Enhancing Safety and Explainability of Reinforcement Learning Agents for Environmental Monitoring Tasks

Luca Marzari[1], Francesco Trotti[2], Francesco Dal Santo[1], Amirhossein Zhalehmehrabi[1], Celeste Veronese[1], Davide Villaboni[1], Federico Bianchi[1], Daniele Meli[1], Alberto Castellini[1] and Alessandro Farinelli[1]

[1]*Department of Computer Science, University of Verona, Verona, Italy*
[2]*Department of Engineering for Innovation Medicine, University of Verona, Verona, Italy*

Abstract

Mitigating pollution in aquatic ecosystems is among the most pressing challenges in environmental sustainability applications. While effective monitoring and intervention activities are key to safeguarding water quality, protecting biodiversity, and supporting industries (e.g., aquaculture), this is traditionally done by human operators—making the process costly, time-consuming, and often inadequate for capturing timely environmental changes. In this work, we focus on safe, explainable design and deployment of autonomous reinforcement learning (RL) agents for environmental monitoring tasks. In particular, we present our recent contributions to: i) safe RL techniques, ii) Neurosymbolic RL, iii) formal verification of deep RL policies, and iv) designing robust control strategies for safe deployment.

Keywords

Safe Reinforcement Learning, Formal Verification of Neural Networks, Explainable and Neurosymbolic AI, Safe Deployment

1. Introduction

Supporting environmental sustainability efforts, such as pollution monitoring and protecting biodiversity in aquatic ecosystems [1], is among the most pressing challenges in environmental sustainability applications, also highlighted by the United Nations in *Goal 14: Conserve and sustainably use the oceans, seas and marine resources* of the Agenda 2030 (https://www.un.org/sustainabledevelopment/oceans/).

Traditionally, these tasks have depended on manual human intervention, which is often costly and unable to adapt effectively to real-time environmental changes. Autonomous aquatic robots offer a promising solution to these limitations by enhancing monitoring capabilities and assisting human decision-making [2]. In these contexts, safety is paramount, as autonomous robots often involve expensive hardware and human interactions. Specifically, robots' decisions (e.g., why is the drone sampling from a certain location, or how is it adapting to environmental changes) must be understandable by human operators, requiring the integration of novel explainable artificial intelligence (AI) methods.

In this work, we summarize our recent advancement in developing a safe and explainable autonomous reinforcement learning (RL) agent for environmental monitoring tasks. Specifically, we first present the problem formulation, highlighting the main challenges in solving the task. We then introduce our novel solutions for safe autonomous navigation, exploiting both verification-inspired, neurosymbolic-AI, and, in general, machine learning techniques (i.e., Gaussian process) for enhanced safe exploration and explainability of RL policies in realistic aquatic scenarios. Importantly, our focus is on deploying safe and reliable autonomous agents, and to this end, we show how we can exploit recent advancements in verification of neural networks to identify unsafe regions of operation, which are then used to construct a safe control barrier function (CBF)-based control layer applicable to arbitrary policies. Our

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23–24, 2025, Trieste, Italy
luca.marzari@univr.it (L. Marzari); francesco.trotti@univr.it (F. Trotti); francesco.dalsanto@studenti.univr.it (F. D. Santo); amirhossein.zhalehmehrabi@univr.it (A. Zhalehmehrabi); celeste.veronese@univr.it (C. Veronese); davide.villaboni@univr.it (D. Villaboni); federico.bianchi@univr.it (F. Bianchi); daniele.meli@univr.it (D. Meli); alberto.castellini@univr.it (A. Castellini); alessandro.farinelli@univr.it (A. Farinelli)

results demonstrate that the proposed methods enable safe and scalable autonomous robotic navigation, achieving zero violations of safety restrictions at deployment while increasing navigation effectiveness.

2. Environmental Monitoring

In this section, we introduce the problem formulation we aim to address. In detail, we consider an aquatic drone that has to navigate in an unknown environment using only sensor measurements. Formally, we define the aquatic drone state and control vector (Fig. 1 (left)) following the state-of-the-art formalism [3] as $\boldsymbol{x} = [v_1 \quad v_2 \quad v_3 \quad \omega_1 \quad \omega_2 \quad \omega_3 \quad p_x \quad p_y \quad p_z \quad \phi \quad \theta \quad \psi]^T, \boldsymbol{u} = [\delta_l \quad \delta_r]$.

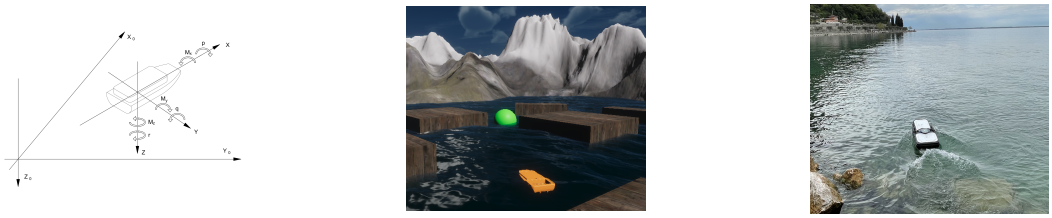


Figure 1: On the left an illustrative representation of the dynamics components of an aquatic drone. In the center the realistic simulator used for the mapless navigation tasks. Finally, on the right, a deploy of our solution for a real-world environmental monitoring task.

Hence, we formalized the aquatic mapless navigation task as a Markov decision process (MDP). Hence, the autonomous agent aims to maximize the expected discounted return for each trajectory. In this context, online model-based learning methods embedding hierarchical control architectures have been used to deal with uncertainty in various navigation tasks [4, 5, 6]. Nevertheless, these methods rely on extensive domain knowledge, which could be difficult to obtain in practice. To address this issue, deep reinforcement learning (DRL) algorithms have been employed to learn how to navigate in complex and unknown (i.e., mapless) environments [7]. However, DRL-based policies are modeled as deep neural networks (DNNs), which are known to be vulnerable to adversarial inputs—small perturbations to the input state leading to unexpected and unsafe actions [8]. To address safety, MDPs are typically extended to constrained MDPs (CMDPs) [9] and incorporate a set of constraints during the training using one or more cost functions to encode safety desiderata [10, 11]. Hence, constrained DRL algorithms aim at maximizing the expected return while maintaining costs under hard-coded thresholds. However, CMDPs present non-negligible limitations, such as the necessity of a parametrization of the policy (i.e., they are not applicable to value-based methods) [10]. To address this issue, in the next sections, we introduce novel RL techniques to enhance the safety aspect of learning agents.

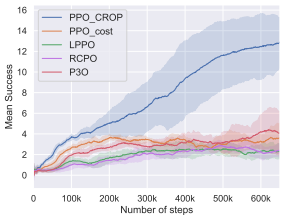
3. Enhancing Reinforcement Learning Techniques

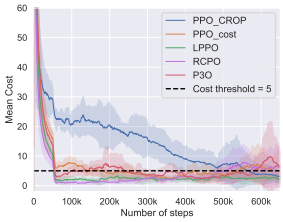
Methods for enhancing exploration, safety, and explainability of RL techniques are summarized here.

3.1. Online Collection and Verification of Safety Property

To enhance safety during training in [12], we propose a novel approach that collects and refines, during the training, safety properties (also referred to as retrain areas), i.e., part of the state space where the agent is prone to violate safety desiderata. We then investigate the use of these online safety properties in two ways: i) in [13], after each collision, we introduce an approximate verification approach by sampling and propagating states from the safety properties that contain the unsafe state, estimating the probability of encountering similar unsafe behaviors. This probability, called the *violation value*, is then used to penalize the agent and discourage risky actions. ii) In [14], inspired by human learning, where consistently repeating tasks enhances the learning of a particular behavior, we investigate the impact of switching between the typical uniform restart state distribution and the retrain areas using a decaying factor ε, allowing agents to retrain on situations where they violated the safety desiderata.

Importantly, our experiments (in part presented in Fig. 2) over hundreds of seeds across mapless navigation, locomotion, and power network tasks show that the proposed approaches yield agents that exhibit significant improvements in both safety performance and sample efficiency over existing safety-oriented DRL methods. Moreover, in [14] we demonstrate a monotonic improvement in policy optimization when using retrain areas, indicating that revisiting unsafe regions consistently helps the agent refine its policy toward safer and more effective behaviors throughout training.





	Mean Violation		
	100k	**300k**	**650k**
*PPO_CROP	6.7 ± 1.0	4.8 ± 1.8	$\mathbf{2.2 \pm 1.2}$
LPPO	1.1 ± 0.5	2.1 ± 1.4	3.5 ± 1.2
RCPO	$\mathbf{0.37 \pm 0.23}$	$\mathbf{1.66 \pm 0.97}$	4.6 ± 5.3
PPO_cost	3.7 ± 1.0	2.5 ± 0.4	6.7 ± 5.8
P3O	2.4 ± 1.1	2.2 ± 0.8	5.8 ± 3.9

Figure 2: Results of the approaches proposed in [13, 12] comparing PPO_CROP (our novel approach), PPO_cost (PPO with a fixed penalty upon each collision), RCPO [15], LPPO[10], and P3O[16] in a mapless navigation task. The table on the right reports the mean *violation value* at three different global steps for each method tested.

3.2. Neurosymbolic RL

Classical data-driven RL can learn optimal policies for complex MDP tasks under uncertain perception and environmental modeling. However, it is known to be sample-inefficient, i.e., data and computationally intensive during training; moreover, RL policies are hardly generalizable to unseen scenarios, which significantly hinders the sim-to-real transfer and the application to real environmental monitoring [17]. Neurosymbolic (NeSy) RL aims to mitigate these limitations by leveraging the advantages of symbolic artificial intelligence in terms of abstraction and generalization capabilities for planning. Specifically, starting from some available background knowledge about the domain (e.g., high-level plan specifications), it is possible to express it as a set of logical rules and perform automated reasoning to guide RL training [18]. However, existing approaches are based on reward shaping, which is inherently sample-inefficient [19]. Inspired by the "thinking fast and slow" theory of mind [20], we propose instead a novel NeSy approach, where symbolic reasoning is tightly integrated in the RL algorithm, and not just at the MDP representation level. Specifically, in [21] and [22] we introduced a *soft probabilistic bias* in model-based and model-free RL, respectively, by replacing the classical uniform distribution for action sampling in exploration with a weighted distribution on logically entailed actions. Crucially, while prior symbolic knowledge was defined for simple domain instances, our methodology yielded faster and more efficient training in more complex settings for several benchmark RL tasks, showing promising generalization performance. In more recent work [23], we also generalized this approach to reason over temporal logic specifications, which are highly expressive for planning problems and further improve the generalization and training performance. We also investigated the balance between NeSy exploration and exploitation in model-free RL, by combining the symbolic exploration bias with the re-scaling of learned Q-values for logically entailed actions. This solution significantly improved the training performance in complex benchmarks, involving partial observability which is a major challenge in robotic and aquatic domains. It also proved more robust to imperfect logic specifications than reward shaping, thanks to the soft bias and the progressive reduction of Q-value re-scaling during training (see Fig. 3). Finally, symbolic AI can also be used to explain black-box RL policies, e.g., by learning their logical approximations from training experience [22], yielding enhanced trustworthiness and possibly fostering human-robot interaction (e.g., human control or teaching) [24].

3.3. Safety-aware Model-based RL

Considering the uncertainty about the transition model in reinforcement learning can allow to obtain theoretical guarantees on the policy improvement. In this context, we develop scalable algorithms for

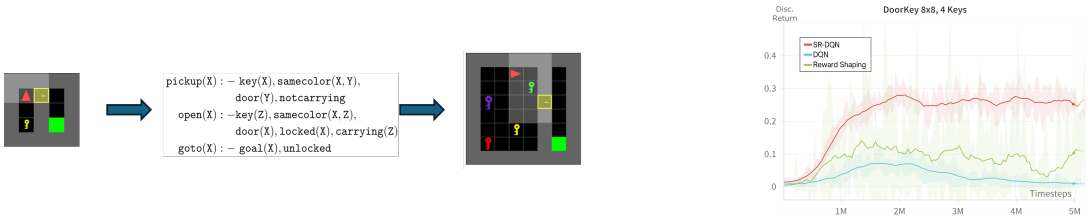


Figure 3: Our NeSy RL approach applied to the partially observable DoorKey benchmark (https://minigrid.farama.org/environments/minigrid/DoorKeyEnv/). The agent must reach the green destination by picking the right key for opening a door. Imperfect symbolic knowledge in 5×5 map with one key (left) wrongly suggests moving to the goal with the open door, preventing initial grid exploration. It yields bad performance with reward shaping in the more challenging 8×8 map with 4 keys (green curve, right).

Safe Policy Improvement in single and multi-agent scenarios [25, 26]. These algorithms, based on Monte Carlo Tree Search (MCTS), assume the availability of a baseline policy and a dataset of trajectories collected by that policy, and they allow us to learn a new policy with theoretically guaranteed improved performance. We also develop methods for learning online the transition model in MCTS [27] and methods that merge MCTS and reactive planning for safe and efficient motion planning in dynamic environments [28]. Finally, we have recently proposed a novel offline RL approach, inspired by Seldonian optimization, returning policies with good performance and statistically guaranteed properties with respect to predefined undesirable behaviors [29].

3.4. Gaussian Process for Enhanced Exploration in Partially Observable Settings

To address RL tasks under partial observability, in [30], we propose a belief state model based on Gaussian Processes (GPs). The model employs local GPs with online updates and incorporates pseudo-observations to improve both agent performance and exploration efficiency. For evaluation, we compared our approach against two baselines: a non-GP model operating under the same partial observability conditions, and a privileged model with full observability. Our results demonstrate that the GP-based belief model was able to recover approximately 70% of the performance gap between the partially observable baseline and the fully observable privileged model.

4. Ensuring Safe Deployment of RL Agents

Achieving safe autonomous navigation systems is critical for deploying robots in dynamic and uncertain real-world environments. In this section, we present our recent hierarchical control framework leveraging neural network verification techniques [31] to design control barrier functions (CBFs) and policy correction mechanisms that ensure safe reinforcement learning navigation policies [32]. Our idea is to compute a safe navigation set by first identifying unsafe regions through offline probabilistic enumeration [33] and then removing them from the state space of the DRL policy. We then design a CBF-based control mechanism to ensure the agent avoids both these pre-identified unsafe regions and any obstacles detected during navigation—ensuring the agent remains within the precomputed safe set. At deployment time, a quadratic programming (QP) optimization incorporating the CBF's safety constraints evaluates the policy's action, determining whether the chosen action keeps the agent within the precomputed safe set or not. If the action violates any safety constraints, a low-level controller modifies the action to ensure the agent returns to the safe set while maintaining effective navigation behaviors. This procedure

Aquatic Evaluation		
Method	**Success (%)**	**Collision (%)**
PPO	84.2 ± 9.75 %	11.22 ± 3.26 %
PPO+CBF	**87.0 ± 12.72%**	**0.0 ± 0.0** %
PPO_penalty	89.7 ± 5.19 %	0.007 ± 0.005 %
PPO_penalty+CBF	**96.7 ± 2.30%**	**0.0 ± 0.0** %
PPOLag	84.0 ± 5.65 %	0.001 ± 0.002 %
PPOLag+CBF	**94.0 ± 4.24%**	**0.0 ± 0.0** %

Table 1: Evaluation of verification-guided CBF.

generates reference velocities, which serve as inputs for an optimal low-level controller based on nonlinear model predictive control (NMPC) [34]. The latter computes the optimal control input for the plant, which evolves using the dynamic, providing both the new state for the controller and CBF, and for the DRL policy. The proposed autonomous safe navigation framework is validated via a thorough evaluation in a realistic aquatic monitoring simulation task (Fig. 1 center). Our goal is to assess, given a DRL-trained policy for a mapless navigation task, whether our verification-guided CBF layer could consistently correct unsafe actions (i.e., ensure zero collisions) while preserving goal-reaching efficiency. The benefits of our CBF layer are clearly highlighted in Tab. 1, where all the given (unsafe) DRL policies combined with our method have a higher success rate while achieving safe navigation with zero collisions.

5. Conclusion and Future Works

In this paper, we present recent advancements in addressing the critical challenge of developing safe and explainable AI systems, with a particular focus on deep reinforcement learning agents for complex environmental sustainability tasks. Our current and future research aims to optimize and rigorously test these algorithms in real-world settings, while enhancing trustworthy and transparent human-AI interaction. To achieve this, we integrate techniques from neural network verification, control theory, neuro-symbolic AI, and explainable AI. Ultimately, our goal is to deploy advanced and sustainable systems that protect both people and the environment.

Acknoledgments

This work has been supported by PNRR MUR project PE0000013-FAIR.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to grammar and spelling check, paraphrase, and reword. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] C. E. Boyd, et al., Achieving sustainable aquaculture: Historical and current perspectives and future needs and challenges, Journal of the world aquaculture society 51 (2020) 578–633.
[2] D. Guihen, The challenges and opportunities for the use of robotic autonomous robotic systems in support of the blue economy, in: International Conference on Offshore Mechanics and Arctic Engineering, volume 86922, American Society of Mechanical Engineers, 2023.
[3] F. Trotti, A. Farinelli, R. Muradore, Towards aircraft autonomy using a pomdp-based planner, in: 2024 American Control Conference (ACC), 2024, pp. 2399–2404.
[4] F. Trotti, A. Farinelli, R. Muradore, A markov decision process approach for decentralized uav formation path planning, in: 2024 European Control Conference (ECC), IEEE, 2024, pp. 436–441.
[5] F. Trotti, A. Farinelli, R. Muradore, Path re-planning with stochastic obstacle modeling: A monte carlo tree search approach, in: 2024 IEEE/RSJ IROS, 2024, pp. 8017–8022.
[6] F. Trotti, A. Farinelli, R. Muradore, An online path planner based on pomdp for uavs, in: 2023 European Control Conference (ECC), IEEE, 2023, pp. 1–6.
[7] J. Ji, B. Zhang, J. Zhou, X. Pan, W. Huang, R. Sun, Y. Geng, Y. Zhong, J. Dai, Y. Yang, Safety gymnasium: A unified safe reinforcement learning benchmark, in: NeurIPS, volume 37, 2023.
[8] G. Amir, D. Corsi, R. Yerushalmi, L. Marzari, D. Harel, A. Farinelli, G. Katz, Verifying learning-based robotic navigation systems, in: 29th Int. Conf., TACAS 2023, Springer, 2023, pp. 607–627.

[9] E. Altman, Constrained markov decision processes, in: CRC Press, 1999.
[10] A. Stooke, J. Achiam, P. Abbeel, Responsive safety in reinforcement learning by pid lagrangian methods, in: International Conference on Machine Learning (ICML), 2020.
[11] D. Corsi, L. Marzari, A. Pore, A. Farinelli, A. Casals, P. Fiorini, D. Dall'Alba, Constrained reinforcement learning and formal verification for safe colonoscopy navigation, in: 2023 IEEE/RSJ IROS, 2023, pp. 10289–10294.
[12] L. Marzari, E. Marchesini, A. Farinelli, Online safety property collection and refinement for safe deep reinforcement learning in mapless navigation, in: 2023 IEEE ICRA, 2023, pp. 7133–7139.
[13] E. Marchesini, L. Marzari, A. Farinelli, C. Amato, Safe deep reinforcement learning by verifying task-level properties, in: AAMAS 2023, 2023, p. 1466–1475.
[14] L. Marzari, P. L. Donti, C. Liu, E. Marchesini, Improving policy optimization via ε-retrain, in: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, 2025, p. 1464–1472.
[15] C. Tessler, D. J. Mankowitz, S. Mannor, Reward constrained policy optimization, in: ICLR, 2019.
[16] L. Zhang, L. Shen, L. Yang, S. Chen, B. Yuan, X. Wang, D. Tao, Penalized proximal policy optimization for safe reinforcement learning, in: IJCAI, 2022.
[17] Y. Jiang, J. Z. Kolter, R. Raileanu, On the importance of exploration for generalization in reinforcement learning, NeurIPS 36 (2024).
[18] G. De Giacomo, L. Iocchi, M. Favorito, F. Patrizi, Foundations for restraining bolts: Reinforcement learning with ltlf/ldlf restraining specifications, in: ICAPS, volume 29, 2019, pp. 128–136.
[19] C.-A. Cheng, A. Kolobov, A. Swaminathan, Heuristic-guided reinforcement learning, NeurIPS 34 (2021) 13550–13563.
[20] H. Kautz, The third ai summer: AAAI R.S. Engelmore memorial lect., Ai mag. 43 (2022) 105–125.
[21] D. Meli, A. Castellini, A. Farinelli, Learning logic specifications for policy guidance in pomdps: an inductive logic programming approach, JAIR 79 (2024) 725–776.
[22] C. Veronese, D. Meli, A. Farinelli, Online inductive learning from answer sets for efficient reinforcement learning exploration, arXiv:2501.07445 (2025). HYDRA workshop @ECAI 2024.
[23] C. Veronese, D. Meli, A. Farinelli, Learning symbolic persistent macro-actions for pomdp solving over time, arXiv:2505.03668 (2025). Accepted 19th Conf. Neurosymbolic Learning and Reasoning.
[24] D. Meli, P. Fiorini, Inductive learning of robot task knowledge from raw data and online expert feedback, Machine Learning 114 (2025) 91.
[25] A. Castellini, F. Bianchi, E. Zorzi, T. D. Simão, A. Farinelli, M. T. J. Spaan, Scalable safe policy improvement via Monte Carlo tree search, in: ICML 2023, 2023, pp. 3732–3756.
[26] F. Bianchi, E. Zorzi, A. Castellini, T. D. Simão, M. T. J. Spaan, A. Farinelli, Scalable safe policy improvement for factored multi-agent MDPs, in: ICML 2024, volume 235, 2024, pp. 3952–3973.
[27] M. Zuccotto, E. Fusa, A. Castellini, A. Farinelli, Online model adaptation in monte carlo tree search planning, Optimization and Engineering (2024).
[28] L. Bonanni, D. Meli, A. Castellini, A. Farinelli, Monte carlo tree search with velocity obstacles for safe and efficient motion planning in dynamic environments, in: AAMAS 2025, IFAAMAS, 2025.
[29] E. Zorzi, A. Castellini, L. Bakopoulos, G. Chalkiadakis, A. Farinelli, Seldonian reinforcement learning for ad hoc teamwork, Reinforcement Learning Journal, RLC 2025 2 (2025).
[30] A. Zhalehmehrabi, D. Meli, F. D. Santo, F. Trotti, A. Farinelli, Depth-constrained asv navigation with deep rl and limited sensing, arXiv preprint arXiv:2504.18253 (2025).
[31] C. Liu, T. Arnon, C. Lazarus, C. Strong, C. Barrett, M. J. Kochenderfer, et al., Algorithms for verifying deep neural networks, Foundations and Trends® in Optimization 4 (2021) 244–404.
[32] L. Marzari, F. Trotti, E. Marchesini, A. Farinelli, Designing control barrier function via probabilistic enumeration for safe reinforcement learning navigation, IEEE Robotics and Automation Letters 10 (2025) 9630–9637. doi:10.1109/LRA.2025.3596431.
[33] L. Marzari, D. Corsi, E. Marchesini, A. Farinelli, F. Cicalese, Enumerating safe regions in deep neural networks with provable probabilistic guarantees, in: AAAI, volume 38, 2024, pp. 21387–21394.
[34] L. Grüne, J. Pannek, L. Grüne, J. Pannek, Nonlinear model predictive control, Springer, 2017.