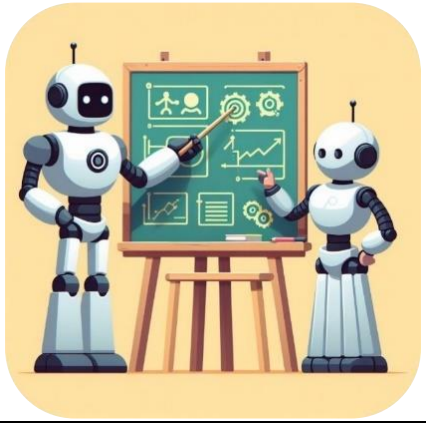
SSM
Scuola Superiore Meridionale

# MERC PhD Project Proposal 2023/2024

## Title of the research project

**Controlling Reinforcement Learning:**
**Control Theoretical Strategies and Methods to Enhance Reinforcement Learning**



## Keywords

Reinforcement learning; Control theory; Data-efficiency; Formal guarantees

## Supervisors

*Supervisor 1:* **Marco Coraggio** (marco.coraggio@unina.it)
http://marco-coraggio.com
Expertise: Control theory, Complex networks, Data-driven control

*Supervisor 2:* **Mario di Bernardo** (mario.dibernardo@unina.it)
https://sites.google.com/site/dibernardogroup
Expertise: Control theory, Complex systems, Dynamical systems

*Supervisor 3:* **Giovanni Russo** (giovarusso@unisa.it)
Dept. of Information and Electrical Engineering & Applied Mathematics at University of Salerno, Italy
www.sites.google.com/view/giovanni-russo
Expertise: Theory of decision-making, Data-driven systems, Control Theory, Complex Cyber-physical systems, Network systems

## Project description

***Introduction***

*Context and motivation.* *Reinforcement learning* (RL) has proven remarkably successful in many challenging applications, ranging from predicting a protein's 3D structure [Jumper, 2021] and controlling plasma in fusion reactions [Degrave, 2022], to developing effective behavior for collaborative and competitive games, such as 2-player sports, Go, and modern videogames (e.g., [Won, 2021]). RL has even been recently used to fine-tune GPT 3.5 and 4. Nonetheless, several notable limitations still affect these methodologies. Namely, (i) very long execution times are required when the state and action spaces are large—preventing wider accessibility of the technology—and (ii) guarantees on the learned

policies are typically missing—making RL unsuited for critical and/or high-performance operations.

In this project, we aim to exploit methodologies and solutions from control theory to overcome these limitations.

*State of the art.* Recent research has shown that combining components from control theory and reinforcement learning can yield effective control strategies. For instance, in [Zanon, 2021], reinforcement learning algorithms have been used to vary the parameters of the model and of the objective function in model predictive control, whereas in [Gu, 2016] local linear models fitted iteratively with exploration data were used to accelerate learning. Notably, in [De Lellis, 2023a], the so-called *control tutor* assumes that a (even imperfect) feedback control law is available for the task at hand and uses it to occasionally suggest actions during exploration in the learning process. This was demonstrated to reduce learning time, both with tabular and deep value-based learning algorithms. Regardless, several important questions remain. Namely, (i) is the benefit consistent when applied to policy-based learning algorithms? (ii) In a multi-agent scenario, can multiple control tutors coexist, or will they conflict? (iii) How can this benefit be quantified and proven formally? Additionally, in [De Lellis, 2023b] a *reward shaping* method was applied to provide guarantees of performance on the learned policy, by assessing the value of the cumulative reward function. Such guarantees are crucial, for instance, with safety-critical systems. The drawback of the method is that the shaped reward is made *sparse*, which can hinder learning the optimal policy, especially with deep reinforcement learning.

**Objectives**

*The workplan and objectives are flexible and will be adapted depending on the inclination of the student and the results obtained in the early phases of the project.*

O1. Develop and validate formally control-theoretical based methodologies to increase data efficiency in reinforcement learning.
O2. Develop practical control-theoretical based methodologies that provide stability and performance guarantees on the learned policy in reinforcement learning.
O3. Validate the developed strategies on the problem of agent navigation.

**Methodology**

The project will start with a thorough classification of available methods combining reinforcement learning and control theoretical tools, detailing advantages, disadvantages, and open problems in each.

Next, the project will focus on achieving Objective O1, to provide an analytical proof of validity of the increased data-efficiency yield by control tutors [De Lellis, 2023a] (already successfully validated numerically in numerous scenarios), for a tabular value-based RL algorithm. To do this, we will first determine how a uniform probability distribution is altered by the suggestions of a control tutor; then, we will extend the classical proof of convergence of the *Q*-learning algorithm, in the case of simple discrete Markov decision processes, assuming the use of this altered exploration probability distribution, rather than a uniform one. This analysis also has the purpose of quantifying the relation between the quality of the control tutor (in terms, e.g., of the optimality of its suggestions) and the decrease in learning time. Subsequently, we will investigate the multi-agent case study of heterogeneous robotic agents coordinating to explore an unstructured environment, coping with different specializations of the robots (e.g., land/air unit, fast/slow, etc.); the aim will be to assess whether the benefit observed in a single tutor scenario immediately extends to one where multiple tutors are present, or if different schemes are necessary, e.g., with a leader tutor coordinating the others, by altering their suggestions.

To achieve Objective O2, research will start from the reward shaping method presented in [De Lellis, 2023b], which, although promising, suffers from sparse shaped reward functions, in the sense that its values can change significantly in the state-action space, which makes learning difficult when approximators (e.g., neural networks) are used to learn the value function. To overcome this problem, the PhD student will experiment with modified reward shaping procedures, focusing on those that alter the reward through smooth functions, rather than the discrete one used in [De Lellis, 2023b]. A starting point for these shaping functions will be energy-like potential functions. Moreover, we will consider also the possibility of providing both deterministic guarantees (using the framework in [De Lellis, 2023b]) and probabilistic ones.

Finally, in the context of Objective O3, the combination will be investigated numerically of different techniques developed in the project, to obtain *both* reduced learning time and guarantees of performance. The best performing algorithms developed in the project will be validated on the task, for a ground mobile robot, of reaching a specified region, while avoiding unsafe regions, mimicking vehicle driving or robot navigation on an extraterrestrial planet.

## Relevance to the MERC PhD Program

***Relevance and beneficiaries***

The project has a strong methodological component. The integration of control theoretical methodologies with reinforcement learning techniques is prospected to deliver two main advantages: (i) the reduction of learning time and (ii) the certification of properties on the learned policy. The former benefit will contribute to the democratization of reinforcement learning techniques, enabling the resolution of particularly complex problems even without the need for advanced supercomputers. The latter benefit will contribute to the process of allowing the use of RL in critical applications such as those involving agent navigation (e.g., automated driving, search and rescue, extraterrestrial exploration).

***Relevance to the MERC PhD program***

The project is highly interdisciplinary, combining dynamical systems and control theory with machine learning to develop novel practical and methodological solutions to challenging complex control problems.

*Skills.* During the project, both tutored by the supervisors and through self-study, the student will develop skills in several fields, including:

- Machine learning algorithms and methods, with a particular emphasis on reinforcement learning,
- Dynamical systems and Markov decision processes, with a focus on stability analysis,
- Advanced computer programming, including state-of-the-art machine learning libraries in modern programming languages.

Additionally, tutoring will also focus on sharpening the student's technical writing and presentation skills, and developing the ability to study scientific literature swiftly and effectively.

## References

***Key references***


- F. De Lellis, M. Coraggio, G. Russo, M. Musolesi, M. di Bernardo, "CT-DQN: control-tutored deep reinforcement learning," in Proceedings of the 5th Annual Learning for Dynamics and Control Conference, PMLR, pp. 941–953, 2023a,

- F. De Lellis, M. Coraggio, G. Russo, M. Musolesi, M. di Bernardo, “Guaranteeing control requirements via reward shaping in reinforcement learning.” arXiv.2311.10026, 2023b.
- S. Gu, T. Lillicrap, I. Sutskever, S. Levine, “Continuous deep Q-learning with model-based acceleration,” in International Conference on Machine Learning (ICML’16), pp. 2829–2838, 2016.
- M. Zanon and S. Gros, “Safe reinforcement learning using robust MPC,” IEEE Transactions on Automatic Control, 66(8):3638–3652, 2021.


***Additional references***


- J. Degrave et al., “Magnetic control of tokamak plasmas through deep reinforcement learning,” Nature, 602(7897):414–419, 2022.
- J. Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” Nature, 596(7873), 2021.
- J. Won, D. Gopinath, and J. Hodgins, “Control strategies for physically simulated characters performing two-player competitive sports,” ACM Transaction on Graphics, 40(4):146:1–146:11, 2021.


## Joint supervision arrangements

The student will meet at least weekly with at least one of the supervisors. The whole team will meet at least once every 1 or 2 months for a progress update.

## Location and length of the study period abroad (min 12 months)

The student will be able to spend a research period (or research periods) at the lab of a member of the board or of a scientist with whom a collaboration is active.