# Ph.D. in Information Technology
# Thesis Defenses

**July 14th, 2023**
**at 14:00**
**Aula Seminari Alessandra Alario, Ed.21 and online by Webex**


## **Pierre Etienne Valentin LIOTET** – XXXV Cycle


## **Delays in reinforcement learning**


Supervisor: Prof. **Marcello Restelli**



**Abstract:**

Delays are inherent to most dynamical systems. Besides shifting the process in time, they can drastically impact their performance. For this reason, it is usually valuable to study the delay and account for it. Because they are dynamical systems, it is of no surprise that sequential decision-making problems such as Markov decision processes (MDPs) can also be affected by delays. The latter processes are the foundational framework of reinforcement learning (RL), a paradigm whose goal is to create artificial agents capable of learning to maximise their utility by interacting with their environment.

RL has achieved strong, sometimes mind-blowing, empirical results, but delays are seldom explicitly accounted for. The understanding of the impact of delay on the MDP is limited. In this dissertation, we propose to study the delay in the agent's observation of the state of the environment or in the execution of the agent's actions. We will repeatedly change our point of view on the problem to reveal some of its structure and peculiarities. A wide spectrum of delays will be considered, and potential solutions will be presented. This dissertation also aims to draw links between celebrated frameworks of the RL literature and the one of delays. We will therefore focus on the following four points of view.

At first, we consider constant delays. Taking a psychology-inspired approach, we study the impact of predicting the near future in order to estimate the impact of the agent's actions. We will highlight how this approach relates to models from the RL literature. It will also be the occasion to formally demonstrate a seemingly evident fact: longer delays involve lower performances. The experimental analysis will conclude by showing the validity of the approach.

As a second point of view, we will consider the simple approach of imitating an undelayed expert behaviour in the delayed environment. The delay will remain constant at first, but we will extend our study to more exotic types of delay, such as stochastic ones. Although simple, we demonstrate the great theoretical guarantees and empirical results of the approach.

Changing for a third time our point of view on constant delay, we consider adopting a non-stationary memoryless behaviour. Although it seemingly ignores the delay, the approach treats the delay's effect as an unobserved variable that guides its non-stationarity. Building on this idea, we provide a theoretically grounded algorithm for learning such behaviour that we test in realistic scenarios.

Finally, our last point of view will consider a broader model than that of constant delay, which includes constant delays as a special case. This model will enable actions to affect multiple future transitions of the environment. Its theoretical properties will be examined to understand its specificities. Based on these properties, some RL algorithms will be ruled out, while others will be tested in various empirical studies. Being a more general model for delays, its understanding has implications for the constant delay frameworks of the previous chapters.

**Luca SABBIONI** – XXXV Cycle

## Exploiting hyperparameter optimization and control frequency in reinforcement learning

Supervisor: Prof. **Marcello Restelli**

**Abstract:**

Reinforcement Learning (RL) has driven impressive advances in artificial intelligence in recent years for a wide range of domains, from robotic control to financial trading.
However, the performance of current RL methods is strongly dependent on the hyperparameters of the algorithms, which practitioners usually need to tune carefully, and on the environment design, where the control frequency plays a dominant role. The consequent engineering procedures are prone to errors and are time-consuming, especially if they are started from scratch for each task modification.
The subject of this dissertation is the development of automatic techniques to enhance the learning capabilities of RL algorithms in a twofold direction.

In the first part, we address the Hyperparameter Optimization (HO) problem, with a particular focus on policy-based techniques for RL: indeed, they rely on strong theoretical guarantees that play a very important role but do not help in the selection of the hyperparameters. To enhance the learning capabilities of this class of algorithms, we frame HO as a Sequential Decision Process and design a solution that allows selecting a dynamic sequence of hyperparameters adaptive to the policy and the context of the environment. Hence, the reward function of the learning process is performance gain, and the action consists in the hyperparameter selection. With this problem definition, it is possible to adopt RL algorithms on a more abstract level to optimize the progress of the whole learning instance.

The second part is devoted to improving RL agents by leveraging the frequency of the agent-environment interaction, which has a deep impact on the control opportunities and the sample complexity of the learning algorithms. We introduce and discuss the concept of action persistence or action repetition: leveraging theoretical results and bounds on the performance loss incurred while employing persistence, we provide algorithmic contributions to detect the most promising frequency. As a conclusive contribution, we employ a new operator that allows for effective use of the experience collected at any time scale to learn a dynamic adaption of the persistence or, in other terms, the best duration of each action.

**PhD Committee**
Prof. **Francesco Trovò,** Politecnico di Milano
Prof. **Samuele Tosatto**, Department of Computer Science And Digital Science Center- Universität Innsbruck
Prof. **Carmine Ventre**, King's College London