Assured Multi-Agent Reinforcement Learning Using Quantitative Verification

Joshua Riley[1][a]
[1]*Department of Computer Science, University of York, York, UK*

1 RESEARCH PROBLEM

When looking at multi-agent systems (MAS), we can define them as a collection of agents, physical or virtual, which have shared responsibilities within the domain they are working within. These systems have the potential to be utilised in a myriad of different settings, such as defence, industry, agricultural, and more (Fan et al., 2011). These systems can be composed of a range of different agents, from complex heterogeneous robotic teams (Rizk et al., 2019) to homogeneous ariel drones (Yasin et al., 2020), the diversity of the construction of these systems allows them to be used in a wide range of problem domains.

There is a clear advantage to using these systems over their single-agent counterparts, one of these being their ability to scale to complex problem domains, in which, single agents would be inefficient within, in several ways. When viewing these systems from a safety engineering standpoint, they have great potential within safety-critical domains, and with complex agents being used within these systems, even have the ability to replace humans, who otherwise would have to be put in a position of risk to carry out jobs.

There are examples of these systems within safety-critical domains, such as within search and rescue operations(Gregory et al., 2016), and within situations that see agents within irradiated places. It has been seen during the disaster at the Fukushima nuclear power plant that these systems have been used for this exact purpose cite (Schwager et al., 2017)

However, with added functionality brings added complexity to the systems themselves, and the behaviours that they can exhibit as a system. This added complexity, within complex domains, can be incredibly time consuming and complicated for programmers to work with, causing issues with performance and reliability. Therefore, a substantial research area has emerged which focuses upon the pairing of MAS with a machine learning technique known as reinforcement learning (RL), to allow systems to learn to work together efficiently (Boutilier, 1996).

RL is a promising technique which enables agents to learn how to achieve system objectives efficiently (Patel et al., 2011), specifically within sequential decision-making problems. Agents with no knowledge of the problem environment will use a mixture of exploration (choosing an action at random), and exploitation (choosing the action with the largest reward), to find state action pairings that amount to an optimal policy. An optimal policy is the groupings of state-action pairs that maximise a reward.

There are many works which focus upon MAS and RL, otherwise known as multi-agent reinforcement learning (MARL). These works discuss the benefits of MARL, such as efficiency, and robustness through the division of labour. While also detailing the challenges, such as ensuring reliable communications, and increased complexity (Buşoniu et al., 2010).

MARL has been proposed to be a solution to a wide range of problems; this includes being used within the inspection of nuclear power plant (Bogue, 2011), and other hazardous environments, removing the need for humans. One large deterrent in the practical use of MARL in safety-critical scenarios is due to the inherent nature of RL and MARL, which is stochastic. There will be hazards within any problem domain that is classed as safety-critical, and these hazards must be given the highest of attention, failing to do so could lead to damage being sustained to humans, valuable resources, or the system itself.

This lack of concern for hazards is a continued issue with traditional RL, as guarantees of the learning agent needlessly performing risky behaviours are not provided. This concern comes from the fact that RL is used to maximise a reward function, and often safety concerns will be counter-productive to this aim, meaning they will not be considered (Garcia and Fernández, 2012). The lack of consideration dramatically limits the potential for these systems to be used in practical applications, as without guarantees of the agent avoiding needlessly risky actions, they can potentially cause harm. Merely trying to mitigate this issue by attempting to capture the safety concerns within a reward function is not sufficient, it is not possible to capture such complex safety demands within

[a] https://orcid.org/0000-0002-9403-3705

a simple numerical reward function.

This problem has been addressed in multiple pieces of work. However, minus recent work by (Mason et al., 2017), there are very limited comprehensive solutions to the problem (Garcia and Fernández, 2012), often these solutions involve completely removing behaviours from the agents that can lead to risk, in a safety-critical environment, this could be highly counter-productive. These largely theoretical solutions also have scalability issues, are unable to express non-trivial safety properties, and do not fully satisfy the objective of securing guarantees that certain requirements will be met.

In multi-agent reinforcement learning, there are even fewer pieces of work focusing on a general, non-problem-specific method of guaranteeing safety properties are met. Often this work focuses on specific aspects of safety, or problems, such as collision avoidance (Wang et al., 2016), autonomous cars (Shalev-Shwartz et al., 2016a), or does not satisfy this project's desired outcomes. Meaning at the time of writing there is not a comprehensive approach to safe MARL, as has been introduced to safe RL.

This project, influenced by recent works in safe single-agent RL (Mason et al., 2017), aims to produce an approach to safe MARL, which will provide solutions which are guaranteed to meet a myriad of safety requirements, while learned producing behaviour that adequately meets the functional requirements of the problem.

2 OUTLINE OF OBJECTIVES

The use of MARL within safety-critical environments, and safe MARL, has several issues with its current state. The first of these include a lack of approaches that can be applied to a broad range of systems and domains, many papers which could be defined as perusing safe MARL, focus on specific issues, domains, or agent types. The second issue is a lack of ability to describe non-trivial safety requirements to the MARL system accurately; it is not possible to describe complex requirement within a reward function, which is commonly used within MARL. The third issue involves the lack of guarantees which can be currently made about successfully following safety requirements, while the system learns to reach the functional requirements of the domain.

This project aims to mitigate these issues by introducing an approach which is depicted in figure one. This approach offers a solution to these issues by introducing a multi-step, interchangeable method for producing assured MARL policies. The main objectives which this approach are as follows.

1. Assured MARL should be applicable to a broad range of problems, domains, and systems, allowing users to follow the steps of the approach while altering the tools and techniques used freely, without compromising the integrity of the approach.
2. Assured MARL should be capable of allowing domain experts to express and implement a myriad of non-trivial safety and functional requirements to guide the MARL system.
3. Lastly, assured MARL should produce guarantees that the safety requirements will not be violated, while still allowing the MAS to complete the functional requirements efficiently, in regard to the constraints set by the safety requirements.

The aforementioned approach is shown in figure one; it is comprised of three main stages. The first stage all unneeded information is abstracted away from the problem in the form of an abstract Markov decision process (AMPD). This is done to ensure that the model is small enough for effective QV by abstracting out all unneeded information a model which is small enough for effective and efficient QV whilst retaining sufficient knowledge for meaningful policies can be obtained. The abstraction of MDPs is a well-known approach and is well established within safety engineering for reducing problems complexity (Cizelj et al., 2011).

In order for this stage to be done effectively, all properties which inform on the agents' safety and functional requirements must be included in the model, for example, states, actions or events, rewards, or costs. It is also necessary for this stage for the domain expert to define the safety and functional constraints which are desired to be met. This can be done with relative flexibility with the constraints being defined with probabilistic computational tree logic (PCTL) (Ciesinski and Größer, 2004).

PCTL, as the name suggests, is a temporal logic and can be used to express functional and safety specifications which need to be met as concise formulae.

The second stage involves analysing the AMDP that was obtained in step one using QV. Using a model checker such as PRISM, the domain expert can describe the AMDP in a state-based language (Parker and Norman, 2014). The QV tool will be able to verify the AMDP and make guarantees based on functional and safety constraints we defined previously.

If the AMDP has been described correctly, the QV tool will synthesis a policy or multiple policies, that will govern the MARL in the domain.

It may be necessary to return to stage one, if a policy cannot be found, as the description of the problem

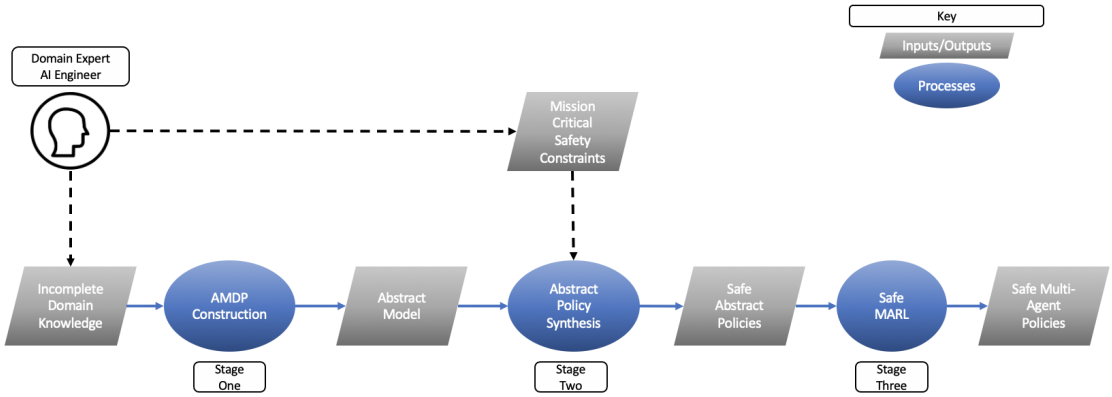

Figure 1: The three stages of our assured MARL approach

or the abstraction of the MDP may have contained flaws. It is also important to note if the AMDP is insufficient the guarantees made by the QV tool may not hold. It is vital, therefore that the problem is abstracted appropriately.

This ability to derive policies for which guarantees are possible is significantly different to other forms of RL and MARL, and also most other forms of safe RL, minus (Mason et al., 2017), a paper for safe single-agent RL, by which this project is largely influenced.

The final part of this approach sees the policy, which was synthesised in stage two, implemented into the actual domain problem in the form of behaviour constraints and task partitions, disallowing agents from performing needlessly risky actions.

The agents will learn within their partitioned task spaces, overlapping with each other when the responsibility of tasks should be shared and will be unable to enter risky situations unnecessarily.

Learning under these constraints will allow the agents to enter into risky situations, but with the use of QV, it is guaranteed that this risk will be bounded. This learning stage does not aim for optimality as it is common in safety engineering for the optimal strategy to be removed due to added constraints. This approach, however, does guarantee a degree of safety while also, in many cases, increasing the speed of learning.

3 STATE OF THE ART

3.1 Preliminaries

Reinforcement learning (RL) is a branch of machine learning that allows an agent to learn which action to choose in relation to the current state it is within. The premise of RL is the use of past experiences to influence an agent's behaviour within the future. As the agent moves around the problem space, it will receive numerical rewards based on the effectiveness of the action it has taken.

Domains which house these problems which RL aims to solve are described as the well established Markov Decision Process. The problem is broken down into states and actions that are available within those states. As the agent makes its way through the domain, it may choose between using an action known to be beneficial (exploitation) and those actions about which little is known (exploration).

When an agent takes action in a state, depending on the effectiveness of this action towards the end goal, the agent will receive a reward or punishment which will be associated with the state action pair $Q : (s,a) \rightarrow \mathbb{R}$. A widespread algorithm used to find the optimal value for these state action pairs is known as Q-learning (Patel et al., 2011).

This agent will continue to learn in this way until it has created a policy which will satisfy the objective, and if left to learn for long enough, will find the optimal policy.

A policy, in the case of RL, is a mapping of action to states with the most efficient mapping of these actions and states being known as an optimal policy. Traditional RL algorithms, such as QL, are solely concerned with finding the optimal policy within a problem. However, due to this, the traditional RL is not adequately equipped for safety constraints, and therefore, the optimal policy may not be safe.

Multi-agent reinforcement learning, an extension of RL, is an area of research which focuses on multiple agents working together within a system. These systems will learn to solve problems together in a

shared environment (Boutilier, 1996). As with RL, MARL has an expansive amount of literature detailing the challenges and the benefits that they can offer (Buşoniu et al., 2010). Some of these benefits, which is driving the research to be able to use these systems practically and safely, include robustness, division of labour, and efficiency. However, with these benefits come the issues of increased complexity, which is why this project includes an abstraction stage, as well as issues with ensuring reliable communication. There are three main types of algorithms which have been created for MARL, and these are as follows, independent learners, joint action learners, and gradient-descent algorithms (Buşoniu et al., 2010; Bloembergen et al., 2015).

3.2 Safe RL

The majority of approaches found within single-agent safe RL revolve around several types of approach depending on whether they are focus on exploration or optimisation features of learning (Garcia and Fernández, 2015). Many of these approaches which are relevant to the approach's aims fall under optimisation focus. Here many pieces of work have pushed towards the tailoring of reward functions, and also the restriction or manipulation of the rewards received (Serrano-Cuevas et al., 2019; Kroening et al., 2020).

Within the optimisation focus, there are three main areas of research, these being, worst-case criterion, risk-sensitive criterion, and constrained criterion (Garcia and Fernández, 2015). The types of works that tailor reward functions, as mentioned above often fall into the scope of worst-case criterion, and risk-sensitive criterion, however, the most promising work which complements the aims of the approach that this project proposes, falls into the constrained criterion.

The constrained criterion approaches are based on the premise of constraints being places on which behaviours and agent can act out, and which ones they may not, this can be seen being utilised in (Moldovan, 2012), to avoid irreversible actions, amongst others (Moldovan, 2012; Biyik et al., 2019). One piece of work introduced a novel approach to the constrained approach, this being the approach completed in (Mason et al., 2017), known as Assured RL.

The combination of a QV stage to the RL processes made by (Mason et al., 2017)introduced a multi-step approach to safe RL, allowing the verification of requirements to a single-agent domain. This work produced highly promising results and validated the research direction of creating approaches based on this premise to new areas.

3.3 Safe MARL

The research area of safe MARL appears less directed than that of safe RL, but certain trends can be found to have formed. When looking at recent advancements within safe MARL, there is a clear focus towards anti-collision of MARL systems, which is warranted given the nature of MAS (Zhang et al., 2019; Cheng et al., 2020; Khan et al., 2019). Also, there is a focus upon automated vehicles, as a large push in a research effort that could open many changes to the world (Shalev-Shwartz et al., 2016b), among some other domains, most notably traffic based domains (Rasheed et al., 2020; Guo, 2020; Lemos et al., 2018). However, much like Safe RL until recently, this work has been focused on specific domains problems, and specific techniques, such as specific algorithmic approaches, which may be restricting in broader use. It should also be noted that the reinforcement learning technique, known as deep learning, has become a large focus within this research area.

4 METHODOLOGY

In order to achieve the objectives of this project, we must progress through the steps which are listed below.

1. To begin with relevant literature must be explored, primarily around the research areas of safe RL, safe MARL, and MARL. In this way we can locate the limitations that are currently troubling the progression of MARL being used within safety-critical scenarios. With this knowledge we can more accurately determine the direction that our research and approach will take.
2. It is then possible for us to develop a theoretical approach to mitigating these limitations, and see how this approach compares to the current literature.
3. We can then begin developing a practical approach to solve the limitations based in a theoretical design, the two changing as needed based on practical limitations.
4. Evaluate our approach in two vastly different case studies, making use of unconstrained RL as a benchmark for our constrained RL produced by our approach. Our focus within the evaluation stages will be primarily on the safety requirements and how each type of RL manages to deal

with the problem space in regards to them, while still being able to complete the functional requirements.

5. Extend the scope of the approach by including advanced RL techniques, such as deep learning and evaluating how well the approach can scale to this.
6. Increasingly test the scalability of our approach, with different system sizes, more complex domains.

The approach will be evaluated using two domains created within simulators. The first of these is a patrolling simulator which allows many agents to navigate around problem spaces (Portugal et al., 2019). This ROS simulator, with the addition of RL strategies will allow us to evaluate our approach in a popular MAS problem with traditional RL. The second simulator is a physics simulator which is very widely used within the Deep RL community (Todorov et al., 2012). Here a case study will be extended to allow the approach to be used as a safety domain for Safe Deep RL.

The QV section of the approach will be dealt with using the model checker known as PRISM (Parker and Norman, 2014). This model checker supports the verification of reward-extended PCTL properties which can be used for the purposes of the proposed approach. PRISM has been used in previous works involving unmanned agents (Calinescu et al., 2017; Gerasimou et al., 2017; Gerasimou et al., 2018), and should be more than capable of being used within the proposed approach.

5 EXPECTED OUTCOME

The expected outcome of the project is to produce an assured MARL approach that will enable multiple reinforcement learning agents to navigate a plethora of domains, and problems, while complying to a myriad of safety requirements. This approach will introduce the ability to both express and satisfy diverse, complex, and sometimes conflicting safety requirements to MARL systems that can then reliably learn to meet functional requirements to an acceptable level of efficiency.

Our approach guarantees that safety requirements will be satisfied by making use of QV. This guarantee comes from the constraints placed onto the system, which are verified with QV to ensure the system will meet safety requirements without removing the ability to meet functional requirements.

As mentioned previously, this approach is intended to be used across many domains, but also with different system sizes, different RL algorithms, and in varying domain complexity.

In order for the approach to be capable of this, there are key prerequisites that must be supplied by the domain expert before learning can begin, these being the safety and functional requirements which will be expressed within PCTL, and also a high-level abstract model of the problem domain. The nature of these prerequisites supports flexibility in the safety requirements expressed, as well as the problem domain, as they can be crafted to the needs of the domain.

To this end, it is expected that the system's agents will be restricted in what actions they can and cannot perform within certain states of the system, while they are learning, and after learning is completed. It is also likely, by the nature of the domain space, and MAS, that states within the domain will be divided between agents within the system, potentially reducing the search space for each agent.

Agents learning within these constraints are guaranteed not to violate the safety requirements while also learning the optimal policy to solving the problem, the definition of optimal, in this case, being subjected to the effect the constraints have on the problem domain.

This approach will contribute largely to safe MARL, offering a way to express and satisfy complex and sometimes conflicting safety constraints, while still completing functional requirements with an acceptable level of efficiency.

6 STAGE OF THE RESEARCH

At this stage steps (1-3) have been completed, with substantial progress made in step (4), the approach has been implemented and successfully evaluated on one domain, showing promising initial results, these results, as well as the work, achieved so far can be found in (Riley et al., 2020). The aim which is being worked towards at the time of writing is expanding the approach to a separate domain, and potentially showcasing the approach using deep learning techniques to bridge the gap to the current projection within MARL research, which is heavily influenced by deep learning.

6.1 Completed Work

In order to evaluate the approach that has been presented, a domain has been created that tackles an issue that has been the focus of a number of pa-

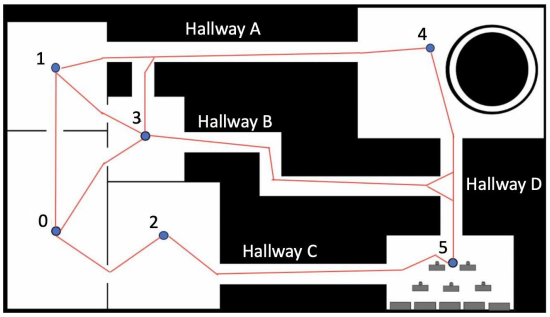

Figure 2: Nuclear reactor map within the simulator, overlaid with states and possible routes.

pers (Bogue, 2011). This domain is a robotic team working within a nuclear power plant. Two robots must learn to navigate the nuclear power plant, patrolling each main area, as shown in Figure 2, which is a view of the ROS simulator which this domain is housed within, with red lines showing the actions the robots have available to them between rooms.

The robots must meet the following constraints:

- C1: Visit each room a minimum of three times
- C2: Complete all tasks without exhausting their batteries

Looking at these constraints, we can see that C1, is a functional requirement of visiting each room in the power plant three times. In comparison, C2 can be seen as a safety requirement, as the robots using too much of their batteries could lead to a human having to come into the hazardous domain to collect them. At the same time, C2 can be seen as a functional requirement, to finish the patrol with the maximum amount of battery remaining. In our example, each action has a corresponding battery cost.

The part of this domain which makes which categorises it as a hazardous domain is the room labelled with the number 4. We suggest that this room is highly irradiated, and while this room should be visited three times, the amount of time that the robots spend within this room should be minimised. Therefore, an additional safety constraint can be added as C3:

- C3: The amount of time spent in room 4 should be minimised

Radiation is a serious safety hazard, that can be detrimental to robotic agents, and human agents alike, given our domain, minimising radiation exposure is a natural safety constraint.

As previously stated, each action has a corresponding value; for example, if a robot were in room three, it would have six available actions available to it. These all being movements to the other rooms via different paths. These paths will be more or less energy efficient based on the distance of travel.

Table 1: Options for entering and leaving room 4 and the corresponding risk of damage.

Entrance	**Exit**	**Exposure Time**	**Risk**
Hallway A	Hallway A	30 (seconds)	0.03
Hallway A	Hallway D	34 (seconds)	0.04
Hallway D	Hallway D	46 (seconds)	0.07
Hallway D	Hallway A	34 (seconds)	0.04

As previously stated, room four has risk associated with it in the form of radiation exposure. The risk associated with these actions relates to the time the robots will spend within the room according to the distance of the path chosen. This is shown in Table 1 and along with the battery costs, were created as reward structures within the PRISM language.

This domain was abstracted into an AMDP, which is common practice within safety engineering, as stated previously. From this abstraction, the rooms, movement actions between rooms, battery usage, and information on risk was left untouched, while everything else was removed from the model.

The functional constraints, C1 and C2 and safety constraints C3 were formally expressed using PCTL and were then used to verify the abstracted MDP using the probabilistic model checker known as PRISM.

After running QV on the AMDP expressed within PRISM, a single policy was synthesised that satisfied both the functional and safety constraints.

This policy was then used to constrain the MAS within the constructed domain, with tasks divided between the agents within the system based on this policy, with the two agents sharing the responsibility of two rooms, and having two rooms each which they are solely responsible. As well as actions being constrained, which stopped the agents needlessly moving around the domain, actions were also constrained according to the policy which caused the agents to needlessly enter into the hazardous area.

The evaluation experiments that were run within this domain were very promising, with the assured MARL constrained approach satisfying all of the constraints, with significantly less battery usage than the experiments run without the assured MARL approach, and in significantly fewer learning episodes.

6.2 Ongoing Work

Work which is presently being undertaken is the continued adaption of the approach to a different domain problem with greater complexity. This will be taking place within the physics engine known as Mujoco (Todorov et al., 2012). This will likely take the

form of a domain which will be ideal for deep learning, fulfilling steps (4 and5). This will allow extensive evaluation of the approach, adding to the validity of all of the objectives which this project aims to complete.

6.3 Future Work

Future work will involve more extensive experimentation of the approach under different circumstances. This will be an ideal way to test how well the approach scales, to different domain sizes and complexity, different system sizes, and learning techniques. These additional experiments will be able to determine to what extend objective one has been achieved.

ACKNOWLEDGEMENTS

This paper presents research sponsored by the UK MOD. The information contained in it should not be interpreted as representing the views of the UK MOD, nor should it be assumed it reflects any current or future UK MOD policy.

The author is grateful to Alec Banks, Radu Calinescu, Daniel Kudenko and Colin Paterson for their guidance and support with this project.

REFERENCES

Biyik, E., Margoliash, J., Alimo, S. R., and Sadigh, D. (2019). Efficient and safe exploration in deterministic markov decision processes with unknown transition models. In *American Control Conference*, pages 1792–1799.

Bloembergen, D., Tuyls, K., Hennes, D., and Kaisers, M. (2015). Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697.

Bogue, R. (2011). Robots in the nuclear industry: a review of technologies and applications. *Industrial Robot: An International Journal*.

Boutilier, C. (1996). Planning, learning and coordination in multiagent decision processes. In *18th Theoretical aspects of rationality and knowledge*, pages 195–210.

Buşoniu, L., Babuška, R., and De Schutter, B. (2010). Multi-agent reinforcement learning: An overview. In *Innovations in multi-agent systems and applications-1*, pages 183–221.

Calinescu, R., Autili, M., Cámara, J., Di Marco, A., Gerasimou, S., Inverardi, P., Perucci, A., Jansen, N., Katoen, J.-P., Kwiatkowska, M., Mengshoel, O. J., Spalazzese, R., and Tivoli, M. (2017). Synthesis and verification of self-aware computing systems. In Kounev, S., Kephart, J. O., Milenkoski, A., and Zhu, X., editors, *Self-Aware Computing Systems*, pages 337–373, Cham. Springer International Publishing.

Cheng, R., Khojasteh, M. J., Ames, A. D., and Burdick, J. W. (2020). Safe multi-agent interaction through robust control barrier functions with learned uncertainties. *arXiv preprint arXiv:2004.05273*.

Ciesinski, F. and Größer, M. (2004). On probabilistic computation tree logic. In *Validation of Stochastic Systems*, pages 147–188.

Cizelj, I., Ding, X. C. D., Lahijanian, M., Pinto, A., and Belta, C. (2011). Probabilistically safe vehicle control in a hostile environment. *IFAC Proceedings Volumes*, 44(1):11803–11808.

Fan, Y., Feng, G., Wang, Y., and Qiu, J. (2011). A novel approach to coordination of multiple robots with communication failures via proximity graph. *Automatica*, 47(8):1800–1805.

Garcia, J. and Fernández, F. (2012). Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research*, 45:515–564.

Garcia, J. and Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480.

Gerasimou, S., Calinescu, R., Shevtsov, S., and Weyns, D. (2017). UNDERSEA: an exemplar for engineering self-adaptive unmanned underwater vehicles. In *2017 IEEE/ACM 12th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*, pages 83–89. IEEE.

Gerasimou, S., Calinescu, R., and Tamburrelli, G. (2018). Synthesis of probabilistic models for quality-of-service software engineering. *Automated Software Engineering*, 25(4):785–831.

Gregory, J., Fink, J., Stump, E., Twigg, J., Rogers, J., Baran, D., Fung, N., and Young, S. (2016). Application of multi-robot systems to disaster-relief scenarios with limited communication. In *Field and Service Robotics*, pages 639–653. Springer.

Guo, J. (2020). Decentralized deep reinforcement learning for network level traffic signal control. *arXiv preprint arXiv:2007.03433*.

Khan, A., Zhang, C., Li, S., Wu, J., Schlotfeldt, B., Tang, S. Y., Ribeiro, A., Bastani, O., and Kumar, V. (2019). Learning safe unlabeled multi-robot planning with motion constraints. *arXiv preprint arXiv:1907.05300*.

Kroening, D., Abate, A., and Hasanbeig, M. (2020). Towards verifiable and safe model-free reinforcement learning. CEUR Workshop Proceedings.

Lemos, L. L., Bazzan, A. L., and Pasin, M. (2018). Co-adaptive reinforcement learning in microscopic traffic systems. In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE.

Mason, G. R., Calinescu, R. C., Kudenko, D., and Banks, A. (2017). Assured reinforcement learning with formally verified abstract policies. In *9th International Conference on Agents and Artificial Intelligence (ICAART)*, pages 105–117. SciTePress.

Moldovan, T. M. (2012). Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810*.

Parker, D. and Norman, G. (2014). Quantitative verification: Formal guarantees for timeliness reliability and performance. *a Knowledge Transfer Report from the London Mathematical Society and Smith Institute for Industrial Mathematics and System Engineering*.

Patel, P. G., Carver, N., and Rahimi, S. (2011). Tuning computer gaming agents using q-learning. In *2011 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 581–588.

Portugal, D., Iocchi, L., and Farinelli, A. (2019). A ros-based framework for simulation and benchmarking of multi-robot patrolling algorithms. In *Robot Operating System (ROS)*, pages 3–28.

Rasheed, F., Yau, K.-L. A., and Low, Y.-C. (2020). Deep reinforcement learning for traffic signal control under disturbances: A case study on sunway city, malaysia. *Future Generation Computer Systems*.

Riley, J., Calinescu, R., Paterson, C., Kudenko, D., and Banks, A. (2020). Reinforcement learning with quantitative verification for assured multi-agent policies. In *13th International Conference on Agents and Artificial Intelligence*. York.

Rizk, Y., Awad, M., and Tunstel, E. W. (2019). Cooperative heterogeneous multi-robot systems: a survey. *ACM Computing Surveys (CSUR)*, 52(2):1–31.

Schwager, M., Dames, P., Rus, D., and Kumar, V. (2017). A multi-robot control policy for information gathering in the presence of unknown hazards. In *Robotics research*, pages 455–472. Springer.

Serrano-Cuevas, J., Morales, E. F., and Hernández-Leal, P. (2019). Safe reinforcement learning using risk mapping by similarity.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2016a). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Shalev-Shwartz, S., Shammah, S., and Shashua, A. (2016b). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*.

Todorov, E., Erez, T., and Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE.

Wang, L., Ames, A., and Egerstedt, M. (2016). Safety barrier certificates for heterogeneous multi-robot systems. In *2016 American Control Conference (ACC)*, pages 5213–5218. IEEE.

Yasin, J. N., Mohamed, S. A., Haghbayan, M.-H., Heikkonen, J., Tenhunen, H., and Plosila, J. (2020). Navigation of autonomous swarm of drones using translational coordinates. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 353–362. Springer.

Zhang, W., Bastani, O., and Kumar, V. (2019). Mamps: Safe multi-agent reinforcement learning via model predictive shielding. *arXiv preprint arXiv:1910.12639*.