning Drives

Deliverable D5.1 of the EU-funded Integrated Project
“IM-CLeVeR – Intrinsically Motivated Cumulative Learning Versatile Robots”,
Contract n. FP7-ICT-IP-231722

Lead Contractor for this Deliverable: UU
Other Contributing Beneficiaries: CNR, FIAS, IDSIA, CNR-Barto, AU, USFD

Deliverable due: 30 April 2011
Deliverable submission: 30 April 2011
Funding Institution: European Commission (European Union)
Project Officer: Cécile Huet
Funded under: Seventh Framework Programme

Start: 01/01/2009 Termination: 30/04/2013

Cite this document as:
<table>
<thead>
<tr>
<th></th>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>INTRODUCTION ..................................................................................................</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>TASK 5.1: INTRINSIC AND EXTRINSIC MOTIVATIONS FOR CUMULATIVE LEARNING ..........</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>TASK 5.2: NOVELTY BASED FORMATION OF ACTIONS: MODELLING THE “JOYSTICK TASK” .</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>TASK 5.3: NOVELTY DETECTION BASED ON HABITUABLE NEURAL NETWORKS ..................</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>TASK 5.4: NOVELTY BASED DEVELOPMENT OF A REDUNDANT SENSORY-MOTOR SYSTEM ...</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>TASK 5.5: INFORMATION-THEORY INDEX AS SOURCES OF NOVELTY SIGNALS ..................</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>ONGOING COLLABORATION BETWEEN PARTNERS IN THE WORK OF WP5 ........................</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>CONCLUSION ..................................................................................................</td>
<td>49</td>
</tr>
<tr>
<td>8</td>
<td>REFERENCES .................................................................................................</td>
<td>50</td>
</tr>
</tbody>
</table>
0. Introduction to the deliverable

This document presents Deliverable 5.1 of the IM-CLeVeR (Intrinsically Motivated Cumulative Learning Versatile Robots) EU FP7 project. It represents one of two deliverables from Workpackage 5 (Novelty Detection and Drives for Autonomous Learning). The second deliverable, D5.2, is due in month 36 of the project.

Workpackage 5 of the IM-CLeVeR project is divided into five tasks with a total budget effort of 157 person months. All tasks have been active during the 2nd year of the project. The University of Ulster (UU) is the partner leading the WP.

WP seeks to address a number of objectives. Task 5.1, led by CNR, aims to produce a broad theoretical framework on the relationship existing between intrinsic and extrinsic motivations. Task 5.2, led by USFD, attempts to understand the brain mechanisms underlying the development of novel sensorimotor skills and action-outcome associations. In Task 5.3, led by UU, novelty detection algorithms and architectures that lead a robot to focus attention on those parts of the environment that appear particularly interesting are developed; such algorithms will be exploited to focus resources and learning efforts of the robots on objects which may maximize robot’s autonomous learning. Algorithms capable of detecting the level of development of a system’s sensorimotor capabilities in order to trigger new developmental stages of behavioural coordination are explored in Task 5.4. Finally Task 5.5 implements, and demonstrates the viability of a very general and principled approach to curiosity and novelty detection based on concepts from algorithmic information theory.

This document, D5.1, reports on the preliminary achievements of WP5 on novelty detection. Substantial progress is reported on all five tasks. The structure of the document is as follows. A concise summary of the achievements in each task is reported in sections 1-5; Section 6 summarises ongoing collaboration between partners in pursuit of the goals of this WP. Section 7 presents the conclusion to the deliverable while section 8 lists the references.

Throughout the text, research papers reporting on the WP work which have been published or are in preparation by the partners are highlighted in bold.
1. Task 5.1: Intrinsic and extrinsic motivations for cumulative learning (CNR-ISTC-LOCEN, CNR-Prof. Barto, USFD, FIAS)

1.1. Introduction to “intrinsic motivations”

This task is about theories and models on intrinsic motivations and their relation to extrinsic motivations. Before considering the contents of the task, it is useful to present a terminological clarification valid also for the rest of the deliverable. This clarification also allows us to introduce the main types of intrinsic motivation mechanisms used and studied within the project. When we wrote the project, we used the term “novelty” as a proxy for all possible intrinsic motivations and so called the whole WP5 “Novelty detection”. The reason was that the WP was led by UU and Prof. Nehmzow that had developed various systems based on novelty detection mechanisms, that is, mechanisms capable of evaluating the familiarity/novelty of percepts. Now it has become clear within the project that there can be several different possible intrinsic motivation mechanisms. Beyond those based on novelty detection, there are those based on prediction errors: these are related to “surprise”, i.e. the violation of some expectations, rather than “novelty”. These mechanisms are often used in the models of IDSIA, with a particular stress on the decrease of prediction error, and CNR-ISTC-LOCEN and USFD, with a particular stress on the prediction error levels. A third type of intrinsic motivations mechanisms are those based on competence progress, namely the capacity to accomplish a given task. These mechanisms are often used in the models of CNR-ISTC-LOCEN and CNR-Barto.

1.2. Overall objectives of the task

The overall goal of task 5.1 is to investigate the relationships and differences between intrinsic and extrinsic motivations and to provide new hypotheses on how motivations, in particular intrinsic motivations, might be implemented in real brains so as to support cumulative learning. This goal is also important to develop new algorithms and architectures to build artificial systems and robots having a high degree of flexibility and autonomy.

To accomplish these overall goals, the task pursues these specific objectives:

a) Clarifying the conceptual difference between extrinsic and intrinsic motivations;

b) Understanding the brain mechanisms that support both kinds of motivations and the complementary roles that they play in driving animal cumulative learning;

c) Proposing new hypotheses and models on the functions and mechanisms of different kinds of intrinsic motivations.

In order to pursue these objectives, the work of this task relied upon different methodologies including theoretical analysis, the study and review of neuroscientific literature, and the development of computational models with different levels of biological constraints. In particular, the research carried out within this task includes the following seven contributions, which will be described in detail in the following sections:

1. Theoretical analysis of extrinsic and intrinsic motivations (CNR-ISTC-LOCEN): this part explains the functions of the two types of motivations within an evolutionary and neuroscientific perspective.

2. Theory and models of extrinsic motivations (CNR-ISTC-LOCEN): this part clarifies some core concepts on extrinsic motivations, in particular stressing those relevant for the integrated models of the demonstrator CLEVER-B.

3. Theory on functions and mechanisms of intrinsic motivations (CNR-ISTC-LOCEN): this second theoretical work contributes to clarify the functions and mechanisms of intrinsic motivations, and dissipates various false problems related to intrinsic motivations.

4. Models of unexpected events as intrinsic reinforcements for cumulative learning (CNR-ISTC-LOCEN): this work presents a bio-inspired model that starts from some core concepts of the neuroscientific theory of USFD on intrinsic motivations (prediction-error based mechanisms), and specifies/translations them into a computational model that allows a
simplified kinematic simulated robot to undergo basic forms of cumulative learning.

5. Models of TD-error as basis for competence-based intrinsic motivations for deciding when to learn what (CNR-ISTC-LOCEN): this work focuses on competence-based intrinsic motivations, a field of intrinsic motivations that so far has not been studied much. In particular, the work shows how the TD-error of standard reinforcement learning models can be used as an index of competence acquisition, i.e. as an important instance of competence-based intrinsic motivation capable of generating learning signals that can allow a system to dedicate its time and resources to acquire skills for which the learning rate is high.

6. Intrinsic motivation: reinforcement learning, evolution, competence progress (CNR-Barto): this work is complementary and strengthens some of the works of CNR-ISTC-LOCEN along two threads. The first, related to the work 1 above, proposes an analysis on intrinsic motivations based on an evolutionary computational framework and a specific model. This work argues that there is a continuum between extrinsic and intrinsic motivations, in particular that the best reward functions for organisms can incorporate to various degrees fitness-enhancing events (e.g., the ingestion of food) and also other salient events that might pave the way to the acquisition of skills directly related to fitness-enhancing events (e.g., succeeding to open a container). The second thread of research worked on the same problem of the work 5 above, and proposes an alternative algorithm to support competence-based intrinsic motivations based on the change in time of the evaluations of a reinforcement learning algorithm.

7. Inferring rewards in inverse reinforcement learning problems (FIAS): this work has a perspective different from the previous ones in that it explores a novel problem related to intrinsic motivations: how an observer external to a given system can infer the reward function of that systems by relying upon information-theory principles. This possibility is important as it can allow to ascribe a give reward function to a system, e.g. intrinsic or extrinsic, based only on the external observation of its behaviour.

1.3. Research work to date

1.3.1 Theoretical analysis of extrinsic and intrinsic motivations

Background and open issues related to IM-CLeVeR

One of the most important challenges of the project IM-CLeVeR is to clarify what intrinsic motivations (IM) are, and how they relate to extrinsic motivations (EM). This section introduces the theoretical framework that CNR-ISTC-LOCEN is developing to frame IM and establish their relationship to EM. This framework is important to build the conceptual tools used to investigate the specific mechanisms and models related to IM.

Specific goals

The goal of this work is to conceptually clarify the notion of intrinsic (as opposed to extrinsic) motivations.

Approach and methods

The research from CNR on IM is carried out within an evolutionary perspective, similarly to what is done by Prof. Andrew Barto. However, Prof. Barto aims to show the continuum along a spectrum existing between EM and IM, whereas CNR tries to identify the two ends of such spectrum.

Results

Here we just briefly sketch our current view on IM and EM: for the details, see Baldassarre (in press).

Learning signals generated by motivations: IM and EM have a different distance from fitness-enhancing events

Biological fitness is the ultimate cause of brain, body, and behaviour of organisms. Motivations
generate learning signals and also energize and contribute to select behaviour. The learning signals
generated by motivations, on which we focus here, have a variable distance in time from the events
that enhance biological fitness. In this respect, motivations vary along a spectrum that goes from
motivations that work on events closely related to fitness, in particular events related to the regulation
of homeostatic needs, to motivations more distant from fitness, related to the acquisition of general
knowledge and competence (skills) that are only later exploited to increase fitness. We can call the
motivations at the two ends of the spectrum respectively "extrinsic motivations" (EM) and "intrinsic
motivations" (IM). Both biology and modelling (see section 1.3.6) show that between the two
extremes there are various interesting cases that posses mixed features of EM and IM. Here we focus
in clarifying the two extremes.

**Extrinsic motivations** are motivations that drive learning and exploitation of knowledge and
competence on the basis of the levels and variations of homeostatic needs. Such knowledge and
competence have the adaptive function of directly increasing survival and reproduction (fitness) of
organisms.

**Intrinsic motivations** are motivations that drive learning and activation of knowledge and
competence on the basis of the levels and variations of such knowledge and competence themselves.
These knowledge and competence have the adaptive function to satisfy the survival and reproduction
needs of the organisms, but, differently than EM, these needs are not present when such knowledge
and competence are acquired, so IM have to rely upon the success in the acquisition itself instead of
its effects on homeostatic needs.

**Characteristics of Extrinsic motivations (EM):**

1) Evolution: The body/brain structures subserving EM emerged early in evolution. So they are
equally important in all species of organisms, although their complexity is comparable to the overall
sophistication of the organisms.

2) Function 1: EM have the adaptive function (hence emerged) to generate extrinsic learning signals
that guide the acquisition of knowledge and skills directed to decrease homeostatic needs.

3) Function 2-3: EM also have the adaptive function (hence emerged) to bias the selection and to
energize the execution of skills and actions directed to decrease homeostatic needs.

4) Mechanisms: The mechanisms of EM generate the selection and energization of behaviour, and the
learning signals leading to the acquisition of knowledge and competence, on the basis of the effects
that the events caused by them in the environment have on the body. In particular, the learning signals
are generated by rewards (appetitive events) that decrease the organisms' homeostatic needs, and
punishments (aversive events) that increase such needs.

5) Dynamics: EM related to a certain situation or object cease when the homeostatic need that can be
satisfied by them ceases, and resume when the homeostatic need comes back.

**Characteristics of Intrinsic motivations (IM):**

1) Evolution: The body/brain structures underlying IM emerged late in evolution. So they are more
sophisticated and important in higher organisms (mammals, primates, humans).

2) Function 1: IM have the adaptive function (hence emerged) to generate intrinsic learning signals
that guide the acquisition of knowledge and competence.

3) Function 2-3: IM also have the adaptive function (hence emerged/are designed) to energize and
focus action, attention and cognitive resources of organisms on the acquisition of particular pieces of
knowledge and competence.

4) Mechanisms: The mechanisms of IM generate the learning signals leading to acquire knowledge
and competence, and energize and focus the activities related to this acquisition, on the basis of events
taking place in the brain (controller). In particular, the regulations and learning signals generated by
IM depend on the levels and the variations of knowledge and competence that are being acquired.
Importantly, the knowledge and competence so acquired is used only later to reduce homeostatic
needs that, this is the key point, are not present at the time of their acquisition.
5) Dynamics: IM related to a certain situation or object tend to cease forever if the organism cannot acquire any further knowledge or competence in relation to them.

Similar definitions apply to IM in robots by substituting evolution with the designer of the robot.

Figure 5.1.1: A sketch of the key features that characterise IM and EM. On the left, IM operate especially during the childhood of organisms and generate learning signals that allow organisms to acquire knowledge and skills on the basis of indexes that are produced within the brain itself and that are related to the acquisition of knowledge and skills ("intrinsic learning signals"). These knowledge and skills (e.g. the "actions" represented here) can then be reused in a later stage of life, e.g. during adulthood (right part of the picture), to act in the world and increase fitness. Fitness increase happens if what is done in the environment positively affect the body. These effects on body are also used by EM to directly generate a second type of learning signals ("extrinsic learning signals"). Evolution leads to evolve the whole body and brain of the organism and also the mechanisms implementing the IM themselves.

Advancement of work and relation to other tasks

We have currently published a conference paper on this (Baldassarre, in press). The conceptual analysis of what intrinsic motivations are and how they differ from extrinsic motivation is clearly central to the whole project. In fact, this work is furnishing an important contribution to the theoretical debates on IM carried out within the consortium, a theoretical background for more detailed analyses of functions and mechanisms related to IM (e.g., other sections within this task, like 1.3.2), and is helping to design and implement models that involve both EM and IM, in particular CLEVER-B models. Furthermore, this work is directly related to the view of Barto described in section 1.3.6.

1.3.2. Theory and models of extrinsic motivations

Background and open issues related to IM-CLeVeR

Extrinsic motivations involve various portions of the brain, including the hypothalamus, orbitofrontal cortex, and amygdale (Amg). The latter plays a pivotal role for extrinsic motivations as it is the locus of most Pavlovian associations; these allow the brain to associate a multitude of innate reactions (or UR, unconditioned responses: from orienting and attention to approaching and avoidance, from salivation to freezing and startling, from the adjustment of internal body states to the modulation of neuromodulators directed to broadly regulate the whole mode of functioning of brain and contribute to produce learning signals) to innately relevant stimuli (or US, unconditioned stimuli: e.g., food, pain, etc.). It also possesses associative (Pavlovian or classical-conditioning) processes which allow it to transfer the innate reactions to any type of neutral stimulus (or CS, conditioned stimulus) in the environment. These associations are one of the major means through which the brain assigns a subjective value to otherwise neutral world stimuli.

The relevance of this work with respect to the project is threefold (for each of the following points, see Figure 5.1.2): first, the study of the relationships between body state information, Pavlovian
associations, and the regulation of body states is fundamental for the development of the conceptual distinction between extrinsic and intrinsic motivations discussed in the previous section (1.3.1); second, the study of the relationships between Pavlovian associations and brain neuromodulation is fundamental for understanding the brain mechanisms underlying reinforcement learning processes and for developing our new hypothesis (described in section 1.3.4) on the relation between extrinsic and intrinsic reinforcement signals in the brain; third, the study of how pavlovian associations interact with instrumental processes giving raise to goal-directed behavior is at the basis of the modeling work of Task 6.1 which will constitute a key contribution to the CLEVER-B4 final demo, where the Amg will allow the robot to recall the actions acquired through intrinsic motivations on the basis of current biological value assigned to the action's outcome.

Specific Goals

This research aimed at building a general theoretical framework on the role of amygdala in the affective regulation of body, brain, and behavior. This framework plays a fundamental for developing a conceptual distinction between intrinsic and extrinsic motivations, for understanding their relationships within the brain, and for building bio-constrained computational models and algorithms directed to specify both the internal associative mechanisms of amygdala and the role played by amygdala in the regulation of affective responses and cognitive processes.

Approach and methods

The research carried out a systematic theoretical systematization of the literature on amygdala, in particular on: (a) anatomical data; (b) data on behaviour and the effects on it of targeted lesions. On this basis, we built specific bio-constrained models on: (a) the functioning of the internal associative mechanisms of amygdala; (b) the ways in which the amygdala, suitably connected to other brain areas, contributes to implement a number of behaviours (see Figure below).

Results

The theoretical investigation and the specific models allowed us to establish the fundamental hypothesis that Pavlovian associative processes of amygdala are based on three basic mechanisms, which roughly correspond to the three major sub-components of amygdala (see Figure 5.1.2): (a) CEA: associates neutral stimuli (conditioned stimuli) directly to basic, unconditioned responses; (b) BLA: associates neutral stimuli to the unconditioned stimuli that are innately associated to those responses on the basis of co-occurrences experienced during lifetime; (c) MEA: modulates CEA’s and BLA’s representations of stimuli and responses on the basis of internal body states.

Amygdala exploits its associative mechanisms to play an important role within various brain systems. These brain functional systems in which amg plays a role can be grouped in two broad classes on their turn containing various sub-classes: (a) Regulation of affective processes involving: (a.1) Homeostatic regulation of body states via the sympathetic and parasympathetic systems, e.g. for the “optimisation” of energy intake and expenditure; (a.2) broad modulation of brain functioning via the main neuro-modulatory systems supported by the ventral tegmental area (dopamine), locus coeruleus (norepinephrine), and raphe nuclei (serotonin); (a.3) Selection and triggering of unlearned behaviours (mainly implemented via the CEA-brainstem neural pathway). (b) Regulation of cognitive processes that can be subdivided in: (b.1) affective labeling of episodic memory, especially if involving spatial context (mainly implemented via the BLA-hippocampus neural pathway); (b.2) goal-driven/top-down influence of the selection of responses learned with instrumental processes (mainly implemented via the BLA-nucleus accumbens-prefrontal cortex pathway); (b.3) planning and decision making (mainly implemented through interconnections between amygdala and prefrontal cortex).
Figure 5.1.2: A scheme indicating the main functions played by the Amygdala (Amg), one of the most important systems behind the regulation of extrinsic motivations in brain. Notice the core associative mechanisms implemented by the Amg, and the role that Amg plays in the affective regulation of body and brain (body regulation, diffuse brain neuromodulation, triggering of innate behaviours) and in various behaviours (episodic memory, goal-oriented behaviour, and planning). BLA: basolateral Amg; CEA: central nucleus of amygdala; MEA: medial nucleus of amygdala; CS: condition stimulus; US: unconditioned stimulus; UR: unconditioned response. For other acronyms, see Mirolli et al., 2010

Level of advancement of the work and relation to other tasks

We accomplished the aforementioned goals and published papers on a general framework on amygdala (Mirolli et al., 2010) and on specific system-level models of the role played by amygdala in various behavioural processes, for example habits and goal-directed behaviour (Mannella et al., 2010). The general framework and the latter model forms the basis of the modeling work of Task 6.1 and will play a pivotal role in the development of the CLEVER-B4 demonstrator (Task 7.4), where the actions that were learned on the basis of intrinsic motivations have to be selected on the basis of the current value of their outcomes: for example, the value assigned to the food present in one of the boxes of the mechatronic board, processed by the amygdala, will influence the selection of the outcomes in the nucleus accumbens / prefrontal cortex look, thus leading to the robot to press the button which opens the appropriate box for obtaining the food.

1.3.3. Functions and Mechanisms of intrinsic motivations

Background and open issues related to IM-CLeVeR
Different kinds of intrinsic motivations have been proposed both in the psychological literature and in the fields of machine learning and developmental robotics, some of which are based on the knowledge of the agent (on the stimuli that the agent perceives, their properties, and their relationships with the agent's knowledge and expectations) and other are based on the agent's competence (on what the agent is or is not able to achieve through its behaviour).

Specific goals

The goal of this work is to clarify the distinction between knowledge-based and competence-based intrinsic motivations, both with respect to the possible mechanisms that might implement motivations and to the possible functions that they might have in cumulative learning.

Approach and methods

We review both the psychological and the computational modelling literatures on intrinsic motivations from the perspective of the distinction between knowledge and competence, we analyse this distinction at the level of both the mechanisms and the functions of intrinsic motivations and we develop a novel hypothesis on the different functions that knowledge-based and competence-based intrinsic motivations mechanisms might play in the cumulative acquisition of skills of a hierarchical learning system.

Results

The first important point that we make is that the principal function of all kinds of intrinsic motivations consists in allowing the development of a repertoire of actions, rather than of knowledge: knowledge might be important, but only as long as it serves the acquisition and deployment of adaptive behaviour.

As it is typical in the computational literature on intrinsic motivations, we frame the discussion within the computational framework of reinforcement learning, where intrinsic motivations can be conceived as components of the reinforcement signals that drives agent's learning that are not directly related to the task that the agent has to perform. Hence, in this context, we see intrinsic motivations as the task-independent learning signals that are able to drive the cumulative acquisition of a number of different skills that can be then be deployed for maximizing extrinsic (task-dependent) rewards.

But cumulative learning requires not only appropriate learning signals, but also a control architecture that can store skills as soon as they acquired without being subject to catastrophic interference while learning new skills (Baldassarre and Mirolli, 2010). Both the computational literature on multi-skill learning and the psychological and neuroscientific literature on the organization of action in real brains suggest that some form of modularity and hierarchical organization is required for permitting cumulatively. And if one considers the accumulation of skills in a hierarchical system, the problem of identifying good intrinsic learning signals splits in two sub-problems, as different levels of the hierarchy are likely to require different signals. For example, in a system composed by several experts implementing different skills and a selector that arbitrates between them (Mirolli and Baldassarre, in preparation; see also below and the work on hierarchical architectures in WP6 of the project), the problem for each expert consists in identifying which skill to acquire and how, whereas the problem for the selector consists in deciding what to learn and when, i.e. which skill to train in each context.

We contend that both in real organisms and in efficient autonomous robots different kinds of intrinsic motivations might implement these two different functions. In particular, knowledge-based intrinsic reinforcement given by the detection of unexpected events might serve the function of driving the discovery and acquisition of actions by the low-level experts, whereas competence-based intrinsic reinforcements related to the progress in skill acquisition of each expert might underlie the function of driving the decisions of the selector on which skill (expert) to train when.

Advancement of work and relation to other tasks

We have written a paper on this work Mirolli and Baldassarre (in press). Furthermore, beyond building this general theoretical framework, we have been investigating these issues through two complementary lines of computational work: the first (described in the next section 1.3.4) is focused
on how intrinsic reinforcements provided by the detection of unexpected events represents appropriate learning signals for cumulative learning; the second (described in section 1.3.5) is focused on showing how the TD-error signal of a standard actor-critic reinforcement learning architecture can constitute a good competence-based intrinsic motivation for a selector which has to decide which skill to train. This work is also related to the work on hierarchical reinforcement learning architectures developed in Tasks 6.1 and 6.3.

1.3.4. Unexpected events as intrinsic reinforcements for cumulative learning

Background and open issues related to IM-CLeVeR

This work tries to reconcile heated debate in the neuroscientific literature on the biological bases of animal reinforcement learning. This literature is dominated by the hypothesis that the neuromodulator dopamine constitutes a reward prediction error that drives classical and instrumental conditioning phenomena. That hypothesis is based on the striking similarities between the dynamics of phasic dopamine in the basal ganglia and the temporal difference reward prediction error signal (TD-error) postulated by the computational theory of reinforcement learning, and has been successfully driving a huge amount of neuroscientific research on conditioning.

This notwithstanding, the reward-prediction-error hypothesis of dopamine is in contrast with the fact that dopamine is released not only in presence of biological rewards such as food, but also when all sort of unexpected events are detected. On the basis of this fact and of other evidences related to the details of the timing of dopamine release Redgrave and Gurney have proposed an alternative theory according to which dopamine is not a reward-prediction-error driving animal conditioning towards the maximization of extrinsic rewards but rather a sensory-prediction-error signal that drives the discovery of agency (i.e. which are the events that are caused by the animal) and the acquisition of actions. This hypothesis is at the basis of USFD experimental work in WP3 and of the computational modeling work in Task 5.2.

Specific goals

Our work tries to solve the conflict between these two competing views by hypothesizing that both views are (partially) right: dopamine is indeed a reinforcement-prediction-error signal analogous to the computational TD-error, but for a learning system that receives two different kinds of reinforcement: (1) temporary (intrinsic) reinforcements provided by unexpected events and (2) permanent (extrinsic) reinforcements provided by biological rewards. In this way, the same reinforcement learning architecture and learning signal can absolve both crucial functions: (1) driving the discovery and acquisition of novel actions and (2) driving the maximization of extrinsic rewards through the appropriate deployment of instrumental actions.

Approach and methods

To test our hypothesis we developed an experimental set-up in which an artificial embodied system has to solve a task that requires the cumulative acquisition of different skills. In particular, the system is a simulated kinematic robot composed of a two-degrees of freedom arm (shoulder and elbow) and a moving eye, and the task consists of learning to eat the objects (apples) put on a table in front of the robot (see Figure 5.1.3). Since the controller of the arm receives as input what the eye sees (i.e. the position of the object with respect to the centre of the visual field), learning to systematically look at the object is a prerequisite to learning to reach for it, which is in turn a prerequisite for the ability to grasp the object and bringing it to the mouth for eating it.

The controller of the robot has been designed by incorporating several constraints that have been suggested by the biological literature. For example:

- we use actor-critic reinforcement learning architectures, which have been proposed to functionally reflect the action selection and learning mechanisms happening in the basal ganglia;
- system is trained through the temporal difference (TD) learning algorithm, whose learning signal has been proposed to correspond to the phasic activation of the neuromodulator dopamine (DA).
• the modular organization of the controller (one controller for the eye and another for the arm) reflect the modular organization of the basal-ganglia-thalamo-cortical loops implementing reinforcement learning in the brain

• the reinforcement learning signal is unique for both the sub-controller, in accordance with the fact that phasic DA signal is likely to be the same for all sensory-motor subsystems.

In order to test our hypothesis we compare the results of three different experimental conditions, which vary with respect to the reinforcement signals that drive the system's learning:

• only extrinsic reinforcements: i.e. the system is reinforced only when it brings the object to the mouth and eats it;

• extrinsic reinforcements plus reinforcements for each of the sub-tasks: i.e. the system is reinforced also when (a) it foveates the object and (b) touches it with its hand;

• extrinsic reinforcements plus intrinsic reinforcement provided by unexpected events: the reinforcements provided by foveating and touching the object are suppressed by predictors, thus representing intrinsic reinforcements provided by unpredicted events (the activation of the fovea or of the touch sensor of the hand).

![Figure 5.1.3: Set up of the experiment on intrinsic reinforcement for cumulative learning: the system is composed by a two-dimensional arm and a moving eye (dotted square with a fovea at the center). The task is to eat apples (red circle) that are randomly positioned on the table by bringing them to the mouth (red rectangle).](image)

**Results**

The results show that when an agent has to learn a complex skill based on different abilities, if the learning is driven only by extrinsic rewards for the final task, the entire process can be hard and long to learn. However, simply adding rewards for possible sub-tasks doesn't help much because the reinforcements for the sub-tasks can interfere with the learning of the final task. The system that works best is the one which complement the extrinsic rewards provided by the final task with intrinsic reinforcements given by the detection of unexpected events (see Figure 5.1.4). The reason is that unpredicted events are good reinforcement signals for cumulatively discovering and acquiring skills because they are present only when they are needed: as the system learns to systematically produce the events, it starts to predict their occurrence, hence the events cease to be unexpected and reinforcing, and the system can focus on learning something else.

These results support the hypothesis that if a reinforcement learning system learning through the temporal difference algorithm is given both temporary (intrinsic) reinforcements provided by unexpected events and permanent (extrinsic) rewards related to the task the system can cumulatively acquire different skills and solve complex tasks because the intrinsic reinforcements can drive the discovery and acquisition of the actions that are necessary for the maximization of extrinsic rewards.
Figure 5.1.4: Performance (percentage of trials in which the system eats the apple) in the three conditions: A – only extrinsic reinforcement; B – extrinsic reinforcements plus reinforcements for each sub-task; C: extrinsic plus intrinsic reinforcements.

Advancement of work and relation to other tasks

We have published a conference paper (Santucci et al. 2010) and we have produced a draft for a paper to be submitted to a journal (Santucci et al. in preparation). From the biological point of view, this work offers a possible solution to the dopamine puzzle that can be found in the neuroscientific literature on the brain mechanisms underlying animal conditioning as described above. As such, it is related to the experimental work of USFD in WP3 and to the modelling work of Task 5.2. From the computational point of view, this work shows how intrinsic reinforcements provided by unexpected (unpredicted) events might represent powerful learning signals that can guide the cumulative acquisition of increasingly complex skills which we would like to see in our robots. As such, it complements the work done in Tasks 6.1 and 6.3 on hierarchical reinforcement learning and will represent an important input to the integration work done in WP 7, in particular with respect to the CLEVER-B demonstrator (Task 7.4).

1.3.5. TD-error as a competence-based intrinsic motivation for deciding when to learn what

Background and open issues related to IM-CLeVeR

The idea of using (sensory) prediction errors as intrinsic reinforcements had been proposed in the 1990s by Schmidhuber, who abandoned it for the idea of using instead measures of prediction learning progress, on the ground that if prediction errors are reinforcing a system might get stuck in case the environment cannot be predicted. How can a system driven by sensory-prediction-error intrinsic reinforcement signals – which seems to be the case for real animals whose dopamine is triggered by unexpected events – avoid getting stuck in trying to learn and reproduce events that are not predictable and systematically reproducible? We think that the solution can lie in the presence of complementary competence-based intrinsic motivation mechanisms acting at the higher level of the action hierarchy. This work, which is in line with the view presented in section 1.3.3 and complementary to the work presented in section 1.3.4, can potentially provide an important input to the work on hierarchical architectures in WP6 (in particular, Tasks 6.1 and 6.3), and consequently be incorporated in the both the demonstrators CLEVER-B and CLEVER-K.

Specific goals

The goal of this work is to test the idea that the standard TD-error signal might be used not only as the learning signal for the training the experts but also as a competence-based intrinsic motivation for
deciding when to learn which skill. The reason behind this idea is that the TD-error might be a good measure of an expert learning progress in acquiring its skill.

**Approach and methods**

We use a hierarchical reinforcement learning system in which at the low level different actor-critic modules (which we call experts) learn and incorporate different skills, and at the higher level another module (which we call the selector) decides which of the low-level experts controls behaviour and learns. In such a kind of hierarchical system, the problem of unlearnability can be solved at the level of the selector, whose role is to decide which skill to train when. But while the learning of the expert can be driven by external reinforcements, be they extrinsic (and permanent) or intrinsic (and temporary), what is the signal that has to drive the learning of the selector? Our proposal is that the TD-error signal of the selected experts constitutes a very good reinforcement signal for the selector. The reason is that, being a reward-prediction-error, the TD-error can be conceived as a measure of how much an expert is improving in maximizing its rewards, that is, it is a measure of the expert's progress in its competence acquisition. Hence, by being driven by the TD-error of the experts the selector will learn to give the control to the expert that is learning most, thus optimizing the acquisition of skills and avoiding getting stuck in situations where there is nothing that can be learned.

In order to test whether the experts' TD-error as an intrinsic motivation for the selector could indeed drive the system in both maximizing its acquisition of skills and solving the problem of unlearnability we devised the simplest possible experimental set-up. We use a grid-world in which the agent can move north, south, east, and west, and 4 different rewards (one for each of the four system's experts) are given when the agent reaches one of the four corners of the world (see Figure 5.1.5, left). In order to assess our proposal we tested the system in several different experiments and compared it both with a random selector and with a system in which the reinforcement to the selector consists in the absolute value of the expert TD-error, so to show the importance of the fact that the TD-error is not just a normal prediction error (for which the absolute value is normally used), but it is a reinforcement-prediction-error in which the sign is important, as it signals whether the learning system is improving or not.

**Results**

We run different experiments in which we test whether the system is sensitive to several possible sources of variability in skill acquisition. For example, we varied:

- the probabilities of the different rewards for each expert
- the absolute values of the different rewards
- the learning rates of the experts
- the difficulties of reaching each reward
- the relationships between skills (where one depends on the other)

In all these cases, we showed that the selector trained with the expert's TD-error as reinforcement was sensitive to the differences, and appropriately decided which expert to train accordingly. For example, Figure 5.1.5 (right) shows how the system in which different experts have different learning rates starts by training the expert which can learn faster (with higher learning rate); then, as soon as that expert has acquired its skill control is preferentially given to the expert which is learning most (the one with the second higher learning rate); when also the second skill has been completely acquired the system preferentially select the third expert; when also the third skill has been acquired, the system realizes that the last expert does not improve (as it has a learning rate of 0), and hence the selector starts to select the do-nothing action because nothing more can be learned.
Figure 5.1.5. Left: Set-up of the experiments using the TD-error as a competence-based intrinsic reinforcement: the agent can move north, south, east, and west (arrows), and at the four corners of the grid-world there are four different rewards, one for each expert. Center: average proportion of times for each expert of being selected with random selection along trials; ‘No action’ corresponds to the choice of selecting no expert and doing nothing. Right: same data as in the center for the experiment in which the selector is trained on the basis of the TD-error of the selected expert, and the different experts i have different learning rates (lr), where lr4 > lr3 > lr2 > lr1 = 0

Advancement of work and relation to other tasks

When preparing all the results for writing up the paper (Mirolli and Baldassarre, in preparation), we realized that although the system was sensitive to the learning progress of the experts and was able to behave accordingly, not in all cases its apparently clever behaviour resulted in a speed of learning higher than that of a system which selected expert randomly. This is the reason why we do not attach a paper or a draft for this work. We are currently working on the details of the algorithm (in particular on the selector) in order to maximize the efficiency with which it is able to switch between training different skills. Furthermore, we are also planning to test the same algorithm in a robotic scenario with continuous state and action spaces. This work is related to the work by Stout and Barto discussed in section 1.3.6 on another competence progress intrinsic motivation. In particular, while their proposal might provide a more precise estimate of expected learning progress, our proposal is much simpler, much computationally less expensive to calculate, does not require, for each expert, the duplication of the policy (one for exploration and another, completely different, for later exploitation), and is much easier to apply to robotic system with continuous action and state spaces. In any case, the results of this work will be important for the work on hierarchical reinforcement learning architectures in WP6 (in particular Tasks 6.3 and, possibly, 6.1), and may be used in both of the demonstrators: CLEVER-B and CLEVER-K.

1.3.6. Intrinsic motivation: reinforcement learning, evolution, competence progress

Background and open issues related to IM-CLeVeR

Developing a broad theoretical framework for intrinsically motivated artificial agents is a key challenge for IM-CLeVeR. A basic assumption is that computational reinforcement learning (RL) is an appropriate framework for incorporating analogues of intrinsic motivation into artificial agents, although this assumption has not been critically examined in light of past theories of motivation in Psychology. Further, a precise characterization of the distinction between intrinsic and extrinsic reward signals is still a matter of controversy. Taking an evolutionary perspective on this question clarifies some aspects of this controversy, while leaving others still to be resolved. Another open issue is related to the competence-progress view of intrinsic motivation: the problem of choosing, at any given moment, which of a number of skills the agent should attempt to improve.

Specific goals
Research at UMass Amherst, in collaboration with colleagues at the University of Michigan, was directed toward improving our understanding of intrinsic motivation by examining the assumption that RL is a suitable basis for incorporating intrinsic motivations into artificial agents, by continued exploration of the implications of an evolutionary view of the origin of brain reward signals, and by developing a computational model of competence-progress motivation.

**Approach and methods**

Extensive effort was devoted to relating computational RL to psychological theories of motivation, with special attention to theories touching on intrinsic motivation. A chapter was prepared on this subject for publication in the IM-CLeVeR Roadmap book. Further computational experiments were conducted with the optimal reward framework that captures the pressure to design good primary reward functions that lead to evolutionary success across a distribution of environments. Spaces of possible reward functions were defined and searched to find reward functions that produced the highest agent fitness as measured across an ensemble of environments that shared some features but varied in others. In this period, experiments were extended to 1) emphasize the generality of the optimal reward framework by using a model-based learning agent in non-Markovian environments instead of the model-free Q-learning agent in the Markovian environments of earlier experiments, and 2) to demonstrate the emergence of optimal reward functions that are contingent on features of the internal environment of the agent rather than features of the external environment. Related effort was devoted to employing evolutionary programming instead of brute-force search in the optimal reward experiments. Effort was also directed toward improving understanding of competence-progress motivation through experiments with an approach to deciding when an agent should shift its behaviour to practice a different skill.

**Results**

The Roadmap book chapter argues that RL is particularly appropriate for bringing learning together with what in animals we call motivation. It further argues that RL is particularly well suited for incorporating principles of intrinsic motivation into artificial agents. Extensive discussion of past psychological theories of motivation suggests some new direction for computational research. Results from further experiments with the optimal reward framework extended previous results by showing how internal informational variables tend to play large roles in optimal reward functions. Again using a simple, but this time non-Markovian, foraging environment (Fig. 5.1.6), the best reward function positively rewards the activity of eating, but the agent’s internal environment—which is invariant across the distribution over external environments—provides an inverse-recency feature. The best reward function exploits this feature to intrinsically reward activities that lead to the agent experiencing state-action pairs it has not visited recently, leading to systematic and persistent exploration. This exploration, in turn, distally produces much greater fitness than achieved by an agent using the fitness-based reward. The application of evolutionary programming to the optimal reward framework demonstrates the possible scalability of reward function search to be of use in practical problems.

![Fig. 5.1.6.](image)

*Fig. 5.1.6. Each foraging environment is a 3x3 grid arranged in (row) corridors. The food represented by a worm appears at the rightmost end of a corridor. The agent represented by a bird has the usual movement actions in the four cardinal directions as well as an eat action when co-located with the worm. Crucially, once the agent eats a worm, a new worm appears at a random corridor-end location and the agent cannot see the worm unless co-located with it. These foraging environments are non-*
**Markovian unlike the boxes environments of previous experiments.**

Computational experiments with CPM, a mechanism of competence-progress motivation, in a simple simulated domain demonstrates that it outperforms a naive agent, achieving higher competence faster by focusing attention and learning effort on skills for which progress can be made while ignoring those skills that are already learned or are at the moment too difficult. This mechanism is similar to that proposed by Schembri, Mirolli, and Baldassarre (Proc. of the 6th Internl. Conf. on Development and Learning, 2007; Proc. of the 7th Internl. Conference on Epigenetic Robotics, 2007). A key difference between CPM and this work is that CPM uses the ΔV algorithm in each skill (Simsek & Barto ICML 2006), while Schembri et al.’s skills use regular RL. This means that Schembri et al.’s selector will choose (myopically) the skill which is learning the most while trying to maximize pseudo reward, rather than planning and behaving to maximize longer-term competence progress.

**Advancement of work and relation to other tasks**

This work benefits all aspects of the IM- CLeVeR project by providing a new perspective on the nature of intrinsic motivation and why it has evolved. The competence-progress architecture is closely related to that described in section 1.3.5, and was the basis of extensive interaction.

**1.3.7. Inferring rewards in inverse reinforcement learning problems**

**Link and relevance to the project of the presented work**

Our work in this task involves models of motivation and learning inferred from behavioural experiments. In order to do this, we formulated the problem as preference elicitation in a Bayesian setting. More precisely, we observe an agent acting within an environment but we are unsure about its motivations. We analyse those motivations by postulating that the agent is obtaining a (hidden to us) sequence of rewards, which we then infer. To test the effectiveness of the approach, we examined inverse reinforcement learning problems, where our solutions performed uniformly better than state-of-the-art inverse reinforcement learning algorithms.

**Specific topic and problems tackled by the model**

Preference elicitation is a well-known problem in statistical decision theory (Friedman and Savage, 1952). The goal is to determine, whether a given decision maker prefers some events to other events, and if so, by how much. The first main assumption is that there exists a partial ordering among events, indicating relative preferences. Then the corresponding problem is to determine which events are preferred to which others. The second main assumption is the expected utility hypothesis. This posits that if we can assign a numerical utility to each event, such that events with larger utilities are preferred, then the decision maker’s preferred choice from a set of possible gambles will be the gamble with the highest expected utility. The corresponding problem is to determine the numerical utilities for a given decision maker.

Preference elicitation is also of relevance to cognitive science and behavioural psychology, where a proper elicitation procedure may allow one to reach more robust experimental conclusions. There are also direct practical applications, such as determining customer preferences. Finally, by analysing the apparent preferences of an expert while performing a particular task, we may be able to discover behaviours that match or even surpass the performance of the expert in the very same task.

We use the formal setting of preference elicitation to determine the preferences of an agent acting within a discrete-time stochastic environment. We assume that the agent obtains a sequence of (hidden to us) rewards from the environment and that its preferences have a functional form related to the rewards. We also suppose that the agent is acting nearly optimally (in a manner to be made more rigorous later) with respect to its preferences. Armed with this information, and observations from the agent’s interaction with the environment, we can determine the agent’s preferences and policy in a Bayesian framework. This allows us to generalise previous Bayesian approaches to inverse reinforcement learning.

In order to do so, we define a structured prior on reward functions and policies. We then derive two
different Markov chain procedures for preference elicitation. The result of the inference is used to obtain policies that are significantly improved with respect to the true preferences of the observed agent.

Methods and key results

Our most recent work is currently under review (Rothkopf and Dimitrakakis, 2011) or under preparation for submission (Dimitrakakis et al., 2011). Prior work involved results on approximate planning under uncertainty (Dimitrakakis and Lagoudakis, 2008b,a; Dimitrakakis, 2010), where we derive nearly optimal planning algorithms that use a relatively small amount of computation.

Finally, we have some experimental and theoretical results linking intrinsic motivations and attention to decision theory through an interesting class of abstract problems (Rothkopf and Dimitrakakis, 2011; Dimitrakakis et al., 2011).

1.4 Conclusion to task

Work on this task has already provided several important results:

With respect to the conceptual distinction between intrinsic and extrinsic motivations, the debate, both within the project and in the whole community, is still ongoing, as it is to be expected given the fact that we are dealing with a very controversial topic and we are trying to cross disciplinary boundary and put together knowledge and approaches from psychology, neuroscience, machine learning, and robotics. In this context, we have clearly produced significant contributions to the state of the art (see sections 1.3.1 and 1.3.6).

With respect to extrinsic motivations, we have developed a theoretical framework regarding the affective regulation of body, brain, and behavior (see section 1.3.2) which has already proved useful for the development of several aspects of the project, including the development of the conceptual distinction between extrinsic and intrinsic motivations, the development of our new hypothesis on the role of dopamine in reinforcement learning and its relationships with extrinsic and intrinsic motivations, and the development of bio-constrained computational models of the hierarchical organization of goal-directed behavior.

With respect to intrinsic motivations, we have already provided at least three important contributions to the state of the art: (1) we have clarified the distinction between knowledge-based and competence-base intrinsic motivations and proposed that different kinds of intrinsic motivations might play different functional roles in driving a hierarchical learning system in the cumulative acquisition of a number of different skills (see section 1.3.3); (2) we have proposed a new hypothesis (supported by a computational model) on the role of dopamine in conditioning which reconcile the two competing theories currently available in the neuroscientific literature (see section 1.3.4); (3) we have shown that intrinsic motivations based on the (progress of) competence can be effective in deciding what to learn when (see sections 1.3.5 and 1.3.6).

Finally, there is interesting ongoing work on the inference of agents' motivations from their behavior in inverse reinforcement learning problems (see section 1.3.7).
2. **Task 5.2: Novelty based formation of actions: modelling the “joystick task” (USFD, CNR-ISTC-LOCEN)**

2.1. **Introduction to the task**

In many theoretical accounts of action learning, the learning agent is assumed to already possess a small set of discrete and independent actions. Usually, each action affects the environment in some way, and learning algorithms (e.g., reinforcement learning) determine which action to select. However, the animal is capable of executing infinite movements of varying structure, most of which cause no appreciable change in the environment. If the animal is naive, i.e., if it does not know which movements affect the current environment, it must first discover those movements that do affect the environment; then, those movements can be recruited as "actions" to be considered for accomplishing some task. Novelty detection can aid this discovery process. As considered here, "novelty" is an unexpected sensory event, i.e., a sensory prediction error. As detailed in Redgrave et al. (2008), an unexpected sensory event can cause a phasic increase in dopamine (DA) neuron activity that biases the animal—via DA-dependent synaptic plasticity at corticostriatal synapses in the basal ganglia (BG)—to repeat the behaviour that preceded the unexpected sensory event. Such a repetition bias allows for the reliable presentation of sensory and motor information to associative networks in the brain which can construct internal representations of movements (i.e., actions) and their predicted outcomes.

2.2. **Overall objectives of the task**

The "joystick task" (experimental work described in WP3, Tasks 3.4–3.6) investigates this novelty-based formation of actions in human participants. Briefly, most joystick movements have no effect on the environment. However, some (e.g., a movement to a specific "hot spot" or goal location) trigger a visual signal. How does the participant alter his behaviour so that he discovers the specific movements that trigger the signal, and does discovering those movements aid in accomplishing future tasks that may require those movements? The current work task (Task 5.2) develops modelling work meant to enable us to study how brain mechanisms mediate such action formation within a simulation of the joystick task. We hope to use biologically-plausible learning and control mechanisms to produce model behaviour similar to observed human behaviour, and to use the modelling framework to predict how human behaviour would change under different circumstances (e.g., delays in signal presentation, multiple hot spots, etc).

2.3. **Research work to date**

2.3.1. *Modeling the joystick task with biologically-plausible neural networks.*

**Background and open issues related to IM-CLeVeR**

The modeling framework we use is based on that used by Kevin Gurney and colleagues (e.g., Gurney et al. 2001) to study how neurons in the basal ganglia (BG), cortex, and thalamus interact to perform action selection. The models use neural networks of rate-coded leaky-integrator neurons. Using such models restricts functional mechanisms to those thought to be implemented by biological systems and allows us to examine how deficits in biology affect behaviour. This modeling work will contribute to forming a "bridge" between more abstract theories developed within IM-CLeVeR and the experimental work and biological systems that inspire those theories.

**Specific goals**

The specific goals of this work fall in line with the overall objectives of the task, described above.

**Approach and methods**

A highly-simplified representation of the architecture of the current model is shown in Figure 5.2.1. The boxes labeled "Cortex" and "Thal" (thalamus) represent 15x15 grids of neurons, while the box labeled "BG" represents five grids (each 15x15) of neurons and intra-BG architecture of the same
form as most other BG models from the lab. (The BG nuclei are: D1 striatal neurons, D2 striatal neurons, STN, GPe, and GPi/SNr.) In the current model, the activity of each neuron in "Cortex" represents a command to move to a location in the two-dimensional workspace corresponding to the neuron's location in the grid. Cortical activity is read by a simple plant that moves the system (in this case, a point) from its current location to the location represented by cortical activity: the weighted average of all neurons that are above a threshold.

Figure 5.2.1: Simplified illustration of model architecture. Boxes labelled "Cortex" and "Thal" represent 15x15 grids of neurons, box labelled "BG" represents five 15x15 grids of neurons (see text for details). "Context" and "Exploration" represent hand-engineered sources of excitation.

Activity in cortex is determined by 1) an exploration mechanism and 2) positive feedback loops between cortex and thalamus (one-to-one connections). For each movement, the exploration mechanism chooses a neuron, the activity of which corresponds to moving to a location within the workspace, at random. Excitation from the exploration mechanism to cortex evolves over time from exciting all cortical neurons weakly to exciting just the (randomly) chosen neuron strongly. The exploration mechanism is inspired by studies investigating neural activity in parietal areas during perceptual decision-making tasks: neurons representing competing options increase in activity according to the confidence in that option (Gold and Shadlen 2007). Positive feedback from thalamus boosts cortical activity. When movement ends (the system reaches the location represented by cortical activity within a small threshold), inputs to cortex are set to zero and the neural activity settles to resting values, after which another location is chosen and excitation to cortex commences again.

Activity in cortex is also indirectly influenced by activity in the BG: the GPi/SNr of the BG sends tonic inhibitory projections to thalamus (one-to-one connections), which are all of equal strength. As detailed in Gurney et al. 2001, the intra-BG architecture is such that an increase in activity of D1 and D2 striatal neurons in the BG result in a decrease in activity of the corresponding GPi/SNr neurons and an increase in activities of the other GPi/SNr neurons. D1 and D2 striatal neurons in the BG receive one-to-one excitatory projections from cortex (which are plastic and are all initially 0.5) and also one-to-all projections from a "Context" unit (which are also plastic and initially zero). (In the current model, there is only one context.)

 Movements to most locations have no effect on the environment (i.e., they trigger no signal). However, as with the joystick task, there is a predetermined goal area: when movement terminates in the goal, a reinforcement signal, representing phasic dopamine (DA) activity, occurs. When this occurs, weights from Context to striatal neurons and Cortex to striatal neurons increase at a rate in proportion to their activities. When movement terminates at another location, there is no reinforcement signal and the weights decrease.

Results

Naive model behavior is dictated entirely by the exploration mechanism: the system moves from one randomly-chosen location to the next. Plasticity dictated by the basic learning rules described above result in the following behavior: weights from Context to the striatal neurons that correspond to the goal location increase so that their activities are of a greater activity than others at the beginning of a movement, resulting in a lower activity of corresponding SNr neurons and hence less inhibition to corresponding thalamus neurons. Thus, the positive feedback gain between those thalamus neurons and cortical neurons is greater than that of other neurons, and weak excitation from the exploration
mechanism is enough to boost the activities of cortical neurons corresponding to the goal location. Because excitation from the exploration mechanism evolves from weak and uniform to strong and focused, it can boost the activity of cortical neurons corresponding to the goal location even if the exploration mechanism has chosen some other neuron as the focus of excitation. Because of excitatory connections from cortex to the BG (which are also plastic), already excited cortical neurons cause a further increase in corresponding striatal neurons, a further decrease in corresponding SNr neurons, and a further increase in other SNr neurons. Hence, even strong excitation from the exploration mechanism to other neurons is not enough to override the activities of the already-selected cortical neurons.

Figure 5.2.2 shows behavior of a model "frozen" at an early stage of learning (i.e., after it has hit the goal only a few times). (X: movement location, G: Goal location, G_exp: location chosen by exploration mechanism.) The model will execute a movement to the goal (pink square) when the exploration mechanism has chosen a location near the goal (red dots). It will execute movements to locations chosen by the exploration mechanism when the exploration mechanism chooses locations far from the goal (blue dots). It will execute movements near but not at the goal when the exploration mechanism chooses locations at a medium distance from the goal (green dots). As training continues, i.e., as the model repeatedly hits the goal, the range of locations chosen by the exploration mechanism that result in a movement to the goal increases, and eventually movements to the goal will occur in response to excitation from the exploration mechanism for any chosen location. In rough qualitative agreement with human behavior, the probability that the model executes movements that hits the goal increases.

Figure 5.2.2: Results from a partially-trained model. Left: dot location represents location suggested by exploration mechanism. Color represents movement location of model. Red: movement went to goal location (pink square). Blue: movement went to location suggested by exploration mechanism. Green: movement went in between. Right: distance of movement (X) from goal location (G) as a function of distance of location suggested by exploration mechanism (G_{exp}) and goal location.

Advancement of work and relation to other tasks

The current model represents an early stage of our efforts towards WP5, Task 5.2. The mechanisms used in the model are more in line with our current understanding of the BG than mechanisms used in many other models: the BG bias movement by disinhibiting selected neurons in other areas, not by generating movements directly. In other words, very simple learning mechanisms allow the BG to modulate how movement-generation areas of cortex respond to excitation from other areas. Thus, there is an explicit separation of two ways to influence behavior: 1) the exploration mechanism and 2) movement bias that is represented in weights from context to striatal neurons. In the current model, the exploration mechanism chooses locations at random. However, this is a very simple and unsophisticated “default” strategy, implemented in the current model so that we can focus on how movements are biased. The separation allows us to easily-implement other strategies, such as starting off with structured exploration (e.g., a spiral pattern, as observed in some participants of the joystick task) or an adaptive strategy.

The current model was developed using a simple reinforcement learning rule. Short-term future work on the model includes the development of learning rules that capture some of the essential features of
novelty detection and intrinsic motivation, which will be conducted in close collaboration with WP5, Task 5.1. Another area of future work is to include multiple goal locations that each delivers a reinforcement signal when reached to allow for situations where several different behaviors affect the environment in different ways. Preliminary modeling results suggest that the development of behavior in human participants (WP3, Task 3.4–6) would follow a different strategy than that if only a single goal location was used. Also, the current model does not incorporate any uncertainty or any way to deal with uncertainty, but such considerations will allow us to make stronger connections with behavior and current learning theories. The developments described in this and the preceding paragraphs will be conducted in close collaboration with WP3, Tasks 3.4–3.6. The basic infrastructure of the current model will suffice for these developments, which will be the focus of our efforts over the next few months.

Longer-term future work includes using the basic modeling framework to study how other types of behaviors are learned. The current models, and the current joystick task, investigate behaviors that amount to moving to a particular location in space. However, other types of behaviors include learning to make a short sequence of movements, or gesture that can be executed in any spatial location. Development of this work may be informed by another line of future research: using behavior generated by the model—which uses simple learning mechanisms—to train higher-level controllers that are specialized to execute behaviors of certain types, e.g., moving to a goal location in space or performing a particular gesture. Training of the higher-level controllers will help determine the reinforcement signal that biases movements (e.g., in a well-trained higher level controller, the outcome of the movement is expected, not novel, and hence no bias based on novelty will occur). The training of higher-level controllers can also instruct the exploration mechanism to "try out" behavior of a particular form. Ultimately, the trained higher-level controllers represent actions that are recruited in more traditional action selection tasks. Again, this work will be conducted in close collaboration with WP3, Tasks 3.4–3.6. In addition, the development of higher-level controllers within this work task will be informed by work done in WP6.

In parallel to model development, we have begun communications with AU so that the neural network models we develop—which implement biologically-based learning and control processes—can be used to control the AU iCub. We decided to begin such collaborations at an early stage of model development so that any problems with this process can be worked out with a very simple model. (Very) preliminary results are promising in that the hand of the AU iCub can be controlled by the current model. This work is related to WP5, Task 5.4 and WP6, Task 6.2 and, after further development, can help inform other work tasks involved with the iCub.

2.3.2 Development of learning rules that result in 'optimal' behaviour without a cost function.

Background and open issues related to IM-CLeVeR

The detection of an unexpected or novel stimulus is thought to produce a reinforcement signal communicated by phasic DA neuron activity. Would such a signal be treated like a "reward prediction error" as used with temporal difference models of reinforcement learning (Sutton, 1988)? If so, an explicit cost function, including a concept of a task, must be defined in order to determine the value of each state or state-action pair. In other words, there is some notion of optimality: each action incurs a cost, and the task is to reach a goal state while minimizing cost. The estimated value of each state or state-action pair must be communicated to DA neurons to generate a prediction error. While mechanisms that mediate learning in well-defined tasks likely incorporate notions of optimality, would learning based on novelty detection use simpler mechanisms that do not?

Specific goals

To develop learning algorithms which do not rely on notions of a well-defined task and optimality. To characterize circumstances under which resulting behaviour is similar to or differs from that which results from learning algorithms that do rely on notions of task and optimality.
Approach and methods

To make direct comparisons with traditional reinforcement learning (RL) algorithms, we implement learning rules for an agent in a "grid world," a simple discrete state environment resembling a checkerboard. The agent can move to a neighboring state in any of eight directions (four cardinal and four diagonal directions). At each trial, the agent must navigate from a set starting state to a goal state. In our learning rule, a reinforcement signal occurs only when the agent has reached a goal state (a salient event). The tendency to select each action from each state that was visited en route to the goal state is increased towards the same maximum value, but at a rate that decreases with temporal distance from the reinforcement signal (via an eligibility trace):

$$\Delta Q(s, a) = \alpha \lambda^\Delta t [R - Q(s, a)]$$

where $Q(s, a)$ is the tendency to select action $a$ from state $s$, $R$ (= 20) is the reinforcement signal received for reaching the goal state, $\alpha$ is a step-size parameter set to 0.1, $\lambda$ determines the eligibility trace and is set to 0.7, and $\Delta t$ is the number of steps $(s, a)$ was visited before the goal state was reached. Actions are selected at each state from a softmax distribution across $Q(s, a)$ (with a temperature of 1.5). Importantly, there is no cost associated with any action or temporal discount; hence, there is no notion of optimality. This learning rule is essentially a simple Monte Carlo update rule with an eligibility trace. Therefore, we tentatively refer to it as "MC($\lambda$)." Figure 5.2.3 (top) illustrates the rule (each circle represents a visited state-action pair, the star represents the goal state, and the thickness of the line and size of the arrow represent the rate at which $Q(s, a)$ is increased). Behaviour resulting from this rule is compared with behaviour resulting from a simple TD(0) rule (illustrated in Figure 5.2.3, bottom) in which each action incurs a cost (−1 for cardinal actions, −1.41 for diagonal actions), there is no temporal discounting, and actions are selected via the softmax method.

![Figure 5.2.3: Schematic illustration of update rules. Each circle represents a visited state-action pair, star represents achievement of goal state.](image)

Results

Behaviour developed under MC($\lambda$) is very similar to that developed under TD(0) during early to mid stages of learning. Figure 5.2.4 plots, for each learning rule, the average (over 20 simulations) number of actions taken en route to the goal state (from a fixed starting state) as a function of trial number (note the log scale). The thick red (MC($\lambda$)) and blue (TD(0)) lines illustrate that for "greedy" trials only (where action corresponding to the maximum $Q(s, a)$ was chosen), while the thin light red and light blue lines illustrate that for all trials. For each learning rule, the number of actions taken to reach the goal state decreases with experience to a minimum at similar rates. In other words, even though no explicit cost function is used with MC($\lambda$), resulting behaviour approaches "optimal" as defined by typical cost functions. Such behaviour arises from the statistics of the system: behaviour that corresponds to the shortest path is reinforced at a greater rate than other behaviours. If, as suggested in Redgrave et al. (2008) and other formulations of novelty detection, reinforcement decreases as the stimulus becomes predictable, behaviour may stay relatively static. However, if reinforcement continues (as may happen with mental disorders, e.g., Redish et al. 2008), behaviour deviates from "optimal" with extended training (> 10,000 trials in Figure 5.2.4) under the MC($\lambda$) rule. Because there is no notion of optimality, $Q(s, a)$ for each visited $(s, a)$ continues to increase to the maximum level (set to $R$) with experience.
Advancement of work and relation to other tasks.

This work arose from conceptual discussions focused on how learning may occur in the joystick task and through novelty detection in general: How can learning be driven by some notion of optimality when there is no "task" to be accomplished? How would behaviour develop if reinforcement were to occur without any notion of optimality? Because of the general nature of these questions, and obvious connections with RL, we are conducting research within an abstract framework that allows us to make direct comparisons with RL algorithms. We do not suggest that learning rules such as MC($\lambda$) can dictate all behaviour; rather, they may be used in relatively simpler settings such as with novelty detection. This research is still in its early stages; as it is further developed, we will make stronger connections with work done in WP5, Task 5.1. In addition, the preliminary results discussed here suggest that there is a functional reason why reinforcement habituates as a stimulus becomes predictable: otherwise, spurious actions are executed. The preliminary results discussed here will be presented in poster format at the upcoming Computational Neuroscience Meeting (July 2011, Stockholm, Sweden).

2.4 Conclusion to task

The modeling efforts under the current task focus on how novelty detection drives the development of behaviour in biological systems in environments similar to that of the joystick task (WP3, Tasks 3.4–3.6), and how such behaviour can train higher-level controllers that represent "actions" as used in many formulations of action learning. These efforts led to questions that we are addressing in the second project described here. Research work to date has constructed some of the basic infrastructures that will be used in these efforts, but has used only very general reinforcement signals. Short-term future work in both projects described in this work task will incorporate ideas of novelty detection (in collaboration with WP5, Task 5.1).
3. Task 5.3: Novelty detection based on habituable neural networks (UU, IDSIA, CNR-Prof Barto)

3.1. Introduction to the task

The aim of cumulative learning is to provide a system with developmental programs that allow it to evolve and learn through prolonged periods of observation and interaction with its environment. In order to efficiently achieve this, a mechanism that identifies observations that are new to the robot is needed. In previous work (Gatsoulis et al., 2010) we have identified particular characteristics that are important for the effective operation of a cumulative learning system. In particular, a cumulative learning system should be able to detect novel perceptions, learn online and unsupervised, expand when required, cope with noise, and fuse information from different sensors.

3.2. Overall objectives of the task

The goals of this task are to design and implement learning-driven dynamic focussing algorithms and architectures that lead a robot to focus attention on parts of the environment that appear particularly interesting, depending on their degree of novelty. This capability to learn over time will be exploited in integrated systems (e.g. Tasks 6.5, 7.4, 7.5) to focus resources and learning efforts of the robots on those parts of the environment which may maximize robot’s autonomous learning.

![Figure 5.3.1: Block Diagram showing the overall UU Approach](image)

In its primitive form the problem of novelty detection is to identify new, novel patterns that have never been seen before (Markou and Singh, 2003a; Marsland, 2003; Saunders and Gero, 2000). It consists an important ability of a number of biological cognitive organisms as it reduces computational load by selecting and guiding attention to areas of “interest”, and it has seen an increasing interest in the last decade considering the number of works and surveys that have been recently published (Chandola et al., 2009; Hodge and Austin, 2004; Markou and Singh, 2003a; 2003b; Marsland, 2003).

A more formal description of the problem of novelty detection is as follows. An agent is trained on a set of perceptual patterns \( X = x_1, x_2, ..., x_n \) using a training method \( F \) and forming a knowledge database \( K = F(X) \). At time \( t \) an observation \( o \) is considered novel if it differs significantly from what is already known, i.e. from \( K \), using a novelty detection filter \( N \) to identify the level of novelty and the particular parts that are novel. The observation \( o \) is then inserted in the training set \( X \) as a new training pattern \( x_k \), updating the agent’s knowledge \( K \).

Although a number of novelty detection methods have been proposed in the literature, mainly focusing on detecting anomalies and outliers, i.e. identifying patterns that do not conform to expected behaviour (Chandola et al., 2009; Hodge and Austin, 2004; Markou and Singh, 2003a, b; Marsland, 2003), these are unsuitable for the task of cumulative learning as they require beforehand a normal set to be trained on. As such, an effective novelty detector requires learning architectures that support dynamic and incremental expansion of knowledge representation. Most importantly, an effective novelty detector should be able to consider as novel new observations as well as already learnt ones that have the potential for further exploration.
To address these requirements, the approach to novelty detection we investigate at the University of Ulster is based on habituation, which is explained in detail in the next section.

**Habituation as novelty detection**

Habituation is a non-associative form of learning, also called single event learning, that is defined as a decrease in responding following repeated stimulation without this decrement being caused by fatigue or receptor adaptation, and is often considered the simplest and most basic form of learning (Thompson and Spencer, 1966; Thompson 2009; Rankin et al., 2009). It is such a fundamental motivation for learning that it has been found in every organism studied (Thompson and Spencer, 1966; O’Keefe and Nadel, 1978), and it has also been observed that the behavioural rules for habituation are common in all organisms (Rankin, 2009).

From a computational perspective the model of habituation proposed by Stanley (1976) is of particular interest as it has shaped much of the research of habituation in the domain of robotics (Marsland et al., 2005; Marsland, 2009). This model describes the decrease in the synaptic efficacy $h$ by the following first order differential equation, which result for different parameters is shown in Figure 5.3.2:

$$ \tau \frac{dh(t)}{dt} = \alpha[h_0 - h(t)] - S(t) \quad (5.3.1) $$

where,

$h_0$: is the initial value of the habituation level.

$\tau$, $\alpha$: are constants controlling the habituation and recovery rates.

$S(t)$: is the activity of the unit.

![Figure 5.3.2: Behaviour of habituation for different rates](image)

One motivation that we believe is crucial in distinguishing a complex organism from a simple one is the pursuit of knowledge. From the moment they are born animals and humans alike try to acquire as much knowledge about themselves (i.e. motor babbling) as about the world. This is driven by both intrinsic and extrinsic motivations.

Novelty attracts the attention of an agent that tries to investigate the novel stimulus to learn something about the world. This is an extrinsic stimulus. When no new stimulus is present, the agent might decide to look for something new. This search for novel stimuli is not triggered by any external signal and therefore can be categorised as intrinsic.

Intrinsic and extrinsic motivations drive each other. We can imagine an agent sitting in front of an empty table. The agent might stare at the table for some time, being amazed by the texture and the material. Eventually it will become habituated with the stimuli it is receiving and it will look for
something novel. This searching for something novel is the action triggered by the internal stimulus of habituation. If a new object comes into sight of the agent, then this catches its attention. This is now an external stimulus that has been triggered by a previous internal one.

Therefore the work of this task is to explore novelty detection based on habituation, as a method of emulating in robotic systems, intrinsic motivations as observed in biological entities.

3.3. Research work to date

3.3.1. A novel integration of habituable neural networks with bag-of-words models

Background and open issues related to IM-CLeVeR

The effective operation of an IM-CLeVeR novelty detector relies on a reliable visual recognition system of objects that the robot has learnt. One of the current state of the art methods in visual perceptual learning and recognition is the bag-of-words approach, which is explained further below. The novel integration of habituable neural networks within the bag-of-words models, would enhance the performance of the learning system.

UU has firstly suggested this approach in the workshop in novelty detection hosted at UU in November 2010. The results of this research have been published in Gatsoulis, et al. (2011).

Specific goals

- To extend and integrate the habituable neural-network of Marsland, et al. (2000) with the bag-of-words model for effective perceptual learning and object recognition, resulting in a novel system.
- To validate the performance with real world visual perceptions.

Approach and methods

The bag-of-words (BoW) model has its roots in natural language processing where it was used to represent and classify documents according to the frequency of particular words existing in a dictionary. The produced histograms are then the representations of the documents. The BoW technique has been adopted by the machine vision research community to describe and classify images in the same manner, by using histograms of the frequencies of “visual words” from a dictionary that exist in the image. The bag-of-words technique consists of the following generic steps, also shown in Figure 5.3.3.

1. Extract a set of feature descriptors, such as SIFT, SURF, etc., from a perceived image.
2. Learn a visual vocabulary, by training an unsupervised structure (e.g. k-means, SOM, etc.) to the extracted features of the perceived image.
3. For a given vocabulary and a set of feature descriptors of an image, compute the histogram of the frequency of visual words that match these feature descriptors.
4. Train a classifier (e.g. support vector machine, SOM, etc.) with the produced histograms.

**Figure 5.3.3: Generic and extended bag-of-words model with habituable neural networks**
For the vocabulary we used a typical Kohonen map while for the classifier we used a Kohonen map with habituable synapses similar to the one used in Marsland et al. (2000). For the experimental data set we used three categories of kitchen objects: 32 forks, 22 mugs and 8 plates. Three quarters were used for training and the remaining one quarter was used for validation.

Results

The experimental results have shown that the classifier with the habituation module was able to classify the validation dataset of objects correctly at a rate of 81.25%. However, we have identified particular limitations to the original methodology of the bag-of-words that prevents it from being efficiently used in a cumulative learning task. These limitations are:

- The vocabulary of the BoW was constructed offline and in batch mode. This is restrictive for real world operation as it assumes that the perceptions are known a priori.
- The size of the vocabulary was fixed. This is a restriction for cumulative learning use of the system.

These issues are addressed in Section 3.3.2.

Advancement of work and relation to other tasks

We have successfully integrated a habituable neural network with current state of the art in visual learning and recognition of objects, and we have identified the limitations of this system that prevent it from being effectively used in the IM-CLeVeR cumulative learning tasks.

We have also tested this system with the visual system of the iCub during our visit at IDSIA for a week in April 2011. This module is related to Tasks 5.4 and 5.5. regarding novelty detection; it can provide input to Task 6.5 dealing with hierarchical representations of perceptual data; and it is also related to the visual pre-processing in WP4.

3.3.2. A novel online expandable bag-of-words used for unsupervised cumulative learning

Background and open issues related to IM-CLeVeR

In previous work (Section 3.3.1) we identified that our novel integration of habituable neural networks with bag-of-words models is limited for cumulative learning tasks, due to the requirements of the bag-of-words. This work provides a novel approach to the bag-of-words model to resolve these issues. This work has been submitted and is under review (Burbridge et al., 2011).

Specific goals

- Research and develop a novel expandable bag-of-words model that is capable of cumulatively learning new objects in an online and unsupervised manner.
- The validation of the proposed novel solution in a real world scenario with a physical robot.

Approach and methods

The expandable bag-of-words is based on an expandable self-organising map, inspired by the work of Fritzke (1994) and Marsland et al. (2002). A node \( n \) in the expandable neural network \( N \) describes a set of feature descriptors belonging to a set of objects, with its weights vector \( w \), and an index of objects \( i \). The network initially consists of one node \( n_0 \) which its weight vector \( w_0 \) randomly initialised, and belonging to no objects, as none has been learnt yet. The system operates in two steps: a) an inspection step, and, b) a training step. During the inspection step the system classifies the perceived object to one of the known objects or as a new object, using a voting mechanism. A training step succeeds the inspection step if the object is unknown, or if it is classified as one of the known objects but it has a high number of unseen features. For each one of the features of the objects, the best matching node of the network is computed, and if its error is high enough then a new node is added to the network, otherwise the best-matching node is further trained. The object is re-inspected
again to decide whether it requires further training, and the two steps of inspection and training are repeated in a loop until the object is learnt sufficiently. A detailed and formal description of the proposed novel system can be found in the submitted paper (Burbridge et al., 2011).

The robot inspects an object from all directions by driving around it and taking images of it, providing a 360° perception of the object. As with many BoW implementations, robust SURF (Bay et al., 2008) features are extracted from the images of the objects to provide scale, rotation and partial illumination invariance. As new images are taken, the data is dealt with immediately and then discarded. The robot learns online and cumulatively new objects in an unsupervised manner. The operational process of the system, shown in Figure 5.3.4, consists of the following phases that are continuously performed in a repetitive loop:

- Inspection loop, during which the robot carries out a single 360° visual inspection of the object. Decision, during which the robot decides based on the inspection loop whether the perceived object is:
  - a well known object which requires no further training;
  - a known object but with still many features to be learnt, and as such requires further training;
  - a new object to be learnt from the beginning.
- Training loop, during which the robot circles around the table perceiving and learning in an online, continuous and cumulative manner the 360° perception of either a previously known object or a new object, depending on the outcome of the preceding decision step. While the robot is training on an object it is said to be in “training mood”. It will remain in training mood until it enters a “do nothing” phase, indicating that the object has been sufficiently learnt.

![Figure 5.3.4: Operation process](image)

Results

The system was tested with a physical robot and 10 real-world objects, shown in Figure 5.3.5. The robot was a MetraLabs Scitos-G5 differential drive mobile robot equipped with a SCHUNK 7 degrees of freedom manipulator and a Microsoft Kinect camera attached to the end-effector. The inspection step was applied on a 360° perception of the inspected object. The training step was also done online and incrementally while observing the 360° view of the object. The robot was programmed to autonomously drive around a table. One object was placed on the table at a time, and the manipulator was programmed to maintain the object in the camera view at all times. The surface of the table was a random sheet of gift wrap paper to make the world as realistic as possible. The experimental set-up with one of the objects on top of the table is shown in Figure 5.3.6.
The experimental procedure consisted of a training phase and a validation phase. In the training phase the robot inspected, decided and learnt the training objects autonomously. In the validation phase the robot is asked to recall the 10 objects it has previously learnt in the training phase, by performing 5 consecutive inspection-decision loops without any further training. Figure 5.3.7 shows the results of training the system on the 10 objects shown in Figure 5.3.5. In total, 64 training and inspection cycles were required to learn the 10 objects, and the size of the network grew to 63000 nodes. The most substantial jump in network size occurred when training on Object B (a toy robot). This was due to the higher number of SURF features found on this object. Figure 5.3.7 shows that the ratio between known and unknown features quickly dropped as the training on each object proceeded. This shows the stability of the SURF features used in the system. However, extreme changes in lighting were found to affect the object learning. This is apparent when training on Object I (epochs 47 – 54), where a spike in the ratio of known features to unknowns can be seen at epoch 52. This occurred when bright sunlight saturated the image. Regardless of such real-world online difficulties, the system was able to learn all 10 objects successfully in a completely unsupervised manner. As each new object was presented to the robot, it was correctly identified as unknown and a new object class was created.

The same 10 objects as had been learnt by the system were each placed in front of the robot 5 times and an inspection loop executed. Figure 5.3.8 shows the classification results of all 5 loops for each object. Each object was correctly identified despite there being a large number of unknown features also discovered, particularly on Objects B, I and J. The unknown features are the result of erroneous
features in the image background (around the object edge), illumination changes causing features to not be matched to previously seen features, and new features on the object that had not been detected while training. Since the system works with online images in real-time, unknown features will always be present. However, the system demonstrated reliability even under these conditions. The overlap in features between objects is most observable between Objects C and H (two books), where although each object was identified correctly, there is evident similarity between them. On the other hand, the percentage of votes for objects other than Object B when presented with Object B were low. This is again due in part to the large number of features present on Object B, reflecting its distinctiveness from the other objects.

Figure 5.3.8: Graph showing the mean histograms of 5 inspection loops for each one of the 10 learnt objects, during the validation experiment

The results from these experiments have demonstrated the dynamic learning capabilities and high accuracy of our proposed novel system. A first limitation is that as the current system is setup, it is capable of perceiving and learning only one object at a time. It would be an improvement if it was able to cope with multiple perceived objects and decide between them which one to focus its attention. A novelty detector would be needed in this case to provide the intrinsic motivation for selecting which object to focus on and to prevent the system from training.

These issues are addressed in the work described in Section 3.3.3.

Advancement of work and relation to other tasks

We have successfully implemented a novel expandable bag-of-words capable of cumulative learning objects in an online and unsupervised manner, and then being able to recall them. The system has demonstrated excellent performance, being able to correctly identify and recall all objects that were learnt in a real-world scenario with a physical robot.

This work has relations to the novelty detection Tasks 5.4 and 5.5, with the hierarchical structures in Task 6.5, and with the visual pre-processing tasks of Task 4.2.

3.3.3 A novel intrinsically motivated perceptual learning system based on expandable bag-of-words and habituation

Background and open issues related to IM-CLeVeR

A key component in the IM-CLeVeR project is a novelty detector module that directs the learning of the robot. This can be decomposed in an intrinsically motivated perceptual learning system driven by novelty detection, and in an intrinsically motivated action learning system driven again by novelty detection. In this work we are focussing on the perceptual side, and we have researched and
developed a novel system module capable of that. We are preparing these results for submission to ICRA’12. This work has significant contribution and is a major milestone to the IM-CLeVeR project.

Specific goals

- To research and develop a new system that allows a physical robot to learn the description of objects, driven by the intrinsic motivation of novelty detection based on habituation.
- To validate the model with a physical robot with real-world objects.
- To integrate this module with the contributions of AU and IDSIA.

Approach and methods

The work presented throughout this section is integrated together here into one complete novel solution. The learning structure presented in Section 3.3.2 is used as the classifier system. The issues of having a modular novelty detector assessing the novelty of the objects themselves, have been addressed here. The system consists of a separate stand-alone novelty detector that measures the novelty of the perceived objects according using Stanley's (1976) habituation model presented in Section 3.1. It maintains along the classifier a database of objects with their novelty parameters as determined by Equation 5.3.1 (in Section 3.1), such as previous novelty value, parameters controlling the levels of habituation and dishabituation, the thresholds of engaging and disengaging attention, etc.

The complete system operates as shown in Figure 5.3.1 (Section 3.1). At every step it takes as input a list of regions of interest of the objects lying on a table from Task 4.2, which are classified accordingly. Then the novelty detector assesses their novelty and if one of them is found to be novel enough the robot focuses its attention on it. While the robot focuses its attention to an object this object is further learnt until it habituates below a level signifying that the robot has no more interest on this object. All objects in the memory of the system slowly dishabituate.

Results

We have initially validated the system with a physical Willow Garage PR2 using real-world objects. Figure 5.3.9 demonstrates the capability of the system in learning to detect the objects that the novelty detector module suggests. It can be seen in Figure 5.3.9(a) that one of the objects is already known and has a low novelty value, while two objects are new and have maximum novelty values. The system chooses one of them to train on. As it learns the description of the project, it also habituates on it, as shown in Figure 5.3.9(b). After training on this object is completed (Figure 5.3.9(c)) the system scans again the area for objects, recognises two of them (notice the dishabituated value on object 0), and the third one is still unknown. After all objects have been learnt (Figure 5.3.9(d)) the system focuses on the object with the highest novelty that has exceeded its “interestingness threshold”.

Advancement of work and relation to other tasks

We have successfully implemented a novel system that gives the capability to a novelty detector which is based on Stanley's model of habituation presented earlier in order to provide the system with the intrinsic motivation of selecting the most novel object.

We have initially validated this novel component with a physical robot and a number of real-world objects. We are currently testing the system thoroughly, and will present the results in the review meeting. We are also preparing a paper describing these results for ICRA'12.

This work is related to the novelty detection Tasks 5.4 and 5.5, takes inputs from Task 4.2, and is related to Task 6.5. During a visit of two UU researchers to AU, we have successfully integrated this module in the iCub and with the modules developed by AU in joint work, the results of which are presented in Section 6.3. We are in the process of integrating this module into IDSIA’s system as well.
(a) Unknown objects with maximum novelty are identified

(b) As a previously unknown object is being trained its novelty/habituation value decreases

(c) Previously known object has dishabituated, however, the most novel object is found the new unknown

(d) After a while one of the learnt objects has dishabituated and is found novel again

Figure 5.3.9: Cumulative perceptual learning driven by the intrinsic motivation of habituation

3.3.4 Affordances learning

Background and open issues related to IM-CLeVeR

Manipulating objects is a robotic skill that shares many of the challenges and the issues of grasping objects. The robot needs to process the noisy sensory information it receives to extract the 3D position of any object to interact with, together with detailed information about the surrounding environment to execute safe operations.

Specific goals

- To develop a method which implements the actions agreed at the November 2010 workshop on Visual Perception and Novelty Detection held at UU
- To observe the outcome of the actions so that the resultant behaviours can be learnt
Approach and methods

For effective skill learning the link between actions and outcomes needs to be determined. Actions exploited were push, pull, topple and rotate, as agreed at the Visual Perception and Novelty Detection Workshop at UU in November 2010. We conducted a number of experiments in order to test the observation of the initial and final pose of the objects after the performed action using a Willow Garage PR2 robot. The robot is in front of a table with objects on the table at reachable distances. RANSAC is used to segment a plane (the table) out of the stereo reconstructed point cloud. The robot chooses a random object, the robot head focuses on it and the narrow stereo cameras create a new, denser, 3D reconstruction of the object. PCA is applied to this newly created point-cloud to find the major and minor axis of the object (Holz et al., 2011). After an action the robot obtains an estimate of the displacement effect of the performed action by comparing the current position of the object with the previous one. In order to allow the robot to use this information to build up knowledge of action-consequence links, the recognition of actions based on semantic graphs was investigated as described in (Erdal et al., 2010). To quickly test the applicability of the algorithm we used a pre-recorded video stream of a human arm manipulating tabletop objects. In the real experiments a robot arm will be used to grasp and manipulate an object as described in (Ratnasingam & McGinnity, 2011a; 2011b). In this work so far, the action can be recognised based on the object which is being moved and the stability of the state as a consequence of an action can also be recognised.

Results

UU has developed methods which enable a physical robot to carry out actions on perceived objects and identify the outcome of these actions so that basic affordances of the objects can be associated with particular events. A typical image from a resulting action image sequence is shown in Figure 5.3.10. In this example an object was toppled by a push action from the robot and the robot was able to correctly observe that the object had toppled as a result of the push action.

![Figure 5.3.10: Pushing action resulting to detected toppling](image)

Advancement of work and relation to other tasks

UU has implemented the agreed actions on our physical robots and has implemented methods to identify a change of state due to a performed action. This work is relevant to Task 4.2, Task 5.5, Task 6.5 and Task 7.5.

3.4 Conclusion to task

In this section UU presented a summary of this year’s research on novelty detection in Task 5.3. Original developments in techniques for visual perception based novelty detection and learning and exploiting habituation are reported. UU has successfully tested their developments on physical robots.
which currently include a SCHUNK 7DOF manipulator mounted on a Scitos base and a PR2 platform. Implementation on an iCub robot is planned to occur during June –July 2011.

UU also described their novel studies which have been published or are under review. This has led to a delivered component to the IM-CLeVeR partners that is capable of learning the descriptions of objects and recognise them in the future, driven by the intrinsic motivation of novelty detection based on the theories of habituation. UU has also implemented the agreed actions on physical robots and investigated how a change of state due to a performed action can be identified.

Our plans for year 3 are to further optimise the perceptual learning/novelty detector module and integrate this module with the research on unsupervised action sequence learning, according to the goals of the IM-CLeVeR project. We will validate our results with physical robots in real-world experiments and collaborate with our partners in extending the research to implementation on the iCub Clever-K demonstrators.
4. Task 5.4: Novelty based development of redundant sensory-motor systems (AU)

4.1. Introduction to the task

When implementing a robot it is necessary to establish a source of motivation, otherwise the agent would be almost entirely passive. We believe it is important, for experimental purposes, to keep the motivational driver(s) of the agent as transparent and as simple as possible. This is because we do not want to tangle the role of motivation with the other processes involved in development and learning. Thus, we avoid as much pre-structuring as possible, including any predetermined goals or goal structures. We have initially assumed a simple model of novelty, essentially defined as any unexpected event. We include all possible events here: sensory, motor, new bodily states, anything not experienced before. Our intrinsic motivator is simply based on the rules "any novelty is stimulating" and "act towards stimulation". We have found that this simple motivation is enough to drive quite complex developmental behaviour.

4.2. Overall objectives of the task

To investigate the impact of novelty on the development of sensorimotor competencies, both for exploration of the sensorimotor space and in the shaping of the developmental trajectory.

4.3. Research work to date

We have been refining our ideas of novelty as a driver for development in our robotic systems. Most of our work has focused on preparing the iCub for future experiments, by implementing the sensory-motor development processes designed earlier. This means our recent work has concentrated on solving various theoretical issues which will facilitate future learning research. With the iCub system now having developed basic sensorimotor control of eyes, head and inchoate arm action, we are beginning to implement and experiment with different novelty detectors.

4.3.1. Novelty as a driver for staged development

Background and open issues related to IM-CLeVeR

Intrinsic motivation is a central feature of this project. The approach adopted assumes intrinsic motivation will be manifest as a curiosity drive, and, hence, novelty will be a key component. This is in accord with early infant development where novelty is a prime motivator and can be seen in the many hours spent in play activity. We aim to employ more sophisticated models of novelty, based on those being developed elsewhere in Work Package 5, at a later stage in the project when results become available.

Specific goals

In this task we are investigating the use of novelty-based systems to drive learning of sensorimotor coordination through motor babbling and play. By implementing different novelty detectors on our systems we will investigate how different motivators cause the robot to develop along different developmental trajectories.

Approach and methods

The link to action is important: when a new object or event is experienced the agent should try to apply actions to it that have been relevant in previous experience. Thus, new items become further explored. When experience is poor or even lacking entirely then the action may take the form of motor babbling. Rather than view motor babbling as a purely random activity we have investigated it further. Motor babbling is a form of spontaneous action, (that may actually have some structure), and is very much a part of developmental learning. As the infant grows, so motor babbling behaviour
changes into the more obviously structured behaviour known as "play". Play is very important in infant development (Bruner et al., 1979) and continues to use novelty as a driver - again even one-shot novelty can trigger extensive periods of play. We have developed our ideas into a theory of play as "intrinsic action" in which behaviour is not a goal following but a goal finding process. This hypothesis is described further in (Lee, 2011). Play is not only important in the understanding of objects and tools, but also social interaction (Bekoff and Allen, 1998), and symbols and language (Leslie, 1987).

Using these ideas we are continuing to implement our developmental approach on our iCub humanoid robot. Rather than use metrics to release constraints in a structured manner (by monitoring development and then triggering new scope for learning) we have explored the possibility of behavioural stages emerging internally when sufficient structure has been created to support another stage of behaviour (Huelse and Lee, 2010). The goal must be to achieve qualitative advances in behaviour without structural change. We are making some progress on this challenge by exploring how constraints can be lifted by internal emergence (Huelse and Lee, 2011).

With our simple novelty based motivator we use excitation and habituation methods to direct attention to objects and events. This has been implemented on the iCub and has allowed sensory-motor mappings to be learned for eye movement control, head control, and eye-head interaction and compensation. This has resulted in a robust gaze space that can represent the direction of gaze relative to the robot’s ego-sphere. We have also implemented a similar mapping learning system for the robot arms (which have considerably more degrees-of-freedom) and this generates a reach space as another component of the ego-sphere. Motor babbling in the form of hand regard is being used to coordinate the gaze and reach spaces, after which the robot will be able to reach to any gaze point and gaze to any reach point.

Novelty is likely to impact upon the developmental trajectory whether behavioural stages emerge from internal structure or are imposed by a system of constraints. Since novelty drives learning in our systems, a preference for certain types of novelty will focus learning on the respective type of stimuli. This will result in a system which is tuned to particular novelty sources, and will more readily develop related skills. As the system develops, the higher level skills that are acquired will reflect the novelty preferences used to learn the lower level sensorimotor mappings. It follows that, in otherwise identical robot systems, different novelty preferences will motivate the robots to learn, and therefore develop, in different ways. We are aiming to implement novelty detectors from UU in our future work, and to compare the effect that these have on the developmental trajectory to that caused by our own simple novelty detectors. This is explained in more detail in section 6.

**Results**

We have explored the ideas of motor babbling and play in infancy, and the importance of novelty as a driving mechanism (Lee, 2011). We have also begun to implement systems from UU and USFD to explore directed search of sensorimotor space. For these investigations we require our developmental systems to be in place on the iCub, which has been the focus of this year’s work. Now these systems are in place, we are set to begin conducting experiments on novelty in year 3.

**Advancement of work and relation to other tasks**

This work will be used to drive learning of the maps and mappings described in Task 4.4, and the constraints lifting process of Task 6.2. We are also currently integrating with systems from UU and USFD, with the aim of carrying out further experimentation in year 3. These works will be utilised in the CLEVER-B demonstrator of Task 7.4.

**4.4 Conclusion to task**

Novelty is a key motivator in learning and development. We have outlined our ideas on how novelty impact upon motor babbling and play, and how developmental stages may emerge through novelty-driven exploration.
5. Task 5.5: Information-theory indexes as sources of novelty signals (IDSIA-SUPSI, CNR-ISTC-LOCEN)

5.1. Introduction to the task

Most of IDSIA-SUPSI’s work on novelty and curiosity for WP 5 focuses on compression progress (also called “learning progress”) as the primary method for generating a novelty signal. The theory of compression progress, as described most recently by Schmidhuber (2010), attempts to concentrate learning effort not just on what is unknown to the agent, but on what is unknown and most easily learnable. While the theory is formally well specified, only few and limited experiments have been done to investigate its validity and applicability. Also insufficiently investigated is the question of which types of learning/compression algorithms and what measures of compression progress are most appropriate. Most of the work in this section was designed to investigate, test, and develop the theory of compression progress as an intrinsic motivation.

5.2. Overall objectives of the task

The goal of this task is to implement and demonstrate general and principled approaches to curiosity and novelty detection using concepts from algorithmic information theory, such as compression progress, in reinforcement-learning tasks. This year, our primary goal was to investigate, deeply and rigorously, the principle of compression progress as an exploration method and to devise implementations that could be used in combination with general learning systems. As the ideas had not yet been studied broadly, there were (and remain still) many ramifications to be explored.

5.3. Research work to date

IDSIA-SUPSI’s Work on novelty and curiosity for WP 5 is described in the following five sections:

1. Artificial curiosity with planning,
2. Theoretically optimal Bayesian exploration,
3. Using curiosity to direct reinforcement learning with visual images,
4. An investigation of learning progress as a curiosity drive,

Each of these sections is explained in detail below. For each section, there is at least one accompanying article.

5.3.1. Artificial Curiosity with Planning for Autonomous Development

Background and Open Issues related to IM-CLeVeR

Reinforcement-learning (RL) agents with high-dimensional visual observations require an effective perceptual system that can simplify the complex visual inputs to make learning tractable and effective. An intrinsic reward signal such as artificial curiosity (Schmidhuber, 2010) when combined with planning can drive the agent to actively improve its perceptual system by seeking out interesting samples in its environment.

Specific Goals

The main challenge is to integrate the external reward signal with the intrinsic reward signal (Figure 5.5.1). The external reward signal is stationary while the curiosity signal changes rapidly. As a result TD-based, incremental RL approaches fail to adapt the state values quickly enough to respond effectively to the ever-changing reward landscape.
Approach and Methods

We introduce a novel, artificial-curiosity system that uses planning to exploit a quickly changing intrinsic reward function (Luciw, et al., 2011). Our implementation used least-squares policy iteration (LSPI) (Lagoudakis and Parr, 2003) augmented with a learned model of the environment. We used vector quantization to build an internal state layer. The transition tables of this internal layer were then learned as the agent moved through the environment. Each internal state-action pair was associated with a pair of predictors. Both predictors kept a weighted average of the sensory reconstruction error and the state transition error, but one predictor kept a long-term average and the other a short-term. The reducible error—the difference in the two averages—provided a measure of interestingness (Schaul, et al, 2010). That is, states with opportunities for learning were considered most interesting (Figure 5.5.2).

Results

The system was tested in a noisy environment with high-dimensional visual observations: an overhead view of a 2-dimensional maze (Figure 5.5.1). The system was able to choose actions that improved both its internal representation of its observations as well as its value-prediction mechanisms. This enabled the agent to quickly learn how to act so as to maximize external reward. Unlike most other implementations (e.g., Oudeyer, et al., 2007) our system explores the environment successfully without relying on random action selection (Figure 5.5.2).

Advancement of work and relation to other tasks

As the iCub has a very large action set, it faces an extremely difficult RL task. Our developmental system potentially allows the iCub to (1) model new parts of its environment as it finds them (continual learning), and (2) explore the results of its actions in a principled manner (artificial curiosity). A logical next step is to adapt the system to the iCub.

Figure 5.5.1. (left) Task with high-dimensional observation: The agent sees its own position (white square in the upper left) and the entire environment, including goals (upper and lower right), walls (other white areas), and passable areas (black areas) as a noisy image taken from above. The agent can move up, down, left or right. Upon reaching a goal area, the agent teleports to the upper left. The upper right goal is hard to find with random exploration alone. (right) External and intrinsic state values for an agent that has reached both goals. Values based on external reward are stable, but those based on intrinsic rewards change constantly. The agent handles the changing reward landscape using a planning technique (LSPI).
5.3.2. Optimal Bayesian Exploration in Dynamic Environments

Background and Open Issues related to IM-CleVeR

The primary purpose of intrinsic motivation is to promote efficient exploration of complex environments. However, so far little theoretical work has been done that directly addresses curiosity as a mechanism for exploration. We investigate information gain as a curiosity mechanism and explicitly consider its usefulness for driving exploration (Sun, et al., 2011).

Specific Goals

Our goal was to provide clear theoretical results on the use of curiosity values (information gain) as a mechanism for optimal exploration in general dynamic environments. We carried out a careful theoretical investigation of the principle of optimal exploration in these environments with no external rewards. In particular, we studied how an agent should plan its moves to efficiently improve its knowledge of the environment.

Approach and Methods

We studied the question of optimal exploration within a classical framework, where the agent improves its model of its environment through probabilistic inference, and learning progress is measured in terms of Shannon information gain. The key idea is the introduction of the curiosity Q-value, defined as the cumulative expected information gain over future time steps:

\[ q^\pi_t (h, a) = \mathbb{E}_{o|h} \mathbb{E}_{a_1, o_1, \ldots, a_{T-1}, o_{T-1}} | hao \| haoa_1o_1 \ldots a_{T-1}o_{T-1} | h, (5.5.1) \]

where \( a \) is an action; \( o \) is an observation; \( h \) is a history of alternating actions and observations; \( q(h, a) \) is the Q-value for history \( h \), action \( a \); \( \pi \) is the policy used for exploration; \( T \) is the planning horizon; and \( g(h'|h) \) is the KL divergence:

\[ g (h'|h) = KL (p (\theta|h') \| p (\theta|h)) = \int p (\theta|h') \log \frac{p (\theta|h')} {p (\theta|h)} d\theta, \quad (5.5.2) \]

with \( \theta \) summarizing the knowledge about the environment.

Results

We show that the curiosity Q-values satisfy the following recursion:
\[ q^T_{\pi} (h, a) = \mathbb{E}_{o|h,a} g(hao|h) + \mathbb{E}_{o|h,a} \mathbb{E}_{a'|hao} q^{T-1}_{\pi} (hao, a') . \] (5.5.3)

This equation has an interesting interpretation: since the agent is operating in a dynamic environment, it has to take into account not only the immediate, one-step expected information gain of performing the current action, i.e., \( \mathbb{E}_{o|h,a} g(hao|h) \), but also the expected curiosity value of the agent’s resulting state, i.e., \( \mathbb{E}_{o|h,a} \mathbb{E}_{a'|hao} q^{T-1}_{\pi} (hao, a') \). As a consequence, the agent must choose actions balancing both factors to improve its total expected future information gain.

We showed that the agent can, at least in principle, optimally choose actions based on previous experiences, such that the cumulative expected information gain is maximized, and we proved the existence of an optimal exploration policy. We also proposed a dynamic-programming-based approximation to the optimal exploration policy, assuming that the environment is finite and Markovian. The proposed algorithm outperforms a number of existing algorithms.

Relevance of the results for the problem and for the project
The results strike to the heart of the theory on intrinsic motivation. They show how a curiosity signal that drives action selection towards maximizing future information gain forms a theoretically critical component of optimal exploration in dynamic environments. Our algorithm that approximates this behavior could potentially be adapted for use by the iCub in exploring its sensors, effectors, and environment.

5.3.3 Using Curiosity to Direct Reinforcement Learning with Visual Images

Background and Open Issues related to IM-CleVeR
The iCub relies on human-like senses, and we therefore expect vision to be crucial for its development. But visual data presents difficulties for reinforcement-learning algorithms due to its high dimensionality. We studied the use of curiosity signals in RL domains with visual input to increase efficiency of learning by guiding exploration (Cuccu, et al., 2011).

Specific Goals
Our goal was to investigate the use of novelty signals to improve learning efficiency in RL tasks with visual input. We used the signals first to choose images for training, and second to bias an evolutionary algorithm (EA) towards those individuals most successful at discovering unusual images, forcing exploration of places where novelty could be discovered.

Approach and Methods
In the spirit of Neto and Nehmzow (2007) and Marsland et al. (2002), we used vector quantization (VQ) as a data compressor. It generated two outputs: a tabular encoding of the image and a novelty value derived from the image reconstruction error. Training was done by policy search using an evolutionary algorithm (EA), which evaluated multiple individual controllers over a series of generations. For our test case, we used the well-known Mountain car task (Moore, 1991). Input to the agents was a high-dimensional visual image of the car on the mountain (Figure 5.5.3). Training of the VQ compressor occurred after each generation, using images generated during evaluation of the controllers. As the controllers improved, they explored new territory, generating new images for training. A novelty value was calculated from the reconstruction error (or "distortion") of each image with respect to the VQ’s closest stored centroid. The VQ was trained on the images with greatest novelty. This novelty value was also used during the fitness search of the EA, which attributed greater fitness to the individuals that generated images with greater novelty.
Results

By introducing novelty, vast plateaus of the previously flat fitness landscape thereafter exhibited contours, providing a gradient that lead towards the least-explored parts of the plateau. Unlike pure novelty-based systems (e.g., Novelty Search, Lehmen and Stanley, 2008), the search is not directed towards the most unexplored areas, but towards the regions of space the system already knows about, but that are most dissimilar to previous experience, i.e., where learning occurs fastest.

Relevance of the results for the problem and for the project

The next step, already underway at IDSIA-SUPSI, is integration with the iCub simulator. We will perform a similar experiment on a basic RL task using the iCub’s cameras. We expect the benefit of the novelty signal in shaping the fitness landscape and improving learning efficiency will increase drastically with the complexity and dimensionality of the problem. This work is also relevant for IDSIA-SUPSI’s joint, ongoing work on novelty detection together with UU.

Figure 5.5.3. (left) The agent’s observation was a 15x30 image from the perspective shown. (right) Averages over 20 runs, showing steps to goal over successive generations. (a) Basic test, with normal gravity and a lenient time limit. The method driven by intrinsic motivation and the one driven only by fitness are about equally good. (b) Test with increased gravity and a short time limit, close to the limit of feasibility. A higher percentage of failures cause a larger plateau in the fitness landscape: without good gradient information, the intrinsic reward from the compressor is shown to be a major advantage.

5.3.4 Investigating Curiosity Drives for Developmental Learning

Background and Open Issues related to IM-CleVeR

The principle of compression progress has been described formally (Schmidhuber, 2010), but little rigorous work has been done to test it empirically. We therefore attempted a solid empirical investigation of this principle as a mechanism for exploration (Ngo, et al., 2011). Understanding the principle better helps us apply it to exploration on the iCub.

Specific Goals

Our goal was to perform a rigorous empirical validation of learning progress (a.k.a., compression progress) as an exploration mechanism, especially in the setting of online interactive learning.

Approach and Methods

We devised a simple task where the regularities embedded within an environment are represented by a set of functions, and the agent chooses which one to learn at each time step. The environment is embedded with functions \( f_i (i = 1..N) \), initially unknown to the agent. Each function maps an observed feature vector \( x = (x^1, ..., x^m) \) to a discrete outcome \( y = f_i(x) \). The agent’s developmental stages are reflected through its learning of the environmental regularities represented by these functions. The agent uses Multi-Layer Perceptrons (MLPs) as predictors, and simple back-propagation as the online learning method. We let the agent learn four types of functions: constant,
linear, nonlinear, and random. We chose these for their learning complexity with respect to the MLP predictors, thus better illustrating the effectiveness of the framework.

Results

Simulation results showed that after some initial exploration steps, the agent shifted its attention toward functions it expected to be compressible (predictable) through further learning. For easy patterns (constant and linear functions), the predictors could learn quickly; thus progress was high at the beginning but diminished rapidly. For hard patterns (XOR function), more samples were required before the predictor began to make progress, but once learning started, the curiosity reward was higher, and learning progressed quickly. For the random patterns, the predictor quickly got bored, as expected by the theory.

Relevance of the results for the problem and for the project

Our results illustrate how systems with artificial curiosity can learn to deal with the unavoidable limitations of their predictor learning algorithms, by temporally focusing computational resources on those parts of the world that make learning easy given the previously learned knowledge. It validates in a simple but clear way that the principle of learning progress can be used to speed learning in environments where the agent can choose what to learn next.

We are now focusing on developing a more precise measurement of learning progress that also incorporates the minimum description length principle (MDL). Since the goodness-of-fit and the complexity of the network model are closely related and need to be jointly optimized, our future plan is to take into account the network complexity, so as to quantify compression progress more accurately.

Figure 5.5.4. (left) Exploration behavior in the first 500 steps. The simpler functions are the fastest to be learned and therefore get chosen most often. (right) Exploration behavior in the first 5000 steps. After initially choosing the simpler functions, the agent gets bored with them and focuses on the more complex XOR problem.

5.3.5 Measures of interestingness and their use in optimization

Background and Open Issues related to IM-CleVeR

Artificial curiosity relies on a measure of interestingness as intrinsic feedback for the action the agent takes. The measures studied so far have assumed an autonomous learning agent, but that may not be necessary. We studied the practical use of incorporating a measure of interestingness to guide exploration in classical optimization problems: One possible way of breaking down the problem of learning on the iCub to a more manageable level is to consider the case where complete complex action sequences are grouped into macros. In this case, curiosity-based exploration can be seen as a multi-armed bandit problem: at each moment, the robot needs to choose which among its (possibly infinite) macro options is most interesting and/or rewarding to execute. This setup can directly be framed as a costly black-box optimization problem.
Specific Goals
Our first goal was to develop a new measure of interestingness based purely on compression, and not on the learning capabilities of an agent. Our next goal was to incorporate this measure into black-box optimization techniques.

Approach and Methods
We introduced “coherence progress,” a novel, general measure of interestingness, independent of a particular robot or learning algorithm (Schaul et al., 2011a). Coherence progress considers the increase in the coherence of the data as obtained by any compressor when adding an observation to the history of observations thus far. We also put the idea of coherence progress into context, considering how such a measure of interestingness could help robots increase the autonomy of their exploration (Graziano et al., 2011).

We then presented “Curiosity-driven optimization” (CDO) a novel approach to costly black-box optimization, which incorporates a measure of interestingness derived from artificial curiosity (Schaul et al., 2011b). The algorithm uses Gaussian process regression to efficiently model both the reward structure and the interestingness in the space of macro options, making explicit the trade-off between exploration (interestingness) and exploitation (reward).

Results
Because of its applicability to any type of compressor, coherence progress allows for an easy, quick, and domain-specific implementation. We demonstrated its ability to satisfy the requirements for qualitatively measuring interestingness on a Wikipedia dataset. The CDO approach makes the exploration-exploitation trade-off explicit, and permits maximally informed data selection. We illustrated the robustness of the approach in a number of experimental scenarios, including the Branin benchmark (Jones, et al., 1998).

Advancement of work and relation to other tasks
This work has yielded a deeper understanding of curiosity and its relationship to exploration. We found that measures of interestingness can successfully be based on learning-independent criteria. They are also more versatile in their use than previously thought, which allows for better exploration in pure optimization tasks.

5.4 Conclusion to task
Curiosity signals are essential for many of the project tasks, especially the hierarchical learning mechanisms in WP6 and the detection of abstract features in WP4. Conversely, the visual inputs of Task 4.5 will be of use to the algorithms in Task 5.5. Much of the work above can be adapted to the iCub and to the construction of hierarchical behaviours (Task 6.3, particularly with respect to the Mot system and the option library). In practical applications, we have made significant progress in using curiosity for principled exploration. Our theoretical and empirical results validate the key insights of the theory of compression progress, and demonstrate that with this principle, agents can learn to deal with the unavoidable limitations of their learning algorithms, focusing computational resources on parts of the task where learning is easiest. We now turn to extending the application of curiosity to the iCub and the algorithms of Task 6.3, and to developing more precise measurements of learning progress that incorporate the minimum description length principle (MDL).
6. Ongoing collaboration between partners

6.1. CNR – various partners

LOCEN-ISTC-CNR has been promoting discussions on the conceptual analysis of intrinsic (vs. extrinsic) motivations with the whole consortium (see section 1.3.1). Interaction on this is particularly intense with Prof. Barto (see section 1.3.6).

On the different kinds of intrinsic motivations, and in particular with respect to the distinction between knowledge-based and competence-based intrinsic motivations (see section 1.3.3), there have been particularly intense debate between LOCEN-ISTC-CNR and IDSIA.

There have been fruitful discussions between LOCEN-ISTC-CNR and USFD regarding dopamine and its supposed relationship with the TD-error of computational reinforcement learning (see section 1.3.4).

There have been fruitful interactions between LOCEN-ISTC-CNR and Prof. Barto's team on possible models of competence-based intrinsic motivations (see sections 1.3.5 and 1.3.6).

Task 5.2 is being conducted in close collaboration with WP3, Tasks 3.4–3.6. We construct our models in an attempt to make connections with the joystick task and to suggest experiments to be conducted with the joystick task in humans.

Task 5.2 will make stronger connections with WP5, Task 5.1. The current work done on Task 5.2 focused on building the infrastructure that allows us to study how novelty detection drives behaviour. Now that much of that infrastructure has been developed, we will incorporate reinforcement signals that conform to ideas developed in Task 5.1.

Task 5.2 has connections with WP5, Task 5.4, WP6, Task 6.2, and looser connections with other tasks involving the iCub. We are adapting our code, in collaboration with AU, to enable AU's iCub to perform a variant of the joystick task. Much of this initial work is focused on the communication between the iCub and high-level Matlab code we (and other psychologists) use in constructing neural network models.

Task 5.2 has general connections with tasks of WP6 in that future work in Task 5.2 is aimed at how behaviour generated by simple BG-mediated mechanisms can train higher-level controllers that represent "actions" as conceived in many formulations of action learning. This approach is "bottom-up" in that much of the research will focus on the lower-level movements that will compose an action that can be recruited as a discrete unit, rather than how high-level skills can be constructed from combining those discrete units. However, general properties of hierarchical skill construction are relevant and will inform our research.

6.2. UU - IDSIA-SUPSI joint, ongoing work on novelty detection

In WP4 we investigated the problem of scene analysis for world understanding. Once we have a model of the surrounding environment we can use this information to estimate a novelty index and, in this way, drive exploration. This goal can be achieved by (a) detecting parts of the perceived world where it is possible to find interesting objects, and (b) analyzing the novelty of the information contained in them.

We proposed an approach based on the "Bag of features model" (BoF) (Sivic et al., 2005), where the novelty of an object is described by a set of features that represent its appearance. The BoF is an extension of “Bag of words” (BoW), a method commonly used in natural language processing (NLP) for representing documents. BoW ignores word order, allowing dictionary-based modeling, such that each document looks like a "bag" which contains words from a common dictionary. A similar idea can be used for object representation and recognition, treating the features extracted from an image as the "words" of a document. Typically, three steps are involved: feature detection, feature description and codebook generation (Fei-Fei and Perona, 2005).
We proposed a particular feature detector and descriptor called “Speeded-up robust features” (SURF) (Bay, et al., 2008). This is a performant interest-point detector and descriptor that is scale and rotation invariant and allows identification and recognition of interesting points in an image in real-time (one of the most important requirements in robotics applications). Our method has comparable performance to (and occasionally outperforms) previously proposed schemes (e.g., Harris Affine, SIFT) with respect to repeatability, distinctiveness, and robustness, yet is much faster computationally.

Our improvement in computational performance is achieved by relying on integral images for image convolutions, building on the strengths of the leading existing detectors and descriptors (using a Hessian matrix-based measure for the detector, and a distribution-based descriptor) and simplifying these methods as much as possible. The codebook is generated using the habituated Neural Network proposed by UU, clustering similar features and describing each object with a list of “visual words” and the novelty of an object is estimated by comparing its image with all the objects perceived in the past.

6.3. UU–AU ongoing collaborative work on intrinsically motivated saccade learning driven by novelty detection

The need for this collaboration came to light during the workshop on Visual Processing and Novelty Detection hosted in UU in November 2010. Discussion in the workshop revealed that the saccade learning system developed at AU (Task 4.4) could benefit from the object learning and novelty detection developed at UU (Task 5.3). This collaborative work integrates Tasks 5.3, 4.4 and 5.4 on the iCub platform. It also integrates work from Task 4.2 regarding image segmentation and identification of the regions of interest.

As a result of further discussions, two researchers from UU visited AU during the period of 13th - 15th June to realise the integration of the individual modules of the two partners. Specifically, the novelty detector and region of interest selector developed at UU was integrated with the saccade learning system developed at AU. This integration required the development of bridging software in order to allow interoperability and communication of the individual modules of the two partners. Through the combined effort of both the UU and the AU researchers, the necessary software environment for interoperability of current and future modules is now in place.

This on-going collaboration has so far resulted in the successful integration of the individual modules on the iCub robot platform. In particular, the iCub was equipped with the ability to identify the most novel object on a table in front of it, as shown in Figure 6.3.1. The result of the novelty detection module developed by UU was then passed to the saccade learning module developed at AU. The saccade learning module was then able to learn to successfully control eye movement in order to centre on the most novel object, as shown in Figure 6.3.2.

Upon successful saccade learning, as shown in Figure 6.3.3, the visual characteristics of the object were learnt by the novelty detector so that the object is able to be recognised in the future, as shown in Figure 6.3.4. At the same time the novelty detector module habituates on the learnt object. Upon successful completion of the object learning, the novelty detector suggests to the saccade learning module either a new novel object (Figure 6.3.5) or an already learnt one that has become novel again due to dishabituation (Figure 6.3.6). This process is continuously repeated and is summarised by the flowchart diagram shown in Figure 6.3.7. Further details on the implementation of the individual modules are given in Task 5.3 and Task 4.4.

In conclusion, the three intense days have been highly productive and the outcome of this collaboration was the successful integration of the individual work of the two partners. It also demonstrates cross-collaboration between Clever-B and Clever-K partners. The necessary software for future collaborations has been created, and further collaboration is planned to extend the system even further.
Figure 6.3.1: Detecting most novel region of interest

Figure 6.3.2: Learning saccade

Figure 6.3.3: Saccade learning map

Figure 6.3.4: Training on the object after successful saccade learning
Figure 6.3.5: Novelty detector selecting an unknown object

Figure 6.3.6: Novelty detector selecting an already known object that has dishabituated

Figure 6.3.7: Integration flowchart of the modules developed by Ulster and Aberystwyth
7. Conclusion to the deliverable

WP5 on novelty detection and drives for autonomous learning has been investigated by different partners under 5 different tasks.

The overall goal of Task 5.1 (‘intrinsic and extrinsic motivations for cumulative learning’) is to investigate the conceptual difference between extrinsic and intrinsic motivations, to understand the brain mechanisms that support both kinds of motivations and the complementary roles that they play in driving cumulative learning in animals and to propose hypotheses and models that facilitate the ability of cumulative learning, which can be implemented in artificial systems. The research on this task has contributed several important results. With respect to the conceptual distinction between intrinsic and extrinsic motivations, clearly significant contributions have been made. With respect to extrinsic motivations, the developed theoretical framework has already proved useful for the development of several aspects of the project. With respect to intrinsic motivations, important contributions have been provided regarding the distinction between knowledge-based and competence-base intrinsic motivations, hypothesis on the role of dopamine in conditioning and intrinsic motivations based on competence.

Task 5.2 (‘novelty-based formation of actions: modelling the joystick experiment’) investigates and develops modelling work meant to enable us to study how brain mechanisms mediate behaviour. The "joystick task" enabled to study how humans develop this behaviour as most joystick movements result in no signal. The modelling framework used here in this task is based on that used by Kevin Gurney and colleagues to study how neurons in the basal ganglia, cortex, and thalamus interact to perform action selection. Important contributions and research outputs have been achieved in this task.

The Task 5.3 (‘novelty detection based on habituable networks’) reports on original developments in techniques for visual perception based novelty detection and learning, exploiting habituation. The habituable neural-network of Marsland has been extended with the bag-of-words model for effective perceptual learning and object recognition, which resulted in a novel system. The model has been successfully tested on physical robots. However, a particular limitation to the original methodology of the bag-of-words that prevents it from being efficiently used in a cumulative learning task has been identified. These limitations were addressed in further research. The novel novelty detector integrates an online dynamically expandable bag-of-words that UU researched and developed. Again, it was tested with a physical robot with real-world objects. It was also shown that the system can detect the outcome of actions of basic actions such as push, pull, topple, etc. in tests with both a physical robot and a simulated scenario.

The Task 5.4 (‘novelty based development of redundant sensory-motor systems’) investigates the use of novelty-based systems to drive learning of sensori-motor coordination through motor babbling and play. The task aims to investigate how different motivators cause the robot to develop along different developmental trajectories by implementing different novelty detectors. To date the research explored the ideas of motor babbling and play in infancy, the importance of novelty as a driving mechanism and how developmental stages may emerge through novelty-driven exploration.

Task 5.5 (‘Information-theory indexes as sources of novelty signals’) addresses novelty detection based on information-theory indexes as sources of novelty signals and planning for efficient exploitation of non-stationary rewards. Theoretical investigation has been carried out on the principle of optimal Bayesian exploration in dynamic environments without external rewards. The use of a novelty signal to improve learning efficiency in RL tasks with visual input has been studied. The research focused on empirical validation of learning progress as an exploration mechanism, especially in the setting of online interactive learning. The developed models/systems have been verified in simulation. It has been showed that the agent can optimally choose actions based on previous experiences, such that the cumulative expected information gain is maximized, and proved the existence of an optimal exploration policy.
8. References


Baldassarre, G., Mirolli, M. What are the Key Open Challenges for Understanding Autonomous Cumulative Learning of Skills?, Autonomous Mental Development Newsletter, 7(2): 2-9 (target column with commentaries and response), 2010


Burbridge, C., Gatsoulis, Y., McGinnity, T.M. “Cumulative online unsupervised learning of objects by visual inspection”, Submitted to IROS 2011.


Mannella, F., Mirolli, M., Baldassarre, G. The interplay of pavlovian and instrumental processes in devaluation experiments: a computational embodied neuroscience model tested with a simulated rat. In Tosh Colin and Ruxton Graeme (eds), "Modelling Perception With Artificial Neural Networks", pp. 93-113. Cambridge, Cambridge University press. 2010


Rankin, C. H. “Introduction to the special issue of neurobiology of learning and memory on habituation”. Neurobiology of Learning and Memory, 92, 125-126, 2009.


Ratnasingam, S., McGinnity, M. “A Spiking Neural Network for Tactile Form Based Object Recognition” Accepted in IEEE,IJCNN, San Jose, California, July 31 - August 5, 2011b.


Santucci, V.G., Baldassarre, G., Mirolli, M. (in preparation): Intrinsic reinforcements provided by unexpected events can drive the cumulative acquisition of skills (to be submitted to Neural Networks).


