Safe and Efficient Reinforcement Learning for Environmental Monitoring

Federico Bianchi, Davide Corsi, Luca Marzari, Daniele Meli, Francesco Trotti, Maddalena Zuccotto, Alberto Castellini* and Alessandro Farinelli

Department of Computer Science, University of Verona, Italy

Abstract

This paper discusses the challenges of applying reinforcement techniques to real-world environmental monitoring problems and proposes innovative solutions to overcome them. In particular, we focus on safety, a fundamental problem in RL that arises when it is applied to domains involving humans or hazardous uncertain situations. We propose to use deep neural networks, formal verification, and online refinement of domain knowledge to improve the transparency and efficiency of the learning process, as well as the quality of the final policies. We present two case studies, specifically (i) autonomous water monitoring and (ii) smart control of air quality indoors. In particular, we discuss the challenges and solutions to these problems, addressing crucial issues such as anomaly detection and prevention, real-time control, and online learning. We believe that the proposed techniques can be used to overcome some limitations of RL, providing safe and efficient solutions to complex and urgent problems.

Keywords

Reinforcement Learning, Safety, Water Monitoring, Air Quality Management, Environmental Sustainability

1. Introduction

In recent years, Reinforcement Learning (RL) has emerged as a powerful technique to solve complex problems in a variety of applications, reaching and often outperforming classical algorithms and humans. Successful stories include chess and Go [1], video games [2], and more [3, 4]. One fundamental trend of research in RL addresses the problem of *safety*, i.e., the application of RL solutions to problems and domains involving interaction with humans, hazardous situations and expensive hardware. This is the case, for instance, of robotics and autonomous driving [5]. In fact, standard model-free RL approaches just aim at maximizing a given reward signal, computing an optimal strategy or *policy* for the task. It is then important to provide safety, and correctness guarantees [6].

Moreover, a crucial limitation of RL lies in the overwhelming requirements in terms of data availability, storage, and power, ultimately affecting environmental sustainability and large-scale adoption of such techniques, especially in systems with limited hardware resources.

In this paper, we propose techniques to address the problem of safe RL with limited resources, leveraging symbolic artificial intelligence, formal verification tools, and online refinement of domain knowledge to improve the transparency and efficiency of the learning process, as well as the quality of the final policy.

We apply our solutions in the context of environmental monitoring and preservation, specifically for *exploration and analysis of water catchments with autonomous surface vehicles* and *smart control of air quality indoors.* Both scenarios require safe and prompt decision-making in highly dynamic conditions based on uncertain information from many heterogeneous sensors and in the presence of humans or other living beings.

In the following, we introduce our benchmark domains and discuss how to effectively apply RL to them, addressing crucial issues such as anomaly detection and prevention, real-time control, and online learning.

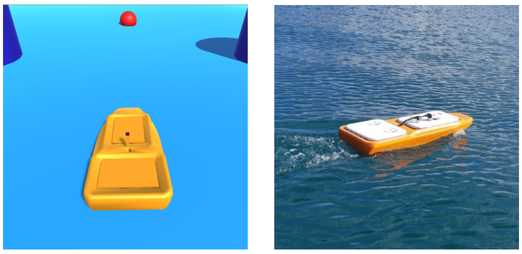
2. Water Monitoring

In this section, we introduce one of the two problems we address in this paper, the *autonomous water monitoring*. This task presents different challenges that range from hardware to software [7]. Fig. 1 shows the drone we used for our experiments, a differential drive platform equipped with sensors and actuators that allows the boat to navigate autonomously in the environment and collect data on the water's quality. In more detail, the platform is based on a hull equipped with two in-water propellers that allow navigation in shallow water with a max velocity of $3m/s$. The set of sensors for localization includes a GPS, compass, accelerometer, and gyroscope, with the addition of a sonar and a stereo camera for collision avoidance. Regarding water quality, the base set includes PH, dissolved oxygen, conductivity, and temperature; however, our drone is modular and different types

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

alberto.castellini@univr.it (A. Castellini)



(a) Training simulator (b) Real aquatic drone

Figure 1: The water monitoring drone and the Unity3D simulator.

of sensors can be integrated, e.g., for heavy metals, as well as a system for collecting water samples.

Our drone is designed to be easy to deploy for non-expert users; therefore, a high level of autonomy is required. The main objective is to provide a platform to be released into a reservoir, able to navigate and monitor water parameters and potential sources of pollution with very limited human intervention [7]. However, autonomous navigation in the aquatic scenario is challenging due to the non-stationarity of the environmental conditions such as waves, wind, light reflections, moving obstacles (both above and below the surface), and more [8]. Hence, safety is of utmost importance to avoid catastrophic failures, e.g., collisions causing water to seep into the hull and severely damaging electrical components [9].

In the following sections, we will propose techniques from safe RL and particularly safe DRL, i.e., RL based on Deep Neural Networks (DNNs), in order to solve critical issues related to real-time control of water drones under dynamic uncertainty [10].

Combining Reinforcement Learning and Control

Against the background introduced in the previous sections, we propose to exploit DRL to generate a DNN-based real-time controller for the water monitoring drone. DRL is a unique paradigm for the training of DNN. Its fundamental characteristic is that it does not require a set of labeled data; in contrast, an agent (the drone in our case) interacts with the environment, with a trial-and-error process, learning from its mistakes and improving the *policy* (i.e., the strategy to solve the task), driven only by a reward function that it aims to maximize [11]. Using model-free DRL to learn low-level control commands for robots is known to be a challenging problem, also due to electrical and mechanical hardware limitations posing safety constraints [12]. Hence, we integrate DRL with a low-level controller for v, ψ, respectively the linear velocity and the heading (yaw) angle of the drone. Specifically, we first consider the following dynamic model of the agent:

$$\begin{aligned} \dot{x} &= v\cos(\psi) \quad & \dot{y} &= v\sin(\psi) \\ \dot{v} &= \frac{T-D}{m} \quad & \dot{\psi} &= w \end{aligned} \tag{1}$$

where x, y are the boat position in earth frame, T is the motors thrust, m is the boat mass, w is the angular velocity in body frame and D is the water drag coefficient.

Let now $\boldsymbol{x} = [x, y, v, \psi]$ the state vector of the system. During DRL training phase, the DRL agent explores different actions $\langle v\psi \rangle$, which are the input to a feedback inverse dynamic control loop incorporating the drone model. In this way, the low-level commands at the drive of the drone are ultimately sent by a model-based controller, guaranteeing their feasibility. The DRL agent then obtains a new dynamically correct state vector with the associated reward value for the specific action. For the reward signal, we adapted a well-known function in the context of mapless navigation, formally:

$$R_t = \begin{cases} 1 & \text{reaches the end} \\ -1 & \text{collision with an obstacle} \\ (d_{t-1} - d_t) \cdot \alpha - \beta & \text{otherwise} \end{cases} \tag{2}$$

where d_t is the distance from the target position at time t, α is a normalization factor used to guarantee the stability of the gradient, and β is a fixed value, decreased at each time-step, resulting in a total penalty proportional to the length of the path.

Deep reinforcement learning (DRL) requires thousands of interactions between an agent and its environment. Training the agent directly on the actual platform is often unfeasible, particularly in a robotic scenario where a failure during the early stages of training cannot be tolerated. To overcome this challenge, we have developed a realistic simulator using the Unity3D engine. Leveraging the built-in physics engine, our simulator can accurately simulate water, waves, and other weather phenomena (Fig. 1 shows a screenshot of our simulator).

DRL training is performed using twp extensions of the Proximal Policy Optimization algorithm (PPO) [13]

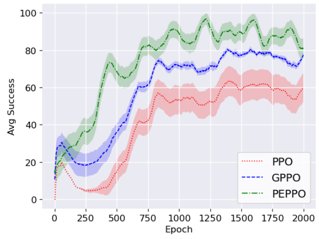


Figure 2: Training results of our reinforcement learning approaches.

(a state-of-the-art approach), called GPPO and PEPPO, that exploit methodologies based genetic approaches to enhance the training phase, showing promising results on our navigation tasks [7]. Fig. 2 shows a plot of the results obtained with our approach; even though *reward-wise* the performance is good, one limitation of the standard DRL approaches is that they are oriented only on the performance, often failing to provide guarantees on high-level safety criteria, which are crucial for water drones. A possible solution is to exploit constrained reinforcement learning approaches, such as Lagrangian-PPO and CPO [14]. In the following section, we propose an approach for ex-post verification of safety after training.

Formal Verification of DNNs

DNNs have shown impressive performance in various tasks. However, the vulnerability of these models to adversarial inputs is a well-documented phenomenon observed across various applications [15]. Formal Verification (FV) of DNNs uses mathematical methods to rigorously prove that a neural network meets certain safety and reliability requirements expressed as input-output relationships [16]. In more detail, the goal of the FV is to provide guarantees that a DNN will behave correctly and predictably under all possible inputs and conditions. This is especially important in safety-critical applications such as autonomous driving or medical procedures, where any error or unexpected behavior of these models could have catastrophic consequences. The FV of DNNs is still a challenging and active area of research. One of the main challenges is the computational complexity of verifying very large networks [17]. Hence, researchers are actively developing new techniques and tools to make formal verification more practical for real-world applications [16, 18, 19]. In particular, we developed an interval propagation-based method called *ProVe* [20]. Our algorithm is more computationally efficient since it performs parallel verification on sub-intervals of the input space. Based on the results obtained with *ProVe*, we extended the classic DNN-Verification problem to a new type of verification called *#DNN-Verification* or *quantitative verification* [21], which aims not only to discover a potentially single unsafe configuration but to count (or enumerate) all the unsafe areas in a particular region of the input space expressed by the safety property. This type of verification allows estimating the probability that the agent violates specific safety properties and selecting fully safe models before final deployment.

Anomaly Detection and Recovery

Water drones may face unpredictable challenges, e.g., battery drainage, unexpected obstacles, and adverse weather conditions. Thus, it is crucial to analyze the real-time stream from onboard sensors and find relevant patterns useful for decision-making and anomaly recovery or prevention. However, interpreting a large amount of heterogeneous data and handcrafting effective reward specifications for RL accordingly is often very challenging [22], especially in complex, uncertain, and dynamic environments or when the reward is sparse [23].

We have addressed the problem of online anomaly detection for water drones using hidden Markov models combined with normalized Hellinger distance [24], resulting in robust anomaly recognition when conducting exploration campaigns over multiple days. The anomaly signal can be used to provide an early reward in RL and prevent possible faults.

One limitation of purely data-driven methods for anomaly detection is the lack of interpretability. To address this problem, we have started to investigate the use of Satisfiability Modulo Theory (SMT) [25] and Inductive Logic Programming (ILP) [26] to automatically identify high-level logical patterns in data. ILP allows to easily incorporate domain knowledge, increasing data and computational efficiency. Moreover, it can provide situational and policy explanations [27], which can be useful to more effectively communicate with on-coast crew in case of emergency. We have preliminarily explored the use of ILP to detect behavioral patterns in paradigmatic simulation tasks involving an autonomous agent operating in an uncertain environment with sparse rewards [28]. Starting from the definition of high-level commonsense concepts about the domain of interest, we have collected traces of normal executions (i.e., state-action pairs) of the agent and identified logical rules matching actions and environmental concepts. Rules have proven useful to guide the exploration process in unexperienced settings, achieving an improvement in computational time and the final reward. Our approach can be extended to RL, using rules learned from normal drone executions to guide optimal policy search and early identifying anomalous regions of the state-action space. Moreover, ILP can be used offline to generate explanations for registered anomalous behaviors and provide useful contrastive explanations [29], as well as to discover high-level safety specifications implicitly embedded in the nominal behavior.

3. Air Quality Management

The second problem we address in this paper concerns managing air salubrity to ensure comfort and safety in closed environments. Air quality and thermal comfort control are important features of modern Heating, Ventilation and Air Conditioning (HVAC) systems. The spreading of the SARS-Cov-2 pandemic highlighted the importance of air quality conditions in indoor environments since it has a positive impact in reducing virus spread.

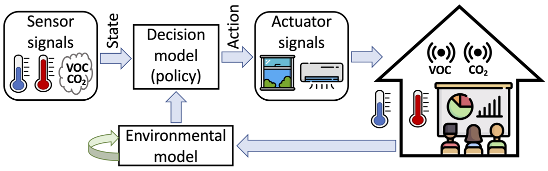


Figure 3: Overview of the air quality monitoring system.

We consider the example of a company meeting room that must be booked in advance for a given number of people. The room is equipped with sensors for measuring concentrations of CO_2 and Volatile Organic Compounds (VOCs), the indoor temperature and the outdoor temperature (from weather forecast). Moreover, actuators to open/close windows and turn on/off vents and sanitizers are available. We assume to have a daily occupation schedule of the room (i.e. number of persons present in the room at each hour of the day).

In the following, we first assume to have an approximate model of the scenario (e.g., provided by experts), expressed as a Markov Decision Process (MDP), and use Monte Carlo Tree Search (MCTS) [30] to compute the policy. Then, we propose a methodology for data-driven Safe Policy Improvement (SPI) online, exploiting large amounts of data available from environmental sensors [31, 32], with the potential to deal with model inaccuracies. An overview of our approach to air quality management is depicted in Figure 3.

Monte Carlo planning

MCTS is a popular algorithm for optimal decision making in large state / observation spaces. It performs forward simulations from the current state of the environment, in order to compute the best action at each step maximizing the expected reward. Our research deals with probabilistic planning and reinforcement learning methods based on MCTS in both completely observable and partially observable [33, 34] environments, and related applications to real-world problems.

Figure 4 shows, for instance, a detail of an example of a control profile generated by MCTS (on the left) and by an expert (on the right) and their effects on an environmental variable, i.e., VOCs concentration. On the first row, we can observe the room occupancy profile, that is, the number of people (blue line) in the room at a given time of the day. In the second row, we represent the evolution of VOCs concentration, and finally in the third row, the actions that are performed. As we can see in the second row, VOCs concentration is always significantly below the maximum threshold (orange line) by applying the policy produced by MCTS (on the left), while it reaches the maximum threshold value by performing the actions suggested by the expert (on the right). MCTS obtains better performance than the expert by leveraging its simulation-based nature. Moreover, as highlighted by green boxes, we can see that the activation of high-intensity ventilation and the sanitization process (third row) causes a sudden decrease in VOCs concentration (second row).

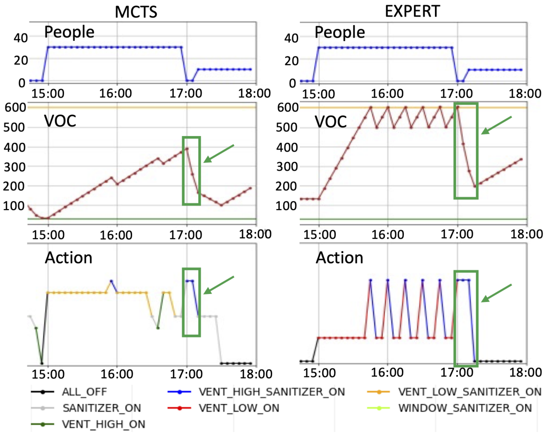


Figure 4: Detail on the effect of the actuator control profile applying the strategy produced by our RL approach and by an expert on VOC concentration.

Safe Policy Improvement

Another topic of our research is safe policy improvement. Algorithms for safely improving policies are important to deploy reinforcement learning approaches in real-world scenarios (e.g., autonomous cars, drones, or industrial plants) where safety, robustness, and reliability of control policies are crucial issues. Safe RL investigates how these issues can be addressed by learning policies that maximize expected return while ensuring minimal performance level or respecting safety constraints.

We focus on Safe Policy Improvement (SPI) [35] for MDPs $M =< S, A, T, R, \gamma >$, where the agent is provided with a baseline policy π_0 and dataset of trajectories $\mathcal{D} =< s_j, a_j, r_j, s'_j >_{j\in[1,N]}$ collected running in the real environment. SPI computes a new policy π_I that outperforms the baseline π_0 *safely* with an admissible performance loss $\zeta \in \mathbb{R}^+$ and confidence level $1 - \delta$, with $0 \leq \delta \leq 1$ and loss $\rho(\pi_0, M) - \rho(\pi_I, M)$.

Safe Policy Improvement with Baseline Bootstrapping (SPIBB) [36, 37] is a state-of-the-art method that considers the worst-case scenario reformulating the percentile criterion [38] to make the search for an efficient and provably-safe policy tractable. SPIBB splits state-action pairs into two subsets: the bootstrapped subset $\mathcal{B} = \{(s, a) : N_{\mathcal{D}}(s, a) < N_\wedge\}$ is the set of state-action pairs that occur less than $N_\wedge$ times in $\mathscr{D}$; the non-bootstrapped set $\overline{\mathscr{B}} = \{(s,a) : N_{\mathscr{D}}(s,a) \geq N_\wedge\}$ is the set of state-action pairs that occur at least $N_\wedge$ times in $\mathscr{D}$. The approach guarantees that π^{spibb} is a ζ-approximate safe policy improvement of the baseline π_0 with high probability $1-\delta$, where ζ depends on $N_\wedge$ and δ.

Our current research deals with extending state-of-the-art safe policy improvement algorithms to enable their applicability to real world problems. This manily requires the scaling of the algorithms to very large state-spaces. We are currently evaluating different methodologies for improving the scaling capabilities of safe policy improvement methods and trying to evaluate the capabilities of such approaches to work on real-world domains with several states, domains having partially observable states and domains with multiple agents.

4. Final Remarks

In this paper, we tackled the pressing issue of exploiting AI, and particularly DRL, to realize advanced solutions for sustainability. Specifically, we examined two complex scenarios: water monitoring with autonomous drones and automatic air quality management indoors. We proposed methodologies prioritizing safety as a crucial requirement for cyber-physical systems operating in real environments in the presence of humans. We combined model-free DRL with tools from formal verification, standard control theory, and model-based planning.

Our current and future research is devoted to optimizing and further testing our algorithms in real-world contexts, and exploring more trustable and transparent interaction with humans exploiting techniques from explainable AI, e.g., ILP and logics. Our ultimate goal is to deploy advanced and sustainable systems for the safeguard of both human beings and the environment.

Acknowledgments

The research has been partially supported by the following projects: SAFE PLACE, POR-FESR 2014-2020, Regione del Veneto; PON "Ricerca e Innovazione 2014-2020", Italian Ministry for University and Research (grant agreement No. 40-G-14702-1); "INEST Interconnected Nord-Est Innovation", Piano Nazionale della Ricerca (PNR).

References

[1] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of go without human knowledge, Nature (2017).

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing Atari with Deep Reinforcement Learning, 2013. Technical Report. https://arxiv.org/abs/1312.5602.

[3] A. Pore, D. Corsi, E. Marchesini, D. Dall'Alba, A. Casals, A. Farinelli, P. Fiorini, Safe Reinforcement Learning using Formal Verification for Tissue Retraction in Autonomous Robotic-Assisted Surgery, in: Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS), 2021.

[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., Highly accurate protein structure prediction with alphafold, Nature (2021).

[5] A. Sallab, M. Abdou, E. Perot, S. Yogamani, Deep reinforcement learning framework for autonomous driving, Electronic Imaging (2017).

[6] D. Corsi, E. Marchesini, A. Farinelli, Formal Verification of Neural Networks for Safety-Critical Tasks in Deep Reinforcement Learning, in: Proc. 37th Conf. on Uncertainty in Artificial Intelligence (UAI), 2021.

[7] E. Marchesini, D. Corsi, A. Farinelli, Benchmarking safe deep reinforcement learning in aquatic navigation, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021.

[8] L. Steccanella, D. D. Bloisi, A. Castellini, A. Farinelli, Waterline and obstacle detection in images from low-cost autonomous boats for environmental monitoring, Robotics and Autonomous Systems 124 (2020) 103346.

[9] R. S. Sutton, D. McAllester, S. Singh, Y. Mansour, Policy gradient methods for reinforcement learning with function approximation, Advances in neural information processing systems (1999).

[10] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT press, 2018.

[11] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, D. Silver, Rainbow: Combining improvements in deep reinforcement learning, in: 32nd AAAI conference on artificial intelligence, 2018.

[12] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, A. P. Schoellig, Safe learning in robotics: From learning-based control to safe reinforcement learning, Annual Review of Control, Robotics, and Autonomous Systems (2022).

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal Policy Optimization Algorithms, 2017. Technical Report. http://arxiv.org/abs/1707.06347.

[14] D. Corsi, R. Yerushalmi, G. Amir, A. Farinelli, D. Harel, G. Katz, Constrained reinforcement learning for robotics via scenario-based programming, arXiv preprint arXiv:2206.09603 (2022).

[15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv:1312.6199 (2013).

[16] C. Liu, T. Arnon, C. Lazarus, C. Strong, C. Barrett, M. J. Kochenderfer, Algorithms for verifying deep neural networks, Foundations and Trends® in Optimization (2021).

[17] G. Katz, C. Barrett, D. L. Dill, K. Julian, M. J. Kochenderfer, Reluplex: An efficient smt solver for verifying deep neural networks, in: International conference on computer aided verification, Springer, 2017, pp. 97–117.

[18] S. Wang, H. Zhang, K. Xu, X. Lin, S. Jana, C.-J. Hsieh, J. Z. Kolter, Beta-CROWN: Efficient bound propagation with per-neuron split constraints for complete and incomplete neural network verification, Advances in Neural Information Processing Systems 34 (2021).

[19] P. Henriksen, et al., Deepsplit: An efficient splitting method for neural network verification via indirect effect analysis., in: IJCAI, 2021, pp. 2549–2555.

[20] D. Corsi, E. Marchesini, A. Farinelli, Formal verification of neural networks for safety-critical tasks in deep reinforcement learning, in: Conference on Uncertainty in Artificial Intelligence (UAI), 2021.

[21] L. Marzari, D. Corsi, F. Cicalese, A. Farinelli, The #dnn-verification problem: Counting unsafe inputs for deep neural networks, ArXiv abs/2301.07068 (2023).

[22] H. Zou, T. Ren, D. Yan, H. Su, J. Zhu, Learning task-distribution reward shaping with meta-learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 11210–11218.

[23] M. Grześ, Reward shaping in episodic reinforcement learning, in: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, 2017, pp. 565–573.

[24] D. Azzalini, A. Castellini, M. Luperto, A. Farinelli, F. Amigoni, Hmms for anomaly detection in autonomous robots, in: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, 2020, pp. 105–113.

[25] G. Mazzi, A. Castellini, A. Farinelli, Identification of unexpected decisions in partially observable monte-carlo planning: A rule-based approach, in: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS), International Foundation for Autonomous Agents and Multiagent Systems, 2021, p. 889–897.

[26] S. Muggleton, L. De Raedt, Inductive logic programming: Theory and methods, The Journal of Logic Programming 19 (1994) 629–679.

[27] A. Drozdov, M. Law, J. Lobo, A. Russo, M. W. Don, Online symbolic learning of policies for explainable security, in: 2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, IEEE, 2021, pp. 269–278.

[28] G. Mazzi, D. Meli, A. Castellini, A. Farinelli, Learning logic specifications for soft policy guidance in pomcp, in: Proceedings of the 22nd Conference on Autonomous Agents and MultiAgent Systems, 2023, p. in publication.

[29] J. Rabold, M. Siebers, U. Schmid, Generating contrastive explanations for inductive logic programming based on a near miss approach, Machine Learning 111 (2022) 1799–1820.

[30] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, S. Colton, A survey of monte carlo tree search methods, IEEE Transactions on Computational Intelligence and AI in Games 4 (2012) 1–43. doi:10.1109/TCIAIG.2012.2186810.

[31] A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, Internet of things for smart cities, IEEE Internet of Things Journal 1 (2014) 22–32. doi:10.1109/JIOT.2014.2306328.

[32] M. Capuzzo, A. Zanella, M. Zuccotto, F. Cunico, M. Cristani, A. Castellini, A. Farinelli, L. Gamberini, Iot systems for healthy and safe life environments, in: 7th Forum on Research and Technologies for Society and Industry Innovation (RTSI), 2022.

[33] A. Castellini, G. Chalkiadakis, A. Farinelli, Influence of State-Variable Constraints on Partially Observable Monte Carlo Planning, in: IJCAI 2019, Macao, China, August 10-16, 2019, ijcai.org, 2019, pp. 5540–5546.

[34] M. Zuccotto, A. Castellini, A. Farinelli, Learning state-variable relationships for improving pomcp performance, in: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, SAC '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 739–747.

[35] P. S. Thomas, G. Theocharous, M. Ghavamzadeh, High confidence policy improvement, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML 2015), JMLR.org, 2015, pp. 2380–2388.

[36] R. Laroche, P. Trichelair, R. Tachet Des Combes, Safe policy improvement with baseline bootstrapping, in: Proceedings of the 36th International Conference on Machine Learning (ICML 2019), volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 3652–3661.

[37] T. D. Simão, R. Laroche, R. Tachet des Combes, Safe Policy Improvement with an Estimated Baseline Policy, in: Proc. AAMAS, IFAAMAS, 2020, pp. 1269–1277.

[38] E. Delage, S. Mannor, Percentile optimization for Markov Decision Processes with parameter uncertainty, Operations Research 58 (2010) 203–213.