Safe Multi-Agent Learning via Shielding in Decentralized Environments

Doctoral Consortium

Daniel Melcer
Northeastern University
Boston, MA, USA
melcer.d@northeastern.edu

ABSTRACT

Multi-Agent Reinforcement Learning can be used to learn solutions for a wide variety of tasks, but there are few safety guarantees about the policies that the agents learn. My research addresses the challenge of ensuring safety in communication-free multi-agent environments, using shielding as the primary tool. We introduce methods to completely prevent safety violations in domains for which a model is available, in both fully observable and partially observable environments. We present ongoing research on maximizing safety in environments for which no model is available, utilizing a centralized training, decentralized execution framework, and discuss future lines of research.

KEYWORDS

Multi-Agent Reinforcement Learning; Safety; Shielding

ACM Reference Format:
Daniel Melcer. 2025. Safe Multi-Agent Learning via Shielding in Decentralized Environments: Doctoral Consortium. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Multi-Agent Reinforcement Learning has gained prominence as a method for solving a variety of tasks, such as RTS games [13], optimizing warehouse robot logistics [9], and robotic soccer [10].

However, many current reinforcement learning methods are not suitable for safety-critical domains. A misspecified reward function could lead to "reward hacking" [2, 5]. Even in environments without hackable reward functions, balancing safety with the agent's objective is difficult—choosing a too high penalty for safety violations may lead to the agent learning to not move at all. Even a successful learned policy is a black box, with no behavioral guarantees.

In contrast, the field of reactive synthesis [6], a subfield of formal methods, provides tools to synthesize systems that are guaranteed to adhere to a given safety specification. However, despite the many optimizations present in tool implementations, many techniques from the formal methods literature still struggle with scalability.

My Ph.D. research focuses on combining the scalability of reinforcement learning and the guarantees of formal methods to achieve safe multi-agent reinforcement learning.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

In particular, we focus on communication-free domains, as reliable communication is not always available. Even in environments where agents usually have access to some method of communication, it is vital to develop a safe backup policy in the event of hardware failure or radio interference.

We choose shielding [1] as a well-understood and simple to implement base to build off of. Each time step, the *shield* computes a set of safe actions for the agent to take. Depending on the exact formulation—each implementation shares the same theoretical guarantees—the agent selects an action from this safe set, the agent ranks each action and the shield chooses the highest-ranked safe action, or the agent proposes an action and the shield intervenes to replace the action if it is unsafe. In the multi-agent setting, each agent should have its own shield. If there is no communication between the agents, then there is no communication between the shields either. The challenge is to construct a set of shields, one per agent, such that as long as each agent acts according to its own shield, the joint behavior of all agents in their shared environment follows a given safety specification.

We analyze the problem domain along two dimensions:

(1) **Model Availability.** In certain environments, a human-provided abstraction of the environment may be available. If so, it is often possible to utilize this model from the start, to completely prevent safety violations even during training. Otherwise, the model must be learned, and safety violations may only be prevented on a best-effort basis.
(2) **Partial Observability.** Even if a model is available to denote safe actions at any given state, agents may not have enough information to determine the current environment state in a partially observable environment. Determining a safe action becomes a complex task when each agent in a multiagent system has a different set of possible states.

Our research has addressed both the fully observable [11] and partially observable [12] model-available domains. We are currently studying the fully observable model-unavailable domain, and plan to address the partially observable variant in future work.

2 THE LEAP TO MULTIPLE AGENTS

A recurring theme in both the fields of reinforcement learning and formal methods is that the addition of multiple agents—especially when those agents are unable to communicate with each other—dramatically increases the complexity of a problem. Single-agent reinforcement learning was used to train an agent that beat the world champion in Go nearly eight years ago [16], but until recently, multi-agent reinforcement learning methods struggled in the toy

Dec-Tiger domain [14]. Likewise, many problems in formal methods become undecidable when set in a decentralized domain [18].

Therefore, our first challenge was to extend shielding to the communication-free multiagent domain. While existing methods address scalability in multi-agent domains by paritioning the environment into small, independent areas [7], such methods require local communication to coordinate actions of multiple agents within one area, or the movement of agents between two areas.

Our initial extension assumes the full observability of all safety-relevant information, so all agents could agree on the current environment state and the set of safe joint actions, without needing communication. The setting is similar to how, upon approaching a stop sign, all agents (drivers) participating in the stop sign interaction have a shared understanding of the set of safe joint actions. In the driving case, all agents can navigate a stop sign based on a shared understanding of which agent receives priority, according to an ordering decided upon in advance.

Rather than requiring a human to create a set of rules for which agent has priority, our method first assigns priority to each agent in a deterministic manner, so that no communication is required for all agents to agree on an ordering. The individual shields then use a deterministic algorithm to iterate through all agents in order of priority, allowing the highest priority agents to perform as many individual actions as possible, while guaranteeing that at least one safe action remains available for low-priority agents. Our algorithm ensures that the result is *maximally permissive*; i.e. it is impossible to enable any additional individual actions to the result while maintaining the safety specification.

In our experiments [11], this approach allows reinforcement learning agents to successfully learn to solve a variety of gridworld-based tasks with zero safety violations, even during training.

3 PARTIAL OBSERVABILITY

Without full observability, any given observation may map to several possible ground-truth states. In certain cases, such as when agents have momentum but can only observe position, it is still possible for all agents to determine the state from only partial information (for example, by computing the difference in position from the previous time step), and so [11] still applies.

However, many partially-observable environments do not have this feature. In the single-agent partially observable case, it is possible to apply a shield based on the set of states that the agent believes the environment may inhabit (the belief support) [4, 8]. This is challenging to extend to the multi-agent domain, as one agent's belief support may not match the other agent's.

As an example, for some observation, Agent 1's belief support may include states A and C. Using the environment model, Agent 1 knows that if the environment really occupies state A, Agent 2's belief support will include states A and B; on the other hand, if the environment actually is in state C, Agent 2's belief support would include states C and D. Even though Agent 1 knows that the environment is *not* in states B or D, it must consider what its own belief support would be in such states, because Agent 2 must take into account Agent 1's beliefs when making its own decisions about what actions are safe. Clearly, this recursive line of reasoning can easily become unmanageable.

Our solution [12] is to avoid an agent-centric reasoning process, and to treat safe action selection as a global constraint satisfaction problem. We introduce three families of constraints: (1) There must always be a safe individual action available for every observation that an agent may encounter; (2) Unsafe joint actions are disallowed at specific environment states; and (3) An encoding of how joint actions at a specific environment state are related to individual actions for a given agent's observation. We encode all of these constraints as a boolean formula, and use a SAT solver [3] to find a set of safe individual actions for each observation. We then apply a post-processing step to these action sets to ensure that the set of enabled actions is maximally permissive. This entire process can occur ahead of time; the results are then provided to each agent, which can then each operate without any further communication.

If the SAT solver fails to find a shield, or if the shield is overly restrictive, we can allow the algorithm to take into account a bounded observation history. However, increasing the bounded history length too much can lead to an exponential growth in synthesis time. As this shield synthesis problem is closely related to the undecidable problem of decentralized reactive synthesis, it is impossible to construct a shield for every environment. Still, our method successfully constructs a shield for a variety of partially observable environments; reinforcement learning agents operating under these shields do not incur any safety violations, and generally learn to perform their tasks well.

4 LEARNING AN ENVIRONMENT MODEL

As previously noted, an environment model may not always be available in advance. Our ongoing research addresses this domain.

Several reinforcement learning methods for centralized training and decentralized execution (CTDE) operate by learning a centralized signal, such as the sum of expected future rewards—the Q-value—of a joint action, and automatically decomposing it into per-agent utility functions that follow the *Individual-Global-Max* principle [15, 17, 19]. These methods are structured such that if each agent independently chooses its highest-utility individual action, the joint action's Q-value is maximized.

We take inspiration from this structure, and adapt it to learning individual safety values for each agent, such that when each agent independently chooses an action that it believes is safe, the resulting joint action is safe. Preliminary results show that this is a promising direction for learning a task with fewer safety violations, compared to unshielded CTDE methods.

My goals for this research include further development of the theoretical aspects of shield learning in large environments without a provided model, as well as more extensive evaluations and comparisons of similar methods. Finally, we plan to extend this method to include environments with arbitrary partial observability.

ACKNOWLEDGMENTS

I would like to thank my advisors, Christopher Amato and Stavros Tripakis, for their guidance and advice. This material is based on work supported by NSF CCF award #2319500, *FMitF: Track I: Safe Multi-Agent Reinforcement Learning with Shielding*. This research was run in part on the Discovery cluster, supported by Northeastern University's Research Computing team.

REFERENCES

[1] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. 2018. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. AAAI Conference on Artificial Intelligence, New Orleans, LA, 10.

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. https://doi.org/10.48550/ARXIV.1606.06565

[3] Armin Biere, Katalin Fazekas, Mathias Fleury, and Maximillian Heisinger. 2020. CaDiCaL, Kissat, Paracooba, Plingeling and Treengeling Entering the SAT Competition 2020. In *Proc. of SAT Competition 2020 – Solver and Benchmark Descriptions (Department of Computer Science Report Series B, Vol. B-2020-1)*, Tomas Balyo, Nils Froleyks, Marijn Heule, Markus Iser, Matti Järvisalo, and Martin Suda (Eds.). University of Helsinki, Alghero, Italy, 51–53.

[4] Steven Carr, Nils Jansen, Sebastian Junges, and Ufuk Topcu. 2022. Safe Reinforcement Learning via Shielding under Partial Observability. arXiv:2204.00755 [cs.AI] https://arxiv.org/abs/2204.00755

[5] Jack Clark and Dario Amodei. 2016. Faulty reward functions in the wild. https://openai.com/research/faulty-reward-functions

[6] Edmund M. Clarke, Thomas A. Henzinger, Helmut Veith, and Roderick Bloem (Eds.). 2018. *Handbook of Model Checking*. Springer International Publishing, New York, NY. https://doi.org/10.1007/978-3-319-10575-8

[7] Ingy ElSayed-Aly, Suda Bharadwaj, Christopher Amato, Rüdiger Ehlers, Ufuk Topcu, and Lu Feng. 2021. Safe Multi-Agent Reinforcement Learning via Shielding. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems* (Virtual Event, United Kingdom) *(AAMAS '21)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 483–491.

[8] Sebastian Junges, Nils Jansen, and Sanjit A. Seshia. 2021. Enforcing Almost-Sure Reachability in POMDPs. In *Computer Aided Verification*, Alexandra Silva and K. Rustan M. Leino (Eds.). Springer International Publishing, Cham, 602–625.

[9] Aleksandar Krnjaic, Raul D. Steleac, Jonathan D. Thomas, Georgios Papoudakis, Lukas Schäfer, Andrew Wing Keung To, Kuan-Ho Lao, Murat Cubuktepe, Matthew Haley, Peter Börsting, and Stefano V. Albrecht. 2024. Scalable Multi-Agent Reinforcement Learning for Warehouse Logistics with Robotic and Human Co-Workers. arXiv:2212.11498 [cs.LG] https://arxiv.org/abs/2212.11498

[10] Zichong Li, Filip Bjelonic, Victor Klemm, and Marco Hutter. 2024. MARLadona - Towards Cooperative Team Play Using Multi-Agent Reinforcement Learning. arXiv:2409.20326 [cs.MA] https://arxiv.org/abs/2409.20326

[11] Daniel Melcer, Christopher Amato, and Stavros Tripakis. 2022. Shield Decentralization for Safe Multi-Agent Reinforcement Learning. *Advances in Neural Information Processing Systems* 36 (2022), 13.

[12] Daniel Melcer, Christopher Amato, and Stavros Tripakis. 2024. Shield Decomposition for Safe Reinforcement Learning in General Partially Observable Multi-Agent Environments. In *Reinforcement Learning Conference*.

[13] OpenAI, :, Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique P. d. O. Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. arXiv:1912.06680 [cs.LG] https://arxiv.org/abs/1912.06680

[14] Thomy Phan, Fabian Ritz, Philipp Altmann, Maximilian Zorn, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. 2023. Attention-Based Recurrence for Multi-Agent Reinforcement Learning under Stochastic Partial Observability. arXiv:2301.01649 [cs.MA] https://arxiv.org/abs/2301.01649

[15] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *J. Mach. Learn. Res.* 21, 1, Article 178 (jan 2020), 51 pages.

[16] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676 (Oct. 2017), 354–359. https://doi.org/10.1038/nature24270

[17] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2017. Value-Decomposition Networks For Cooperative Multi-Agent Learning. arXiv:1706.05296 [cs.AI] https://arxiv.org/abs/1706.05296

[18] Stavros Tripakis. 2004. Undecidable Problems of Decentralized Observation and Control on Regular Languages. *Inform. Process. Lett.* 90, 1 (April 2004), 21–28. https://doi.org/10.1016/j.ipl.2004.01.004

[19] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2020. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062* (2020).