CHIST–ERA Proposal

Project Acronym:

DELTA

Project Title:

Dynamically Evolving Long-Term Autonomy

Addressed Call Topic (LLIS or VADMU): LLIS

Coordinator contact point for the proposal

Name	Anders Jonsson
Institution/Department	Dept. of Info. and Comm. Technologies, Universitat Pompeu Fabra
Address	Roc Boronat 138, 08018 Barcelona
Country	Spain
Phone	(+34) 935422952
Fax	(+34) 935421440
E-mail	anders.jonsson@upf.edu

Partners' people involved in the realisation of the project

Partner Number	***Country***	***Institution/ Department***	***Name of the Principal Investigator (PI)***	***Name of the co-Investigators***	***Name of the other personnel participating in the project***
1 *Coordinator*	Spain	UPF	Anders Jonsson	Gergely Neu, Vicenç Gómez	Post-doc
2	France	INRIA	Michal Valko	Alessandro Lazaric, Emilie Kaufmann	Post-doc, Ph.D student
3	Austria	MUL	Peter Auer	Ronald Ortner	Post-doc
4	Belgium	ULG	Bertrand Cornélusse	Damien Ernst	Post-doc

Duration: 36 months

Summary of the project:

Many complex autonomous systems (e.g., electrical distribution networks) repeatedly select actions with the aim of achieving a given objective. Reinforcement learning (RL) offers a powerful framework for acquiring adaptive behaviour in this setting, associating a scalar reward with each action and learning from experience which action to select to maximise long-term reward. Although RL has produced impressive results recently (e.g., achieving human-level play in Atari games and beating the human world champion in the board game Go), most existing solutions only work under strong assumptions: the environment model is stationary, the objective is fixed, and trials end once the objective is met.
The aim of this project is to advance the state of the art of fundamental research in lifelong RL by developing several novel RL algorithms that relax the above assumptions. The new algorithms should be robust to environmental changes, both in terms of the observations that the system can make and the actions that the system can perform. Moreover, the algorithms should be able to operate over long periods of time while achieving different objectives.
The proposed algorithms will address three key problems related to lifelong RL: planning, exploration, and task decomposition. Planning is the problem of computing an action selection strategy given a (possibly partial) model of the task at hand. Exploration is the problem of selecting actions with the aim of mapping out the environment rather than achieving a particular objective. Task decomposition is the problem of defining different objectives and assigning a separate action selection strategy to each. The algorithms will be evaluated in two realistic scenarios: active network management for electrical distribution networks, and microgrid management. A test protocol will be developed to evaluate each individual algorithm, as well as their combinations.

Relevance to the topic addressed in the call:

RL allows autonomous systems to improve themselves without human intervention, a central subject of the call. The novel algorithms aim to take advantage of previous knowledge each time the environment changes, avoiding relearning from scratch as much as possible. Many objectives will be initially unknown, and the system has to learn to achieve new objectives as they appear. Active exploration of the environment is of fundamental concern to the project, and the algorithms for task decomposition will automatically identify new subgoals to achieve. The project aims at improving on the state-of-the-art in the two scenarios, as well as increasing their flexibility by allowing distribution networks to change over time. The test protocol will require different metrics than those typically used in RL to evaluate algorithms.

1. S/T Quality

1.1 General objectives of the project

From a decision-making perspective, the key to achieving long-term autonomy is the ability to adapt to changing circumstances. This ability is limited in today's autonomous systems, for several reasons. If the decision strategy is not adaptive enough, it will always select the same action in a given situation, even after it should have discovered that another action choice is superior. An environment model with fixed sets of variables and actions may preclude a system from interacting with an object it does not recognize. Predefined objectives may turn out to be insufficient, and prevent the discovery and completion of new objectives. These drawbacks are exacerbated for autonomous systems that remain operational for long periods of time, increasing the likelihood of experiencing changes.
To successfully negotiate a changing environment, an autonomous system needs to address several issues. First, the decision strategy should be adaptive, allowing the system to evaluate and modify its action choices. Second, the environment model should be flexible and, apart from a decision strategy that achieves system objectives, the system should include an exploration strategy that explicitly selects actions with the aim of updating and refining its environment model. Third, the system objectives should be allowed to change over time, and the system should efficiently learn to achieve new objectives on the fly.
In this project, we adapt the framework of reinforcement learning (RL) to changing environments. Existing work on traditional RL uses a rich set of tools for evaluating and updating action choices to identify decision-making policies that lead to high expected long-term rewards. To extend this line of work to deal with changing environments, the project aims to develop several novel RL algorithms that can operate for extended time periods and solve multiple different objectives, some of which are initially unknown. These algorithms should accept changes to the environment model, improving the ability of a system to interact with new objects it encounters. Together, the novel algorithms will increase the applicability of RL to real-world autonomous systems.
The project objectives are defined as follows:

- **Objective 1**: Develop a novel planning algorithm that efficiently achieves a new, previously unknown objective given the current environment model of the system. The planning algorithm should account for the fact that the environment model may change over time.
- **Objective 2**: Develop a novel exploration strategy for RL that automatically and efficiently updates the environmental model, by selecting actions that explore parts of the environment that the system is not yet familiar with.
- **Objective 3**: Develop a novel framework for task decomposition that automatically creates and evaluates tasks, discarding tasks that are not deemed useful. Each task has its own associated decision strategy.
- **Objective 4**: Evaluate the novel planning and RL algorithms in two realistic scenarios: active network management for electrical distribution networks, and microgrid management. Apart from using these scenarios for evaluation, the project also aims at improving on the state-of-the-art in these two applications.

1.2 State of the art

Reinforcement learning (RL) [1] enables a system to optimise action selection from experience. The fundamental elements are *states* that describe the current configuration of the environment and *actions* that modify the environment. A *factored state* is expressed as a joint assignment to multiple *variables*. After taking an action in a state, the environment transitions to a new state and the system receives a scalar reward. A *Markov decision process* (MDP) extends states and actions with *transition probabilities* and *expected rewards* that govern the outcome of taking actions. The aim is to select actions that maximise long-term reward. A selection strategy is often stored as a *policy* that maps states to actions. The *regret* of a policy is defined as the loss of reward when compared to an optimal policy.

Given an environment model (usually an MDP), planning is the problem of finding the best action in the current state. In complex MDPs, exact dynamic programming is intractable. One efficient approximation is Monte-Carlo Tree Search (MCTS). MCTS constructs a search tree and performs rollouts to estimate the long-term reward associated with actions. UCT [2,3] is a particular MCTS algorithm with known practical success [4]. However, it is still only a heuristic which works in some cases and fails in others. Exploring the possibilities uniformly at random is the most trivial example of MTCS which never fails, but with exponential sample complexity (number of calls to the model). The ideal MCTS planning algorithm needs to 1) always find the best actions, 2) have sample complexity adapting to the difficulty of the task, and 3) be computationally efficient. Only the most recent algorithms [21,22] attempt to tackle all three requirements. However, none of them can handle changing environments.

In the more complex RL framework, where the environment needs to be learned, the *exploration-exploitation tradeoff* has to decide either to explore (act as to improve the estimate of the environment model) or exploit (act according to the current best policy). UCRL [7] is an algorithm that balances exploration and exploitation by optimistically choosing a model from estimated confidence intervals for transition probabilities and expected rewards. A feature of UCRL is that its regret is provably bounded giving e.g. problem dependent regret bounds logarithmic in the number of steps taken. For factored states, a variant called UCRL-Factored [8] achieves regret that is polynomial in the number of variables encoding the state.

There exists a variety of methods for automatically identifying subtasks, of which the most relevant to the project are those that assume that the state is factored [9,10,11]. There is not much work in the literature about evaluating how useful subtasks are, but there have been attempts to select the best subset of subtasks from a given set [12,13] and to compute policies that achieve sublinear regret across a range of subtasks [14].

Active network management (ANM) for electrical distribution systems [15] models the flow of electrical power through a given distribution network. ANM can be modeled as a RL problem in which states are the current power levels of devices in the network, and actions are to control the power injected by generators or to reduce the consumption of loads. The aim is to select actions as to avoid congestion and move consumption to anticipated periods of high renewable generation. Microgrid management is a novel application that shares features with ANM. A microgrid is a small distribution network that acts as a single entity with respect to the main grid. At least two associated tasks can be defined: interact with the main grid, or stabilize the microgrid in case of disconnection from or reconnection to the main grid.

RL in changing environments is related to Bayesian RL [16], which has uncertainty regarding the true underlying environment model. However, Bayesian RL typically assumes that there exists a finite set of known candidate models that includes the true model, an assumption

that does not hold when the model can change in arbitrary ways. The factored representation used in the project is related to object-oriented and relational representations for RL [17,18].

1.3 Approach and research method

This section describes the proposed methodology for achieving the project objectives. For each of Objectives 1-3, the outcome will be a software module that implements the novel algorithm developed as part of achieving the objective. To facilitate integration of the software modules as part of Objective 4, they will be based on a common environment model, described next.

To take advantage of problem structure, the environment model assumes that states are factored. Moreover, the model is flexible and allows variables and actions to vary over time. The project considers an object-oriented representation that assigns attributes and methods to object classes. When a system first encounters an object, the attributes and methods of the corresponding object class induce new variables and actions that are incorporated into the environment model, effectively preparing the system to interact with the new object. Likewise, an object can disappear by removing the associated variables and actions. As a result, changes to the environment model are not arbitrary, and the new model strongly resembles the old model in that most of the variables and actions remain the same.

A system objective is defined as a partial assignment of target values to a subset of the variables encoding the factored state. In the case of continuous variables, the target is an interval rather than a single value. The task associated to each system objective is defined on the same set of variables and actions, so for a fixed environment model, only the reward structure changes. However, any change to the environment model affects each of the tasks.

Apart from empirically evaluating the different algorithms in the two scenarios, we also aim to derive theoretical bounds on their performance. The combination of empirical and theoretical analysis should result in stronger algorithms that advance the state-of-the-art in lifelong RL.

Objective 1: Develop a novel planning algorithm for MDPs that efficiently achieves a new, previously unknown objective given the current environment model of the system.

We propose a change of paradigm for MCTS algorithms, enabling them to deal with new nodes, new states, and a new model. Most MCTS algorithms, including UCT, take an optimistic approach to selecting the branch of the tree to descend, using upper confidence bounds on the rewards to be gained, as the UCB algorithm for bandit models [23]. However, it can be observed that new methods [21,22] instead rely on the adaptation of algorithms for best arm identification (BAI) in bandit models (whose behavior is quite different from regret minimising algorithms [24]). We believe that recent optimal algorithms for BAI [19,20] could be successfully used within MCTS procedures, leading to improved performance in terms of sample complexity, with provable guarantees. BAI tools have also recently been used for MCTS in simple games [25], and we plan to extend it to cover any possible type of search nodes. When the environment changes, the objective is to allow to use the information from the tree before the change occurred in order to minimize the number of samples needed to learn the best action in the new situation. For this we will use an adaptive hierarchical partitioning of the space, which is able to reuse data from the time before the change occurred.

Objective 2: Develop a novel exploration strategy for RL that automatically and efficiently updates the environmental model.

Some RL algorithms already exhibit the ability to adapt to changing environments, but they are either restricted to changes of only the rewards, or to a few disruptive and complete changes of the environment. The starting point for achieving Objective 2 is the UCRL algorithm developed by MUL, which – using restarts – can tolerate disruptive changes. We expect that a suitable modification of UCRL and its analysis will shed light on the principal mechanisms, that allow RL algorithms to cope with gradually changing environments.
Whereas any RL algorithm depends on exploration in an unknown environment, we will consider incremental and conservative exploration. This is necessary in very large environments where it takes too long to find the overall optimal policy. Instead, we propose to start with a (nearly) optimal policy for a small part of the environment, and then enlarge this small known part incrementally. Thus the algorithm will provide a growing model of its environment that adapts to changes. This form of exploration may also provide a basis for subtask identification.
To deal with large state spaces efficiently, we will incorporate the hierarchical partition developed as part of Objective 1 into the algorithms for exploration of the environment. This will allow to employ the algorithms in the demanding scenarios considered in this proposal. For example, in active network management, the exploration algorithm will start with investigating the modulation of only a few loads, and then move on to more complex modulation patterns. For an efficient and useful exploration it is essential that a compact representation of the state space can be achieved. Such a compact representation may be provided by a hierarchical partitioning.

Objective 3: Develop a novel framework for task decomposition that automatically creates and evaluates tasks, discarding tasks that are not deemed useful.

In lifelong RL, it is unrealistic to assume that a system always pursues the same task. Moreover, a complex task can be difficult to achieve using a single decision strategy, and in realistic applications it is common to decompose complex tasks into subtasks.
The proposed approach is to maintain a library of known tasks that the system can carry out, and associate a separate decision strategy (i.e. a policy) with each task. Each task in the library achieves a single system objective as defined above. The system has to manage the task library on its own, which involves adding and deleting subtasks as well as updating the policy associated with each task. The policy of a task may select between other tasks, effectively forming a task hierarchy. The task library allows the system to seamlessly switch from one task to another as necessary.
Tasks are added either as a result of encountering a new system objective, or by identifying useful subtasks. As a starting point, we will consider existing methods for automatically identifying subtasks, specifically those that assume a factored state. However, few of these methods work for continuous variables, making it necessary to develop novel methods for subtask identification in the two scenarios. The exploration algorithm developed as part of Objective 2 may also help identify useful subtasks.
The system should also include a mechanism for discarding tasks that are not deemed useful. Repeatedly adding new tasks to the task library will eventually cause problems, both in terms of the memory required to store the policy associated with each task, and in terms of the increasing complexity when selecting between a growing number of subtasks. The project proposes to develop novel metrics for evaluating the utility of subtasks. One way in which we propose to evaluate a subtask is to measure the regret incurred when selecting it across a range of problems, compared to the best action or subtask in each of those problems.

Objective 4: Evaluate the novel RL algorithms in two realistic scenarios: active network management for electrical distribution networks, and microgrid management.

With the increasing share of renewable and distributed generation in electrical distribution systems, Active Network Management (ANM) has become a valuable option for a distribution system operator to operate his system in a secure and cost effective way without relying solely on network reinforcement. ANM strategies are short-term policies that control the power injected by generators and/or taken off by loads in order to avoid congestion or voltage issues. While simple ANM strategies consist of curtailing temporary excess generation, more advanced strategies instead attempt to move the consumption of loads to anticipated periods of high renewable generation.
An adaptative model that can cope with disruptive changes in the dynamics of the system is critical for ANM applications. The dynamics of distribution systems is influenced by several uncertain factors and is constantly changing. These factors include the behavior of thousands of residential consumers and of hundreds of production units. The extent of these evolving factors excludes expert-driven customization of control policies and calls for a procedure that adapts automatically to these changes.
Realtime control of microgrids is another industrial application where adaptiveness to changing environment is of paramount importance. A microgrid is a fully capable small scale decentralized energy network with its own energy supply sources, electrical loads and power transmission methods. A microgrid optimizes its interactions with the system to which it connects (e.g. the main electricity grid). Internally, the consumed electricity can either come from the local production (photovoltaic panels, wind turbine, hydroelectric plant...) or from the regular distribution network. One typically wants to satisfy as much as possible local consumption (demand) with local production (supply) whenever local production is cheaper than the electricity from the regular distribution network. However, the local production is sporadic and does not necessarily match the consumption pattern. With the advent of cheaper electricity production units and storage means, measures can be taken to mitigate the mismatch between the local supply and demand patterns.
In this application, the goal is to devise adaptive action policies that can cope with deviation from a calculated operational plan. Deviations include unplanned events such as a lower energy production than estimated or an erroneous prediction of the electricity retail price. It is also very much desirable that the microgrid controller automatically adapts to changing environmental conditions (e.g. connecting or disconnecting loads, switching from connected to islanding mode).
The design of the experimental settings will be driven by the aim of evaluating the performance of the adaptive policies at controlling realistic and accurate simulators of the studied applications.
In each of the two scenarios, several alternative autonomous systems will be evaluated. For each of Objectives 1-3, we will consider two alternative algorithms: a traditional RL algorithm that does not account for changes in the environment, and the algorithm developed as part of the objective. This way we can swap algorithms in and out to evaluate their individual contribution to the system's performance. One autonomous system will include all novel algorithms developed as part of the project, allowing us to evaluate the full combination of the novel algorithms. Each autonomous system will be evaluated for extended periods of time to test its lifelong learning ability. During evaluation, devices will be added and deleted from the distribution networks, and new objectives will be introduced, effectively testing the ability of the system to adapt to a changing environment.

In order to guarantee that other researchers can reproduce the results of the project, the two scenarios as well as all software modules will be published in a public repository managed as part of the project.

1.4 Targeted outcomes

The main targeted outcome is to develop an integrated autonomous system which is able to successfully interact with its environment during long periods of time without breaking down or failing to achieve its objectives. This in itself would be a considerable improvement over the state-of-the-art, since we are aware of no realistic applications of lifelong reinforcement learning that are capable of operating for extended time periods while adapting to changes in the environment.

The above targeted outcome will be tested empirically in the two simulated scenarios: active network management (ANM) and microgrid management. The simulators developed as part of the project will make it possible to test the autonomous system for long time periods while switching between objectives and introducing changes to the environment, effectively providing a rich, realistic benchmark in which to evaluate the different RL algorithms. Success will be measured using several different criteria: 1) the ability of the system to achieve its objectives, including objectives that were initially unknown; 2) the ability of the system to recover and adapt as a result of changes to the environment model, without having to relearn from scratch; 3) the ability to learn and plan faster over time by taking advantage of the estimated environment model. In RL, the ability to satisfy these criteria can be measured by recording the cumulative reward over time: a higher reward means that the system is better at achieving its objectives.

Another targeted outcome is to improve on the state-of-the-art for ANM and microgrid management. So far, only methods from Mixed Integer Nonlinear Programming or metaheuristics have been applied to these problems. We expect that reinforcement learning should have an advantage over these previous approaches, since it allows the decision strategy to improve over time and since the novel RL algorithms will allow the application to adapt to changes in the environment. Since results are available for previous approaches, it is straightforward to measure the performance of RL algorithms in ANM and microgrid management, and compare them to previous approaches to see which one works better.

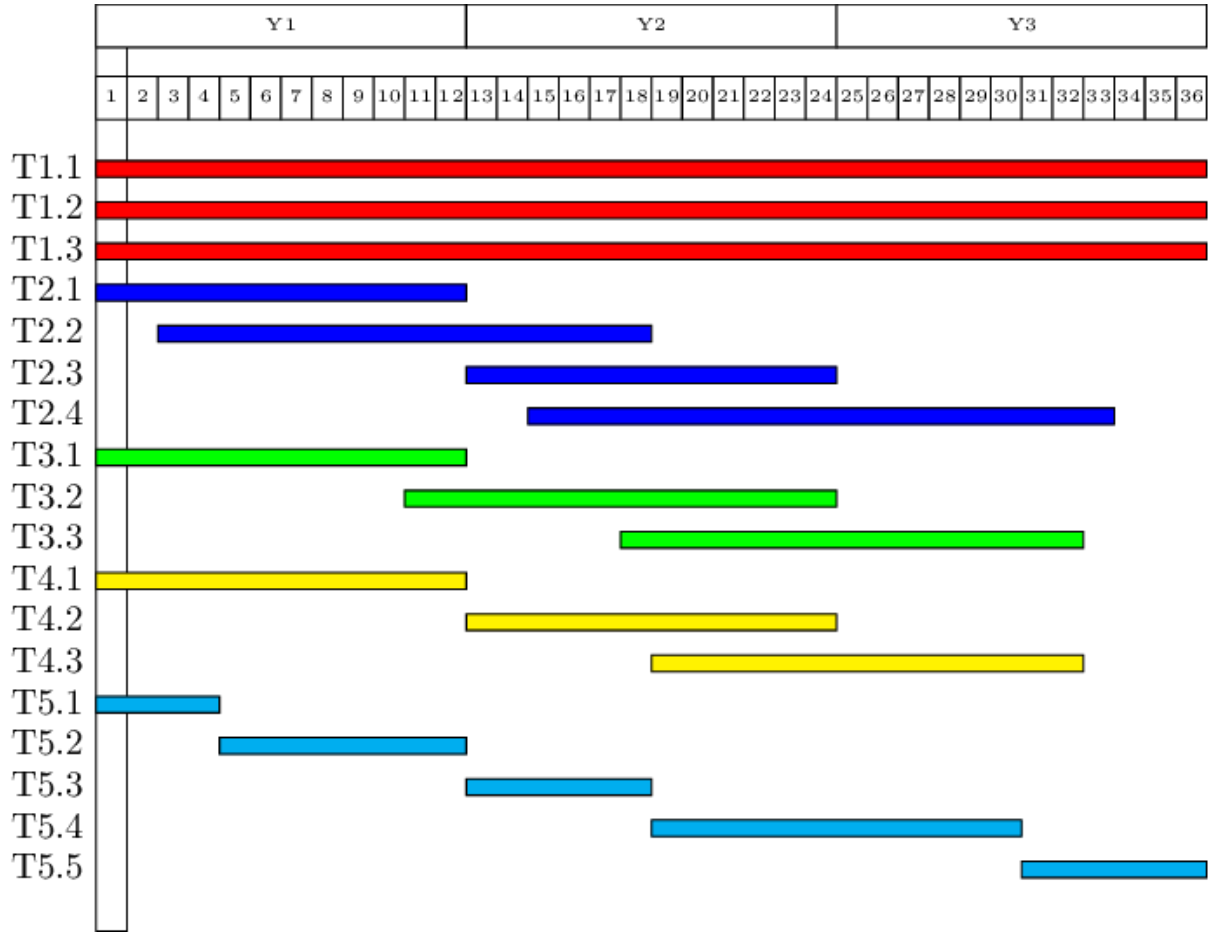
Several of the novel RL algorithms are directly related to the target outcomes specified in the LIIS call. The proposed extension of the UCRL algorithm will be able to actively explore its environment. Exploration will be guided by the current task, either as a result of achieving a new system objective, or as a result of identifying useful subtasks. The proposed method for refining the hierarchical partition will avoid the need for regression and take advantage of existing knowledge about the old environment model. The evaluation protocol proposed in the previous section will help measure the performance of each individual RL algorithm.

Finally, another targeted outcome is to disseminate the novel RL algorithms developed as part of the project, both by publishing scientific papers and by making the resulting software modules available as part of a public repository. In this way we hope to contribute to the progress of lifelong RL beyond the scope of this project.

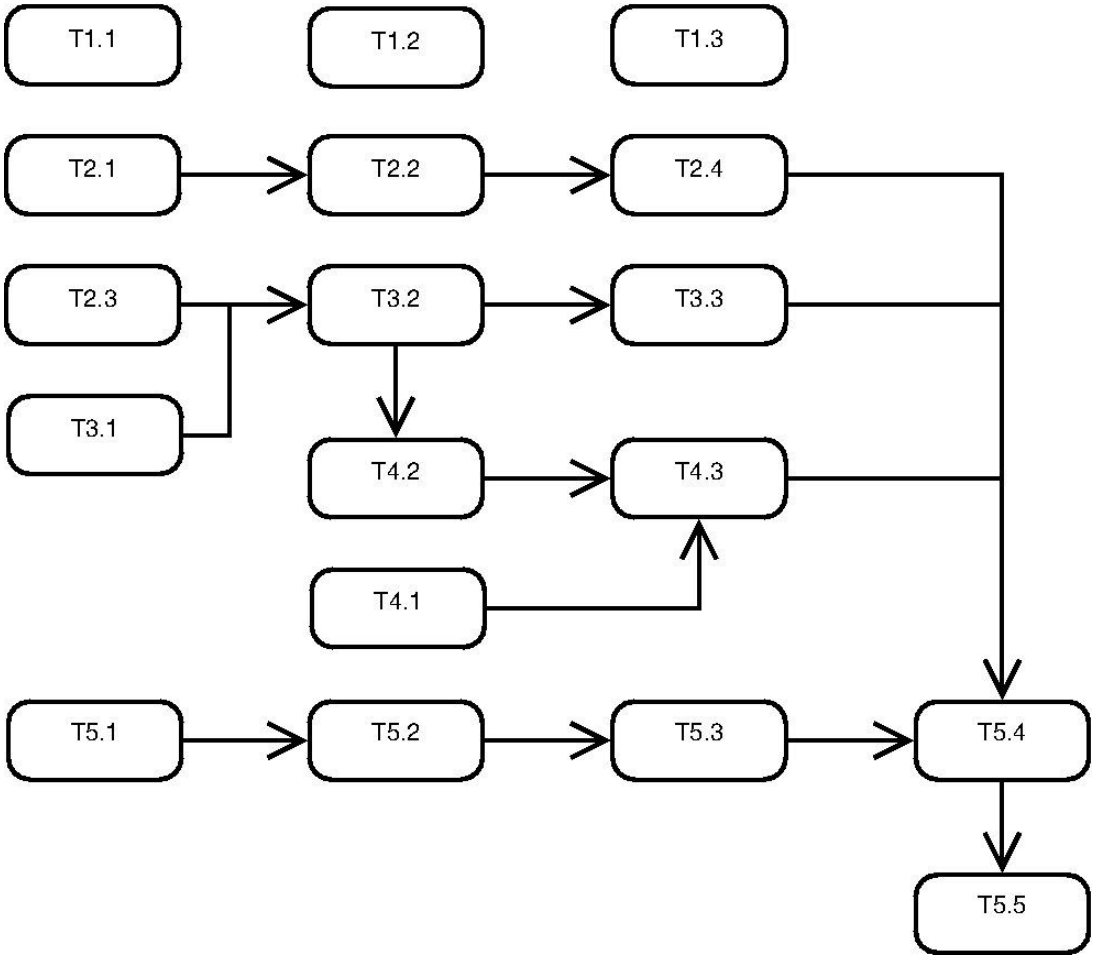
2. Implementation

2.1 Work plan

The DELTA workplan is divided into five workpackages (WPs) comprising a total of 18 tasks. WP1 is concerned with project management and dissemination, while WP2-WP5 correspond to the four general objectives of the project, described in Section 1.1. Specifically, WP2 aims at developing a novel planning algorithm for RL based on Monte-Carlo tree search (MCTS), WP3 aims at developing a novel exploration algorithm for RL based on UCRL, and WP4 aims at developing novel methods for task decomposition. The outcome of each of these workpackages will be a software module to be integrated into the two scenarios. The purpose of WP5 is thus to develop the simulators that will provide the backbone of the two scenarios, and integrate all software components from the other workpackages in order to build the final autonomous system and measure the performance of the RL algorithms in a realistic setting. The resulting autonomous system will be used for dissemination and to sell the project idea to potential business partners.
The timeline of the different workpackages and tasks are shown in the following Gantt chart, which has been color coded (WP1: red, WP2: blue, WP3: green, WP4: yellow, WP5: cyan):



The following Pert chart shows the interrelation of tasks in the project. The main thing to notice is that the final task of WP2-WP4 each feeds into Task T5.4, which integrates all software modules into the final application.



Work package overview (total effort per WP and partner in person months)

Partner	WP1	WP2	WP3	WP4	WP5	Total
1	5			28	4	37
2	1	28	4		2	35
3	1	4	28	2	2	37
4	3	2			33	38
Total	10	34	32	30	41	147

2.2 Work packages

WP 1	**Management and dissemination**					Start month: 1		End month: 36
Contribution of project partners								
Partner number	**1**	2	3	4				
Total effort per partner (Person*months)	5	1	1	3				

Aim of the WP

Coordinating project activities by ensuring effective internal communication and providing adequate information to decision-making bodies ● Managing legal, ethical and administrative resources ● Ensuring compliance with EC rules and the Consortium Agreement.
Disseminating project results through activities such as: ● publishing the research results in leading journals and high profile conferences; ● establishing contacts with relevant research projects, institution, researchers, publishing houses, etc; ● creating awareness of the project in relevant social networks; ● planning tutorials, workshops or teaching activities related to DELTA topic.

Tasks

T1.1	**Communication with the EC (M1–M36; responsible: 1)** The Project Coordinator will collect and integrate all technical, administrative and financial data from partners and prepare documents for the European Commission: management reports, progress reports, final report, cost and financial statements, deliverables, etc. The Project Coordinator is responsible for preparing the Consortium Agreement.
T1.2	**Organisation of management meetings (M1–M36; responsible: 1)** The Project Coordinator will prepare and organise regular meetings of the Steering Committee and the Project Manager Board.
T1.3	**Dissemination plan and activities (M1–M36; responsible: 4, involved: all)** This task is in charge of producing a dissemination plan for the project which will include partner-specific and project dissemination activities in their respective research communities (AI, machine learning, and power systems). Proper dissemination will consist in publications in relevant journals in the partner's fields of expertise and highly regarded conferences than undergo peer review. One of the foreseen activities is to promote the project through a set of dedicated social networking channels such as Twitter, ResearchGate, Mendeley, etc. All partners in the project will be encouraged to disseminate research results and activities through the agreed social networking sites to give visibility to the project. The project will also be disseminated through tutorials, invited talks, seminars, etc. at relevant research institutions and conferences.

Deliverable	**Month of delivery**	**Title of deliverable**
D1.1	M6	Internal Quality Assessment Plan
D1.2	M12;M24;M36	Periodics/Final Project Report
D1.3	M8	Dissemination Plan
D1.4	M18; M36	Dissemination Report (2 versions)

WP 2	Planning in complex changing environments						Start month: 1	End month: 33
Contribution of project partners								
Partner number	1	**2**	3	4				
Total effort per partner (Person*months)		28	4	2				

Aim of the WP

Develop new planning algorithms in general environments: from games to MDPs to changing environments

Tasks

T2.1	**Best-arm identification tools for planning (M1–M12: responsible partner 2)** In this task, we will investigate how optimal Best Arm Identification tools (BAI) can be used within MCTS algorithms, starting on the game example of [25] that will be extended to handle any two player game.
T2.2	**MCTS with general nodes in the state trees (M3–M18: responsible partner 2)** Combining the general game algorithm proposed in T2.1 to the BAI-MCTS algorithm of [22] for planning in MDPs, we will propose a new algorithm for general tree search, with sample complexity guarantees. General tree search may include MIN nodes in addition to MAX and AVG nodes, which extends MCTS to more complex scenarios.
T2.3	**Hierarchical partition of the states of planning (M13–M24: responsible partner 2; involved partners 3, 4)** We plan to build on adaptive hierarchical partitioning for function optimization [5,6] to deliver a method able to adaptively query a model of the environment (or the environment itself) and estimate the value of a (factored) state. We will design an algorithm that can deal with a previously unobserved dimension of a state by reusing known samples.
T2.4	**MCTS planning in changing environments (M15–M33: responsible partner 2; involved partners 3, 4)** In this task we will deal with the changes. We will consider updates in the model precision and or the news states and look for the opportunities to use a previously learned search tree in the new situation (to avoid relearning a new search tree from scratch).

Deliverable	Month of delivery	Title of deliverable
D2.1	M12	Best-arm identification tools
D2.2	M18	MCTS with general nodes in planning
D2.3	M24	Hierarchical state partitions in planning
D2.4	M33	MCTS planning in changing environments

WP 3	**Exploration**				Start month: 1	End month: 32		
Contribution of project partners								
Partner number	1	2	**3**	4				
Total effort per partner (Person*months)		4	28					

Aim of the WP
Develop RL algorithms that explore and adapt to changing environments with provable performance guarantees.

Tasks

T3.1	**RL algorithms for changing environments (M1–M12: responsible: 3; involve: 2)** The goal of this task is the development of RL algorithms that cope with gradual changes of the rewards and transitions while maintaining near optimal performance should the environment not change. The starting point is the UCRL algorithm developed by MUL [7], which – using restarts – can tolerate disruptive changes.
T3.2	**Open-ended exploration in changing environments (M11–M24: responsible: 3)** The goal of this task is to develop strategies for incremental and conservative exploration, starting with a nearly optimal policy for a small part of the environment, and gradually enlarging this small known part. This task will build upon the UCB-Explore algorithm developed by MUL [26] as well as on the results of Task T3.1 using the developed RL algorithms with modified objectives or rewards that lead to the desired exploratory behaviour.
T3.3	**Incorporating state space partitions into exploration (M18–M32: responsible: 3; involve: 2)** This task will incorporate the hierarchical partitioning of the state space developed in T2.3 into the algorithms for exploration of the environment. This will allow to employ the exploration algorithms in the demanding scenarios considered in this proposal.

Deliverable	**Month of delivery**	**Title of deliverable**
D3.1	M12	Reinforcement learning in changing environments
D3.2	M24	Open-ended exploration in changing environments
D3.3	M32	Incorporating state space partitions into exploration

WP 4	**Task decomposition**						Start month: 1	End month: 32
Contribution of project partners								
Partner number	**1**	2	3	4				
Total effort per partner (Person*months)	28		2					

Aim of the WP

Define metrics that measure the utility of subtasks ● Develop novel methods for automatically identifying subtasks ● Develop a framework for creating, evaluating and discarding subtasks.

Tasks

T4.1	**Subtask utility metrics (M1–M12; responsible: 1)** This task will define metrics for evaluating the utility of an individual subtask. The approach is to test a subtask across a range of different tasks, measuring its regret compared to the best subtask in each situation.
T4.2	**Identifying potential subtasks (M13–M24; responsible: 1; involve: 3)** This task will build on previous work in task decomposition to identify subtasks. This work will build on existing approaches to automatically discover subgoals. Another source of information is the exploration algorithms developed as part of task T3.2.
T4.3	**Task library management (M19–M32; responsible: 1)** This task will develop a framework for task decomposition that integrates the components from the other tasks of the work package. The framework will automatically identify and incorporate subtasks into its library, continuously evaluate them using the metrics developed in T4.1, and discard subtasks that are not deemed useful.

Deliverable	**Month of delivery**	**Title of deliverable**
D4.1	M12	Subtask utility metrics
D4.2	M18	Identifying potential subtasks
D4.3	M32	Task library management

WP 5	Integration	Start month: 1	End month: 36

Contribution of project partners								
Partner number	1	2	3	**4**				
Total effort per partner (Person*months)	4	2	2	33				

Aim of the WP

This WP integrates the developments of the other WPs. It will propose two scenarios of application of the algorithmic concepts developed within the DELTA project. Algorithmic ideas developed in the other work packages are implemented for these scenarios.

Tasks

T5.1	**Adaptation of the ANM benchmark (M1–M4; responsible: 4; involved: 1)** T5.1 will adapt the ANM benchmark available at http://www.montefiore.ulg.ac.be/~anm/ to the needs of the DELTA project. In particular, we will create the appropriate actions and parameters to create an environment that evolves over time.
T5.2	**Static RL in ANM (M5–M12; responsible: 4; involved: 1)** This task will consist in applying “classical” RL concepts to the ANM benchmark. So far, only methods from Mixed Integer Nonlinear Programming or metaheuristics have been applied to this problem. This first algorithm is termed “static” because it will not yet embed the research performed by other partners of the DELTA project.
T5.3	**Set up of the microgrid testing environment (M13–M18; responsible: 4)** As described in section 2.8, ULG has created a microgrid laboratory. This task will consist in developing the necessary interfaces to open the microgrid to the algorithms developed within the DELTA project.
T5.4	**Dynamic RL (M19–M30; responsible: 4; involved: all)** This task will apply the newly developed Reinforcement Learning algorithms to the ANM benchmark and Microgrid benchmarks. ULG will handle the development of the algorithms, and will coordinate with the other partners to define the specifications and ensure the integration is in line with the ideas developed in the other tasks.
T5.5	**Final validation (M31–M36; responsible: 4; involved: all)** Generation of the final results and reports based on the developments of tasks T5.1 to T5.4. ULG will handle the final validation, in collaboration with other partners.

Deliverable	Month of delivery	Title of deliverable
D5.1	5	Active Network Benchmark 2.0
D5.2	13	An implementation of static RL for the ANM benchmark, and a report/publication on the performance of static RL for the ANM benchmark
D5.3	19	An interface and a documentation for the microgrid benchmark.
D5.4	31	An implementation of the dynamic RL methods developed within the Delta project targeted to the two scenarios.
D5.5	36	A final report describing the results of the integration.

2.3 Management and Risk Assessment

The general purpose of the project management is progress control of each work package, coordination of the different project activities and implementation of quality control mechanisms. For achieving the above, the DELTA management structure comprises of the following entities:

1. The Project Management Board (PMB) will comprise WP leaders and be chaired by the Coordinator (UPF). PMB will convene at least every 3 months, via videoconference or, in case there is another consortium event running simultaneously, in person. PMB will be responsible for operational management of the consortium efforts, while the WP leaders comprising PMB will be in charge of the coordination, planning, monitoring and reporting on their respective WP.

2. The Steering Committee (SC) will comprise one representative from each beneficiary and be chaired by the Coordinator. SC will act as the main decision making body of the consortium, voting on critical issues defined in the CA. Mostly, however, its operation will involve monitoring of project progress, risk assessment and steering the project so as to proceed towards the pre-defined objectives. SC will meet physically twice a year at main consortium meetings. Extraordinary meetings may be called upon by any beneficiary raising major concerns regarding project implementation. These will usually be held as videoconferences.

3. External Advisory Board (EAB) that comprises experts in scientific and exploitation areas related to DELTA, will be established to enable independent advisory going beyond the vision of the consortium members, ensuring current development in wider field is reflected and work plan regularly reviewed to maximize the relevance and impact of the project.

Risk management: Risk register and list of deliverables and milestones will be monitored closely by the relevant WP leaders, presented at every PMB meeting and updated to represent work progress and potential new mitigation measures identified. Full review of risk register will take place at every meeting, where partner representatives will outline the general strategy to counter any potential drawbacks in project implementation. We will use risk management procedures based on Risk Issue Logs identifying tolerances and thresholds, and preparing contingencies. Although the consortium has already identified sources of potential risks and strategies to deal with them a Risk Assessment and Contingency Plan will be developed before month 12. The Coordinator will report on risk issues to the Project Management Board and update the Risk Assessment and Contingency Plan with each Periodic Activity Report.

List of milestones

Milestone	Delivery month	WP involved	Title
M1	12	WP5	Static RL in ANM
M2	12	WP4	Subtask utility metrics
M3	24	WP3	Open-ended exploration in changing environments
M4	31	WP5	Dynamic RL in ANM and microgrid management
M5	33	WP2	MCTS planning in changing environments

M6	36	WP5	Final application completed

2.4 Description of the Consortium

Partner 1	**Organisation name / Department**
Anders Jonsson	DTIC, Universitat Pompeu Fabra (UPF)

Expertise:

Universitat Pompeu Fabra (UPF) was established in 1990 as a public university with a strong dedication to excellence in research and teaching. It has been recently ranked 15th among the top 150 universities under 50 years old (THE2016) and 5th among young universities in the world that progress most quickly (THE2015, Young Universities Summit). UPF is also ranked 12th in Europe (1st Spanish) of a total of 625 universities (Multirank 2016) and 1st in Spain in teaching and research performance (U-Ranking, BBVA Foundation & Ivie, 2016). UPF takes part in this proposal through the Department of Information and Communication Technologies (DTIC). DTIC has an important track record of active participation in EU projects, including coordination (a total of 66 FP7 projects and 10 other projects in non-FP7, and, up to now, 23 H2020 projects). DTIC is the Spanish university department with the largest number of ERC grants (13), and is part of the FET Flagship initiative "The Human Brain Project". It is the only Spanish ICT department that has been awarded the "Maria de Maeztu" excellence by the Spanish government for the quality and relevance of its pioneering scientific research. DTIC will participate in this project through the AI and ML group.

Dr. Anders Jonsson received his Ph.D in 2005 from the University of Massachusetts Amherst, USA, working on hierarchical representations for reinforcement learning under the supervision of Prof. Andrew Barto. He has been a member of the AI and ML group at DTIC since, first as a post-doc, then as a tenure-track professor, and more recently as an interim tenured professor. His research interests involve sequential decision problems in general, formulated both as AI planning and reinforcement learning. He has worked extensively on hierarchical representations of learning and planning, and has authored more than 30 refereed publications in international conferences and journals. He has also been an investigator on several EU FP7 projects, including APIDIS and SpaceBook.

[1] D. Lotinac & A. Jonsson (2016). Constructing Hierarchical Task Models Using Invariance Analysis. ECAI'16.
[2] J. Segovia-Aguas, S. Jiménez & A. Jonsson (2016). Hierarchical Finite State Controllers for Generalized Planning. IJCAI'16 - Distinguished Paper Award.
[3] A. Jonsson & V. Gómez (2016). Hierarchical Linearly-Solvable Markov Decision Problems. ICAPS'16.
[4] J. Segovia-Aguas, S. Jiménez & A. Jonsson (2016). Generalized Planning With Procedural Domain Control Knowledge. ICAPS'16.
[5] C. Bäckström, A. Jonsson & P. Jonsson (2014). Automaton Plans. *Journal of Artificial Intelligence Research*, 51: 255-291.

Role in project:

UPF will be the leader and manager of the project, leading WP1. In terms of research, UPF will lead the work on task management (WP4). UPF will also contribute to the work on integrating the project modules into the two applications (WP5).

Partner 2	**Organisation name / Department**
Michal Valko	SequeL, INRIA

Expertise:

INRIA Lille is an Inria (the French National Institute for Research in Computer Science and Applied Mathematics) research center created in 2008. It includes around 200 researchers and contains 10 research teams. The research team SequeL (Sequential Learning) is composed of about 20 members (half of them being permanent) working in the field of machine learning, and more precisely on reinforcement learning, multi-armed bandit, statistical learning, and sequence prediction. INRIA Lille is a beneficiary site for the European Network of Excellence PASCAL2. SequeL has been involved in more than 5 national projects and in a number of international initiatives, including a Pump-Priming PASCAL2 project. Notably, from 2009 to 2012, SequeL coordinated the ANR Explo-RA project on the development of exploration-exploitation algorithms for efficient resource allocation. SequeL has also organized tutorials, workshops, and challenges in all the major conferences in machine learning in recent years. SequeL has been also a member of the EC-ICT project CompLACS, and it is manager of the work package responsible of investigating extensions of multi-armed bandit theory to large set of arms and its application to planning.

Dr. Michal Valko received his Ph.D. in machine learning from University of Pittsburgh, USA, in 2011 and habilitation from ENS Cachan, France, in 2016. He holds an experienced junior scientist position at INRIA Lille. Michal is primarily interested in designing algorithms for sequential problems with limited feedback, that would require as little human supervision as possible. He authored more than forty publications in machine learning and related fields, including 3 best paper awards, and regularly serves on program committees of the top-tier conferences and journals. The common thread of Michal's work has been adaptive (often graph-based) learning and its application to real-world applications. He applied his research in statistical techniques for conditional anomaly detection in University of Pittsburgh Medical Center hospitals to monitor the decisions of the physicians. Moreover his research in online semi-supervised learning resulted in an industry transfer to Intel Corp. as an online face recognizer which has been used in enhanced car experience, authentication for mobile devices, and personalized advertisement.

[1] Jean-Bastien Grill, Michal Valko, Rémi Munos: Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning, NIPS 2016
[2] Daniele Calandriello, Alessandro Lazaric, Michal Valko: Analysis of Nyström method with sequential ridge leverage scores, UAI 2016
[3] Tomáš Kocák, Gergely Neu, Michal Valko, Rémi Munos: Efficient Learning by Implicit Exploration in Bandit Problems with Side Observations, NIPS 2014
[4] Alexandra Carpentier, Michal Valko: Extreme Bandits, NIPS 2014
[5] Michal Valko, Alexandra Carpentier, Rémi Munos: Stochastic Simultaneous Optimistic Optimization, in ICML 2013

Role in project:
Inria will lead the effort on adaptive planning with a model that can adapt to changes. Inria will work with MUL on the hierarchical state partitioning.

Partner 3 Peter Auer	**Organisation name / Department** MUL

Expertise:

MUL is an engineering school with an excellent reputation for its research and teaching, with a commitment to both basic and applied research. The Chair for Information Technology (founded in 2003 and headed by Professor Peter Auer) does research in computational intelligence, machine learning, and its application to technical and cognitive processes, including incremental and online learning, reinforcement learning, and learning for cognitive vision. The Chair for Information Technology was partner in various joint projects, such as the EC-IST project LAVA (applying machine learning to computer vision problems), the EC-ICT PinView (on information and image retrieval with online machine learning techniques), and the PASCAL2 network of excellence. Recently the Chair of Information Technology was part of the consortium for the EC-ICT project CompLACS and worked on another project on “Structured and Continuous Reinforcement Learning” funded by the Austrian Science Fund (FWF).

Prof. Peter Auer received his PhD in mathematics from the Vienna University of Technology in 1992. In 1995 and 1996 he was research scholar at the University of California, Santa Cruz. In 1997 he was promoted to be associate professor at the Graz University of Technology, and he accepted the position of a full professor for Information Technology at the University of Leoben in 2003. He has authored or co-authored more than 50 refereed publications in scientific journals and conferences in the areas of probability theory, symbolic computation, and machine learning, and he is a member of the editorial board of the Machine Learning journal. He has been principal investigator in a number of European initiatives such as LAVA, PASCAL2, and CompLACS.

[1] Peter Auer, Chao-Kai Chiang: An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. COLT 2016: 116-120
[2] Ronald Ortner, Daniil Ryabko, Peter Auer, Rémi Munos: Regret Bounds for Restless Markov Bandits. Theoretical Computer Science 558, 62-76 (2014)
[3] Shiau Hong Lim, Peter Auer: Autonomous Exploration For Navigating In MDPs. COLT 2012: 40.1-40.24
[4] Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, Peter Stone: PAC Subset Selection in Stochastic Multi-armed Bandits. ICML 2012
[5] Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, Peter Auer: PAC-Bayesian Inequalities for Martingales. IEEE Transactions on Information Theory 58(12): 7086-7093 (2012)

Role in project:

MUL’s main role in the current project will be to investigate RL algorithms and autonomous exploration strategies for changing environments in WP3, in which MUL has the main responsibility and which is led by Peter Auer. MUL will also contribute to Tasks T2.3 and T2.4 in WP2 on planning with MCTS and use the outcome of these tasks to use hierarchical partitioning of the state space for more efficient exploration of the environment.

Partner 4	**Organisation name / Department**
Bertrand Cornélusse	Université de Liège (ULG)

Expertise:

The University of Liège (ULG) pursues an objective of excellence in training, research and innovation. It caters for more than 20,000 students and employs about 4,300 people including 2,800 teachers and researchers. The "Systems and Modelling" research unit of the department of Electrical Engineering and Computer Science consists of 10 faculty members and more than 50 researchers with diversified and interdisciplinary research interests. The Power Systems team's expertise lies primarily in the development of algorithmic solutions from mathematical programming and machine learning to solve practical problems of planning, operation and control of electrical systems in general and electrical distribution systems in particular. The team is active in several research and development projects, acting either as project coordinator or partner, related to the modernization of planning practices, exploitation, and control of electrical distribution systems.
Prof. Bertrand Cornélusse: Bertrand Cornélusse is Assistant Professor in Smart Microgrids at the University of Liège, Department of Electrical Engineering and Computer Science. His research interests are in optimization with a particular emphasis on operational planning of microgrids and distributions systems, as well as day-ahead electricity market coupling and demand side management. He was a consultant and software developer at N-SIDE, Louvain-la-Neuve (Belgium). He has published around 20 research papers and has experience in teaching, research, and consulting for industries. Prof. Damien Ernst: Damien Ernst is Full Professor in Smart Grids at the University of Liège, Department of Electrical Engineering and Computer Science. His research interests are in control theory and optimizations with a particular emphasis on power system control problems and reinforcement learning. His was a research fellow ate the University of Liège, an assistant professor at University of Rennes, and visiting scholar at MIT (USA) and ETH (Switzerland). He has published over 220 research papers and has co-authored two books and several book chapters.
Publications:

- Cornélusse, B., Vangulick, D., Glavic, M., & Ernst, D. (2015) "Global capacity announcement of electrical distribution systems: A pragmatic approach". Sustainable Energy, Grids and Networks 4 (2015), pp. 43–53.
- Cornélusse, B., Leroux, A., Glavic, M., & Ernst, D. (2015) "Graph matching for reconciling SCADA and GIS of a distribution network". In: Proceedings of CIRED, the International Conference on Electricity Distribution.
- Georges, E., Cornélusse, B., Ernst, D., Lemort, V., & Mathieu, S. (2017). "Residential heat pump as flexible load for direct control service with parametrized duration and rebound effect". Applied Energy, 187, 140-153.
- Gemine, Q., Cornélusse, B., Glavic, M., Fonteneau, R. & Ernst, D., (2016) "A Gaussian Mixture Approach to Model Stochastic Processes in Power Systems." In Proceedings of the 19th Power Systems Computation Conference.
- Gemine, Q., Ernst, D. & Cornélusse, B. (2016). "Active Network Management for Electrical Distribution Systems: Problem Formulation, Benchmark, and Approximate Solution." Optimization and Engineering, 1–43.

Role in project:

ULG will create the simulation environments and apply the methods developed by the other partners of the project. ULG will also help in the dissemination of the results, especially

towards bodies in the power systems' industry with whom the team has excellent relationships.

2.5 Added value of the collaboration, including multidisciplinarity and European dimension

UPF, INRIA and MUL are all among the top research groups in Europe working on reinforcement learning. In this project, their expertise is complementary: UPF brings expertise regarding task decomposition, INRIA brings expertise regarding planning, in the form of MCTS algorithms, and MUL brings expertise related to exploration, in the form of the UCRL algorithm. All of these areas are necessary components of lifelong RL, and it would be difficult for any of these groups to find suitable partners nationally for a project on lifelong RL.
The two applications related to electrical distribution networks provided by ULG has a complexity that goes far beyond the benchmarks that are commonly used to evaluate RL algorithms. Without such a level of complexity it would be difficult to evaluate the lifelong components of DELTA. In this sense, ULG is an essential partner that could not be easily substituted by a different partner in another European country. Conversely, ULG has long seeked to collaborate with research groups that perform cutting edge research in RL, in order to apply RL algorithms to adaptive network management (ANM). One expected impact of the project is that this collaboration will lead to an improvement in the state-of-the-art of ANM by building autonomous systems that achieve superior performance on ANM and related tasks.

2.6 Consortium agreement principles (partner's rights and duties, IPR management)

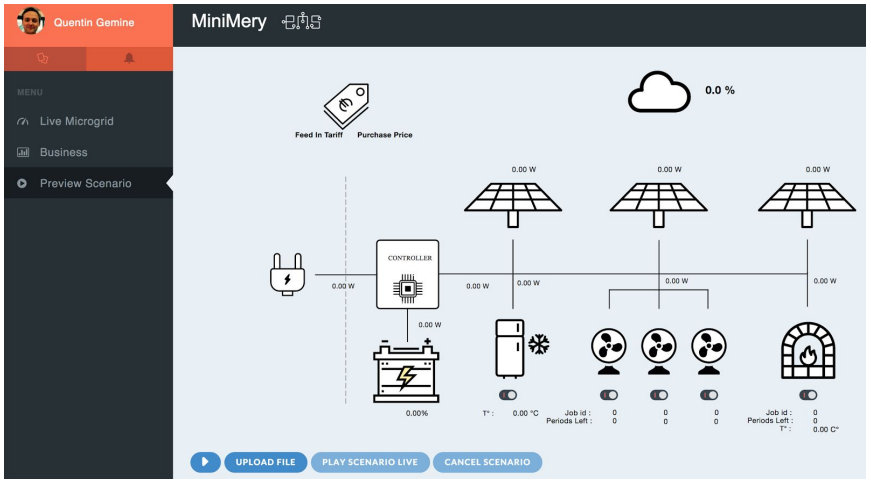
The roles, activities, responsibilities, ownership, commercial rights, confidentiality and IPR issues, as well as other legally binding commitments in the DELTA consortium will be collected and described in a Consortium Agreement (CA). This CA will be negotiated, and signed by each partner before the project starts to constitute a framework for operation of the partnership, defining management structure and procedures, transfers of EC payments, conflict resolution mechanisms, access to intellectual property and use of results. The CA will be based on the DESCA model as the most commonly applied framework, thus ensuring a considerable level of legal certainty.
In accordance to the general rules set out in the Grant Agreement signed with the European Commission, the CA also will contain provisions regarding IPR management and related issues such as dissemination, use and accessibility of the results, confidentiality provisions, arrangements on the settlement of disputes and liability. The Coordinator will lead the task of IPR Coordination to reflect the interests of all partners. In this sense, UPF has an Innovation Unit who advises on IPR matters the UPF research groups. Nevertheless, the IP regime will be governed by the Consortium Agreement, on terms similar to those already used by the parties in other national and European projects. It will ensure agreement at Consortium level, of which knowledge or inventions should be placed in the public domain as the basis for further exploitation and development (and the optimal ways to maximise it). Since IPR issues concern all partners, either as inventors or potential users, IPR and disclosure approvals will form a regular agenda item for Project Management Board decision.

2.7 Data management

Simulation data for the two scenarios will be obtained based on the Active Network Management and microgrid simulation environments. The Active Network Management benchmark will consist only of publicly available and standard distribution networks (such as IEEE networks) models, publicly available environmental data (such as wind speed, solar irradiance, etc), and publicly available consumption data. For the remaining data that we will have to generate for simulation purposes, ULG will do its best effort to select open data sources or to make the data generation process transparent. Regarding the microgrid simulation environment, ULG will also make its best effort to describe at best the equipment and to collect and make available trajectories of the system under several control policies.
Apart from simulation data, the other activities of the project do not directly generate data. In the case of the RL algorithms, they will instead be published as software modules and made available to other researchers so that they can reproduce the results of the project as well as apply the RL algorithms in other projects on lifelong learning.

2.8 Significant facilities and large equipment available to the consortium to perform the project

ULG has built a microgrid laboratory within the Electrical Engineering and Computer Science department (cf. the screenshot of the energy management system frontend below). This laboratory is a low power but full featured microgrid including photovoltaic generation under controlled lighting, a battery storage, the power electronics device that handles disconnection from and reconnection to the grid, and the charge of the battery. Three types of controllable loads are implemented: a fridge that can be switched off but has a temperature constraint, motors with flexible startup and shutdown times, and an oven where a set of jobs (a job is a temperature profile) must be scheduled. All the electrical consumptions, electrical productions, and temperatures are monitored in real-time through a PLC, polled by a backend and served to the front end and controllers.



2.9 Link with ongoing projects

UPF, INRIA and MUL were all involved in the recent CompLACS (Composing Learning for Artificial Cognitive Systems, http://www.complacs.org/) project funded by the European Community's Seventh Framework Programme (FP7/2007-2013), grant agreement 270327. Vicenç Gómez, who is currently a Ramon y Cajal fellow in the AI and ML group at UPF, worked on the development of two novel RL control methods based on Path Integral control. MUL worked on models for intrinsic motivation and curiosity driven exploration and developed the first algorithms for autonomously motivated exploration with theoretical guarantees, which will be the starting point for MUL's research on exploration in DELTA.
Another project involving MUL on "Structured and Continuous Reinforcement Learning" funded by the Austrian Science Fund (FWF project P 26219-N15) has recently ended in 2016 and an application for a successor project is currently under review. The main goal of the FWF project is to improve on the state-of-the-art in reinforcement learning with large or continuous state spaces by exploiting particular structural information. Although the proposed CHIST-ERA project and WP3 pursue different goals with the emphasis on exploration and changing environments, it would obviously benefit from progress concerning RL in large state spaces.
INRIA is in the 3rd year of ANR-funded JCJC project Extra-Learn (n.ANR-14-CE24-0010-01) and in the 1st year of ANR-funded JCJC project BADASS. The related parts of these projects deal with transfer learning (Extra-Learn) and non-stationarity (BADASS).
The research carried out by UPF in WP4 is related to a current project with the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) called "Solvers for General Artificial Intelligence with Applications" (SOLGAIA), TIN2015-67959. A secondary objective of that project is to develop a task management system similar to the one proposed in DELTA. Although both projects aim at decomposing tasks into subtasks, the context is different: in SOLGAIA sequential decision problems and subtasks are formulated in the formalism of AI planning, while in DELTA they are formulated as reinforcement learning.
ULG is currently involved in several microgrid and active network management projects (MeryGrid and GREDOR have the strongest connections with this project proposal):

- MeryGrid, construction of an industrial microgrid with a smart energy management system and an advanced battery + ultracapacitor storage system.
- InduStore, identification of the potentials and utilization of flexibility from industrial sites.
- GARPUR (FP7 – Energy), power system reliability with uncertainty and risk modelling.
- PREMASOL, integration of domestic PV into the electricity grid.
- GREDOR, electrical distribution system management with large share of renewables. Development of operational planning and real-time control algorithms.
- PEGASE (FP7 – Energy), Pan European grid analysis and state estimation.
- OMASES (FP5 – Energy), open market access and security assessment system.
- ExAMINE (FP5 – Energy), monitoring and control of European infrastructure of electrical power exchanges.

In all of these projects, ULG develops algorithms and simulation environments, relying on

1. machine learning for building forecasting models that can be used to sample scenarios of random variables of the system
2. optimization techniques, such as mixed integer nonlinear programming, to solve the resulting stochastic optimization problems.

The group does research in RL though it was never applied to the two scenarios proposed. The group however believes this are promising techniques that are worth being assessed.

2.10 Financial plan

The requested EU contribution results from the accurate estimation that each beneficiary has done following the funding criteria and regulations of its Research Funding Organisation. Therefore the project will take 36 months to achieve its objectives and has a total requested budget of €668.436. Due to the nature of the project, the major costs of the resources committed for the project are related to personnel costs (93% of the total costs). In the case of overhead costs they are covered by the own partners' resources since they are, with the exception of MUL, non eligible expenses. Other direct costs are considered minor and limited to 5% of the requested budget.

The explanation for the categories with a significant relevance of the requested budget is:

Personnel costs reflect the cost of the 147 person months to be spent in the project (WP1-WP5). This cost is based on a realistic plan of the time and level of expertise needed to carry out the tasks described in the proposal. Furthermore, the costs of the personnel involved in the project implementation have been calculated based on average monthly salary costs at each partner organization.

Equipment: Only requested by ULG to fund the replacement of available computers.

Travel cost: [Whole project €24.702] The consortium will meet physically at least once per year in order to closely follow up and build a common understanding of the work plan, as well as the progress of the tasks and resources of the project. These regular meetings will include Supervisory Board meetings and working meetings for WP groups, hosted by one of the project partners to economise on travel costs. In this sense the beneficiaries have a travel budget to ensure adequate representation at the following 6 meetings:
- Kick-off meeting (1) to launch the project and refine plans and arrangements for the initial implementation phase. Barcelona (Spain). Hosted by UPF as coordinator.
- Progress meeting (2) to review progress and discuss any significant problems and deviations, if necessary. 1 in Barcelona (Spain) hosted by UPF and 1 in Leoben hosted by MUL.
- Review meeting (3) to evaluate intermediate and final results. To assess quality, impact and effectiveness of project work. 3 (one per year) in Brussels (Belgium) or suitable project site to be decided in agreement with the Project Officer.
The consortium shall convene as necessary extraordinary working meetings (intra-WP, cross-WP, SB) using remote meeting technologies if circumstances allow it.

Other/Conferences and Publication Fees: [€7.400] Over the 3 years of the project, DELTA plans a total number of publications in major international conferences at least equal to 12 (3 in the first year, 4 in year 2, and 5 in year 3). In addition, some partners plan to publish journal papers in open access journals for a cost of ~1250€ each. The requested budget should also cover registration at conferences as well as accommodation.

Overhead: Only requested by MUL as a compulsory 5% of the direct costs.

3. Impact

3.1 Dissemination and Exploitation of Results

Dissemination will be achieved as follows: (1) Publications in high-impact international journals and conferences (ICML, COLT, NIPS, AISTATS, PSCC). (2) Project demos presented at public and corporate events related to autonomous systems. (3) Workshops. The consortium will organise at least one workshop on lifelong RL. All dissemination activities will be regularly monitored by the coordinator.
Communication with a wider audience beyond the consortium will be achieved as follows: (1) The *project web portal* will provide easy-to-use access to information on a popular science level, links to project partners, descriptions of research teams, a description of the project itself, the project's objectives, innovation and scientific publications arising from the project. Furthermore, DELTA events, tutorials and workshops will be announced on the website. (2) *Business-oriented media*. The project's prototypes and application are of potential interest to industry. We will communicate results through TV, magazines and exhibitions. (3) *Wikipedia*. We will provide a comprehensive description of the DELTA concept in Wikipedia and link it to other concepts, projects, and supporting multimedia content. (4) *Liaisons with other projects and consortia*. Liaison with other projects is a means to co-ordinate the activities of the DELTA project in the context of ongoing activities in other projects. In particular, we envisage strong interactions with CENTAURO, ROBUST, SHRINE and others.
Exploitation will be mainly achieved by ULG, who will handle the development and publication of the interface to the microgrid, and will coordinate with UPF to define the specifications and ensure the integration is in line with the ideas developed in the other tasks. ULG will also organize events with industrial partners, such as electricity distribution system operators for ANM, and events in their microgrid laboratory facility, to promote the research and its applications.

3.2 Expected Impacts

The use of machine learning (ML) is continuously changing the world around us in a myriad of ways. It is transforming our lives visibly with novelties such as self-driving cars, care robots, digital assistants and the omnipresent marketing robots, and less conspicuously by streamlining and optimizing all facets of our lives behind the scenes. Everything from what we eat to the news we read is now guided by ML algorithms. All of this is made possible by the use of ML that is focused on one narrow task at a time. While several tasks in parallel coupled with a human like attributes such as for example the voice of Siri can give the user the appearance of something more, it does not have the ability to do much outside of its limited pre-defined range.
The DELTA project proposes a possible route to step beyond the current ML paradigm, in developing the methods to allow ML systems to autonomously expand on their problem solving capability. The impact of dynamically evolving long-term autonomy cannot be overstated. It would dramatically improve the capabilities of ML, adding completely new uses for a technology that already underpins most of modern society. The effects the DELTA project, if successful, will further the state of the art of ML in a way that will alter the field profoundly, expanding the capabilities and areas of use for ML considerably.
The DELTA project has the potential to profoundly transform ICT and any industry or scientific discipline that depends on it. Instead of painstakingly implementing a specialised

autonomous agent for each specific application, one can design more general problem solving agents and that can adapt dynamically to their environment, finding the best problem solving strategy independently. This will not only cut costs, but the strategies found by autonomous agents have the potential to be better than any hardcoded strategy.
Examples of the uses of dynamic, self-learning ML are numerous, but to name a few:

- A self-learning control system of large scale operations and infrastructure, such as power plants, that can handle previously unknown situations not predicted by the system designers. This is of specific importance regarding low probability situations with catastrophic consequences.
- Robots working in unknown environments (be it in rescue zones, factories or our homes) are likely to encounter problems unforeseen by the manufacturers. It is of vital importance that such a robot can elaborate new problem-solving strategies on the fly.
- Market crashes are often exacerbated by hardcoded trading software (leading to so-called ``flash crashes''). An automatic trader with dynamic problem-solving capabilities can react more flexibly to these uncommon but very costly scenarios.
- Natural language systems can learn new vocabulary and dialogue patterns, leading to more flexible interfaces, automated personal assistants acting independently in the user's best interest, etc.
- In the future, numerous interconnected smart devices will assist users in all walks of daily life, and to function properly these devices have to adapt quickly to unforeseen situations.

Active Network Management could be applied to all distribution networks in countries that are willing to integrate more renewable energy and/or that have changing consumption modes, such as electric vehicles or heat pumps, that will highly impact the flows in these networks. Hence it is a very large potential market. The main challenges for now is to develop tools that are adapting well to evolving conditions (e.g. weather conditions, seasonality). Adaptiveness is kind of an enabling factor.
Microgrids do naturally develop either where there is no existing distribution system (e.g. in Africa), where adaptiveness is also a key factor because it is by definition a small system and adding or removing a device can change a lot the way the system should be controlled, or on the other hand where distribution network exist but high reliability is essential (e.g. a hospital, a military base, etc.).

4. Ethical issues

The partners do not foresee any ethical issues related to the project.

5. References

[1] A. Barto & R. Sutton (1998). *Reinforcement Learning: An Introduction*. MIT Press.
[2] L. Kocsis & C. Szepesvári (2006). Bandit based Monte-Carlo Planning. *Proceedings of the European Conference on Machine Learning (ECML'06)*.
[3] R. Coulom (2007). Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search. *Proceedings of the International Conference on Computers and Games (CG'06)*.
[4] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou & V. Panneershelvam. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529 (7587): 484-489.

[5] S. Bubeck, R. Munos, G. Stoltz & C. Szepesvári (2011). X-armed bandits. *Journal of Machine Learning Research*, 12: 1587-1627.
[6] M. Valko, A. Carpentier & Rémi Munos (2013). Stochastic simultaneous optimistic optimization. *Proceedings of the International Conference on Machine Learning (ICML'13)*.
[7] T. Jaksch, R. Ortner & P. Auer. Near-optimal Regret Bounds for Reinforcement Learning. Journal of Machine Learning Research 11: 1563-1600 (2010)
[8] I. Osband & B. Van Roy (2014). Near-optimal Reinforcement Learning in Factored MDPs. *Advances in Neural Information Processing Systems (NIPS'14)*.
[9] B. Hengst (2002). Discovering Hierarchy in Reinforcement Learning with HEXQ. *Proceedings of the International Conference on Machine Learning (ICML'02)*.
[10] S. Singh, A. Barto & N. Chentanez (2005). Intrinsically Motivated Reinforcement Learning. *Advances in Neural Information Processing Systems (NIPS'05)*.
[11] A. Jonsson & A. Barto (2006). Causal Graph Based Decomposition of Factored MDPs. *Journal of Machine Learning Research*, 7: 2259-2301.
[12] M. Gheshlaghi-Azar, A. Lazaric & E. Brunskill (2013). Regret Bounds for Reinforcement Learning with Policy Advice. *Proceedings of the European Conference on Machine Learning (ECML'13)*.
[13] E. Brunskill & L. Li (2014). PAC-inspired Option Discovery in Lifelong Reinforcement Learning. *Proceedings of the International Conference on Machine Learning (ICML'14)*.
[14] H. Bou Ammar, R. Tutunov & E. Eaton (2015). Safe Policy Search for Lifelong Reinforcement Learning with Sublinear Regret. *Proceedings of the International Conference on Machine Learning (ICML'15)*.
[15] Q. Gemine, D. Ernst & B. Cornelusse (2016). Active network management for electrical distribution systems: problem formulation, benchmark, and approximate solution. *Optimization and Engineering*.
[16] M. Ghavamzadeh, S. Mannor, J. Pineau & A. Tamar. Bayesian Reinforcement Learning: A Survey. *Foundations and Trends in Machine Learning*, 8 (5-6), 359-492.
[17] C. Diuk, A. Cohen & M. Littman (2008). An Object-Oriented Representation for Efficient Reinforcement Learning. *Proceedings of the International Conference on Machine Learning (ICML'08)*.
[18] M. van Otterlo (2012). Solving Relational and First-Order Markov Decision Processes: A Survey. *Reinforcement Learning: State-of-the-art*, 253-292.
[19] S. Kalyanakrishnan, A. Tewari, P. Auer & P. Stone (2012). PAC subset selection in stochastic multi-armed bandits. *Proceedings of the International Conference on Machine Learning (ICML'12)*.
[20] A. Garivier & E. Kaufmann (2016). Optimal best arm identification with fixed confidence. *Proceedings of the Conference On Learning Theory (COLT'16)*.
[21] B. Szörényi, G. Kedenburg, R. Munos: Optimistic Planning in Markov Decision Processes Using a Generative Model. in *Neural Information Processing Systems (NIPS 2014),*
[22] J-B. Grill, M. Valko, R. Munos: Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning, in *Neural Information Processing Systems (NIPS 2016)*
[23] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite time analysis of the multiarmed bandit problem. *Machine learning, 47 2/3:235 – 256, 2002.*
[24] S. Bubeck, R. Munos and G. Stoltz, Pure Exploration in Finitely-Armed and Continuously-Armed Bandits. *Theoretical Computer Science 412, 1832-1852, 2011*
[25] *Maximin Action Identification: a New Bandit Framework for Games, Aurélien Garivier, Emilie Kaufmann and Wouter M. Koolen. in COLT, 2016*
[26] Shiau Hong Lim, Peter Auer: Autonomous Exploration For Navigating In MDPs.

COLT 2012: 40.1-40.24