How much can change in a year? Revisiting Evaluation in Multi-Agent Reinforcement Learning

Siddarth Singh[1], Omayma Mahjoub[1], Ruan de Kock[1], Wiem Khlifi[1, 2], Abidine Vall[3], Kale-ab Tessera[4], Arnu Pretorius[1]

[1]InstaDeep Ltd
[2]National School of Computer Science, Tunisia
[3]National School of Engineering of Tunis
[4]University of Edinburgh

Abstract

Establishing sound experimental standards and rigour is important in any growing field of research. Deep Multi-Agent Reinforcement Learning (MARL) is one such nascent field. Although exciting progress has been made, MARL has recently come under scrutiny for replicability issues and a lack of standardised evaluation methodology, specifically in the cooperative setting. Although protocols have been proposed to help alleviate the issue, it remains important to actively monitor the health of the field. In this work, we extend the database of evaluation methodology previously published by (Gorsane et al. 2022) containing meta-data on MARL publications from top-rated conferences and compare the findings extracted from this updated database to the trends identified in their work. Our analysis shows that many of the worrying trends in performance reporting remain. This includes the omission of uncertainty quantification, not reporting all relevant evaluation details and a narrowing of algorithmic development classes. Promisingly, we do observe a trend towards more difficult scenarios in SMAC-v1, which if continued into SMAC-v2 (Ellis et al. 2022) will encourage novel algorithmic development. Our data indicate that replicability needs to be approached more proactively by the MARL community to ensure trust in the field as we move towards exciting new frontiers.

Introduction

Multi-Agent Reinforcement Learning (MARL) is a rapidly growing field that has gained more attention in recent years, with large amounts of new algorithms in development which have shown the ability to solve complex problem solving tasks like Starcraft (Samvelyan et al. 2019), Hanabi (Foerster et al. 2019; Hu and Foerster 2021) and recently Diplomacy by integrating new developments from other fields like LLMs (Bakhtin et al. 2022). Given these successes there is a growing interest in utilising MARL to solve real-world problems like resource allocation, management and sharing, network routing and traffic signal controls (Vidhate and Kulkarni 2017; Brittain and Wei 2019; Nasir and Guo 2019; Spatharis et al. 2019; Liu et al. 2020; Zhao, Liu, and Cheng 2020; Pretorius et al. 2020; Gu et al. 2021) as these problems naturally lend to being formulated as cooperative MARL systems where a team of agents coordinates to optimise a shared global reward. However, similar to other fields within machine learning, ensuring a high level of scientific rigour and sound experimental methodology has become difficult as algorithms become more complex and computational requirements more extreme. As we move away from synthetic problems like video games towards tasks with real-world repercussions, it is important to ensure that trust in the field as a whole is not eroded over time by limited explainability of our results. Fortunately, these issues have been identified as they appeared and multiple works have suggested approaches to improve this and identify areas of concern before they become too widespread (Colas, Sigaud, and Oudeyer 2018; Henderson et al. 2018; Colas, Sigaud, and Oudeyer 2019; Engstrom et al. 2020; Jordan et al. 2020; Agarwal et al. 2022). An important first step in encouraging trust in the field of MARL is clarity and an accessible means for people to be able to determine the health of the field without requiring individual data gathering. To this end (Gorsane et al. 2022) compiled a dataset of evaluation methodologies used in deep cooperative MARL, from the first paper published in January 2016 up until April 2022. This database is open source and with only a small amount of database manipulation, it is possible on the reader's side to evaluate the health of MARL and determine popular algorithms, settings and interesting new avenues for development. Importantly, it also allows readers to determine which publications have made a strong effort to ensure high rigour and replicability. The authors discovered some worrying trends in research w.r.t. the consistency of reported results. As the number of publications grows in the coming years, it is important to update this dataset with new publications and perform a continuous assessment of the field. Although the original analysis of (Gorsane et al. 2022) highlighted many issues, there was evidence of a more thorough evaluation beginning to gain traction in more recent years, with an increase in the number of ablation studies and an upward trend in the use of uncertainty quantification.

In this paper, we treat the original dataset compiled by (Gorsane et al. 2022) as the historical trends of MARL evaluation and compare this to recent developments since the original publication. We compile a new dataset from the period of April 2022 to December 2022 which covers 29 papers. These papers are sourced under the same standards as the historical dataset and only include works published at high tier conferences including NeurIPS, ICLR and ICML with a

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

focus on deep cooperative MARL. Using this dataset of recent trends, we show in which areas evaluation methodology has improved or stagnated, highlight changes in algorithmic and environment popularity and provide recommendations[1].

Background

Algorithm Training. In cooperative MARL, the majority of algorithms can broadly be split into 2 paradigms; Centralised Training Decentralised Execution (CTDE) and Decentralised Training Decentralised Execution (DTDE). DTDE methods like Independent Q-Learning (Tan 1993) train agents using a policy based purely on their own egocentric experiences in the training environment. These methods are often called Independent Learners (IL) (Tan 1997) as each agent is trained without direct input from the other agents in the coalition. CTDE methods incorporate a centralised critic into their architecture at training time which is trained to perform multi-agent credit assignment and generates more informative reward signals for the agent in the coalition (Sunehag et al. 2017; Rashid et al. 2018; Yu et al. 2021). This centralised critic is only used during training time. During execution time agents in the coalition act in an ad-hoc or decentralised manner like in the DTDE case.

Value Decomposition. In multi-agent systems, typically a global shared reward signal is generated. Training a MARL system from this signal is difficult as we cannot determine the contribution of each member of the coalition from the environment feedback which makes it hard to evaluate how effective each member's learnt policy is. Value decomposition network (VDN) (Sunehag et al. 2017) introduced a new paradigm where a centralised critic is used to split the joint signal into individual agent rewards for better credit assignment. This method has proved very effective for Q-Learning based algorithms and became the most common avenue of development for new cooperative MARL methods (Rashid et al. 2018, 2020; Mahajan et al. 2020; Wang et al. 2020, 2021). For policy gradient (PG) methods, multi-agent proximal policy optimisation (MAPPO) (Yu et al. 2021) has been shown to be highly effective in many cooperative settings. However, while it falls under the CTDE paradigm, it does not make use of value decomposition.

Historical trends vs new data

Depreciation of legacy algorithms. The analysis of recent data indicates that, the landscape of common algorithms has shifted greatly from historical analysis as illustrated in figure 1. Historic algorithms like COMA (Foerster et al. 2018) and MADDPG (Lowe et al. 2017) have lost popularity as baselines with none of the reviewed papers making use of COMA and MADDPG used in only 25.8% of new publications vs a historical use of 35.2%. As these methods are unable to achieve reasonable performance on current common benchmarks their decline reflects researchers moving towards more modern baselines (Papoudakis et al. 2021). Qmix (Rashid et al. 2018) remains a strong baseline for newer cooperative

[1]The updated dataset is available at the following https://drive.google.com/file/d/1psSUrS8ywPAJY1Qm4ySfTYcYYJs0TlLF

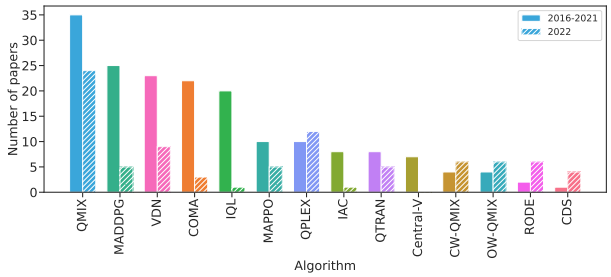

Figure 1: *Comparison between occurrences of algorithms from past years against the most recent year of full data.*

MARL algorithms which take inspiration from it. Importantly, Qmix is still competitive with newer methods when parameterised correctly and maintains relevance on modern benchmarks (Hu et al. 2021).

Despite the discovery of the effectiveness of Proximal Policy Optimisation (PPO) (Schulman et al. 2017) to match the more complex value decomposition methods in performance and sample efficiency (Yu et al. 2021), PG approaches still seem unpopular. With both the CTDE and DTDE variants of the Advantage Actor-Critic (A2C) (Mnih et al. 2016); Independent Actor-Critic (IAC) and Central-V (Foerster et al. 2018) not being present in newer publications. However MAPPO was still used in 16.1% of the new papers and will hopefully gain traction as a common baseline. Another notable development is the decline of Independent Learners (IL). Although normally used as baselines, the findings in (Papoudakis et al. 2021) show that they are important to assess the trade-offs newer method may have in certain settings.

Inconsistencies in performance for well-studied settings. In the early years of MARL, there was naturally a high variance in results even when using the same algorithm. Initially standardised frameworks were uncommon and MARL algorithms suffered from high implementation variance. Fortunately, MARL experienced some standardisation with the release of the the Starcraft Multi-Agent Challenge (SMAC) (Samvelyan et al. 2019) along with the EPyMARL framework (Papoudakis et al. 2021). Since then it has become and remains the most common benchmark in cooperative MARL by a sizable margin as shown in figure 5. However, even with this framework and common testing environment there was still high variance in reported performance as shown in figure 2a. For the most recent data in Figure 2b we can see that these trends persist. Seemingly, variance has only reduced significantly on the 'easy' category of scenarios like 8m which are trivial for newer algorithms. In practice this makes it very difficult to determine the true performance gains of recently developed and future works against the historical baselines they aim to improve on.

Tends in performance reporting. Metric reporting reached a promising peak of 75% in 2021 and seemed to be a rising trend over previous years. However we see this drop to 63% in 2022 in figure 3a. This is alarming given the goal of using MARL in real-world settings where reliability is often more important than absolute performance.

Performance Aggregation. Additionally, aggregation met-

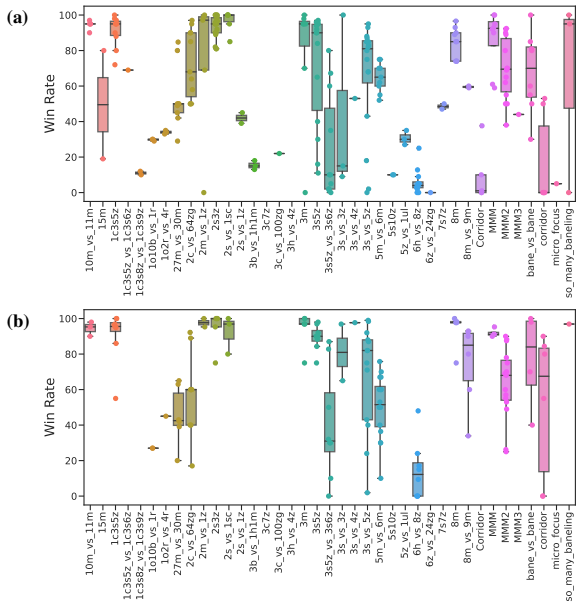

Figure 2: *Comparison between the performance of the Qmix algorithm from 2016-2021 against 2022.* **(a)** Historic performance spread for Qmix in SMAC (2016-2021). **(b)** Recent performance spread for Qmix in SMAC (2022).

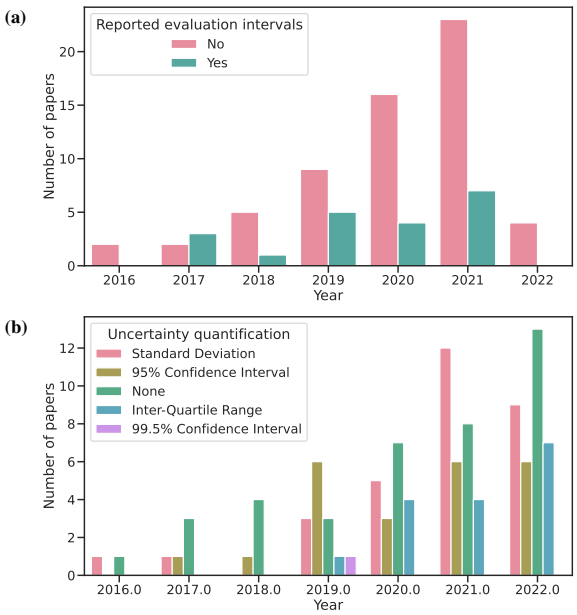

Figure 3: *Historical tendencies for reporting evaluation runs and uncertainty metrics from 2016 to 2022* **(a)** Tendencies for reporting evaluation runs (2016-2022). **(b)** Tendencies for reporting uncertainty metrics (2016-2022).

rics are not present in 20% of recent publications and with the computational complexity of modern MARL it is very difficult to evaluate over enough seeds to account for the variance in performance over different runs (Agarwal et al. 2022).

Measurements of spread and uncertainty. From figures 4c and 4d, it is clear that the tendency to display a measurement of spread has not meaningfully changed going from 33.8% to 31% respectively. Similarly, from figure 3b it is clear that the number of papers not reporting uncertainty are also increasing relative to the total number of published papers each year. This statistic is alarming most deep RL publications rarely use enough environment seeds to account for algorithmic variance during training (Agarwal et al. 2022). Further, for many practical tasks where high safety is a priority, methods with highly inconsistent behaviors are not usable which hampers the development of algorithms that have real-world utility.

Environment usage trends. As mentioned previously, SMAC is the most popular environment in MARL by a large margin followed by MPE which can be seen in figure 5. Evidence suggests that these settings have both reached a point of being over-fit to (Papoudakis et al. 2021; Gorsane et al. 2022) which makes determining algorithmic ranking difficult. Newer settings have generally not gained much popularity and despite the insight RWARE and LBF were shown to provide on MARL algorithms by (Papoudakis et al. 2021) recent data shows they are underutilised. However, environment overfitting on SMAC may be naturally resolved as users move over to SMAC-v2 (Ellis et al. 2022) which provides a new host of challenges through a well-understood interface. Given these findings it may be more effective to make use of explainability frameworks like ShinRL (Kitamura and Yonetani 2021) to determine a human interpreble representation of the learnt policy and from this compare the viability and variability of different methods. Even without a fully-featured framework, using methods like those described in (Agogino and Tumer 2008; Albrecht and Stone 2019; Boggess, Kraus, and Feng 2022) we can uncover exactly what different settings are testing and the limitations of current MARL methods that are difficult to detect from pure performance plots.

SMAC scenario usage. SMAC categorises scenarios into ‘easy‘, ‘hard‘ and ‘super hard‘ difficulties. Over time, scenarios not in the ‘super hard‘ category have become trivial to solve. Due to the large number of scenarios in SMAC, authors will often use a subset of the scenarios to reduce compute time which produces misleading aggregate performance (Gorsane et al. 2022). A helpful development seen in figure 6 is the reduction in use of the ‘easy‘ scenarios as a focus on difficult challenges streamlines the total set of evaluation scenarios required for comparative evaluation. (Gorsane et al. 2022) advocate for environment creators to propose a minimal viable set of scenarios for their settings along with an evaluation guideline to improve evaluation, however, when done by hand this induces human bias and until a setting has been overfit to it can be difficult for authors to determine the relevancy of each scenario. By using methods that explain environment features like (Hu et al. 2022) we can disentangle what competencies are required by each scenario and what exactly what is needed for the algorithm to solve the setting.

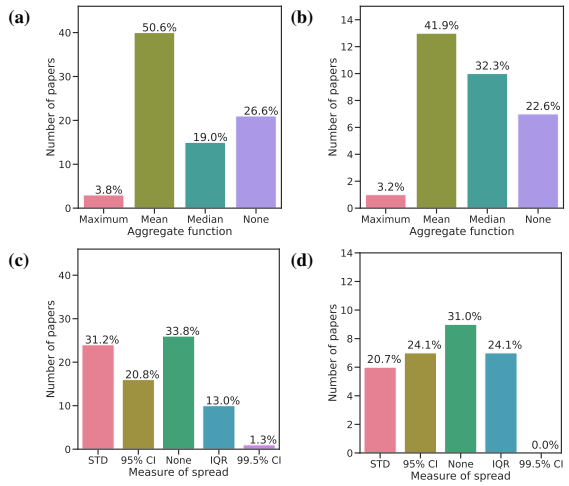

Figure 4: *Historical tendencies for reporting performance aggregation and variance from 2016 to 2022* **(a)** Historical aggregation usage (2016-2021). **(b)** Recent aggregation usage (2022). **(c)** Historical measure of spread (2016-2021). **(d)** Recent measure of spread (2022).

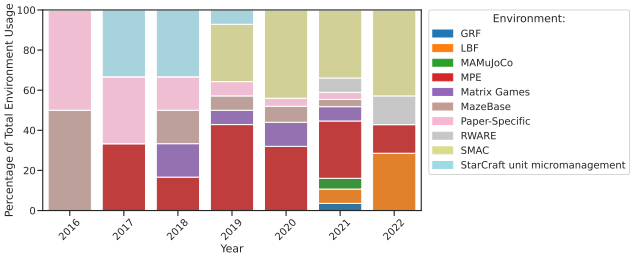

Figure 5: Historic environment usage from 2016 to 2022

Recent positive developments

On environment overfitting and computational expense. Along with the signs of overfitting, the authors of SMAC-v1 have done an analysis of their environment and discovered that due to its deterministic nature, it can be solved purely by memorising timesteps (Ellis et al. 2022). They have noted that algorithms designed for this type of setting at not suitable for the stochastic nature of real-world problems and have noted how in the single-agent setting procedural content generation (pcg) has gained popularity in resolving this issue (Juliani et al. 2019; Küttler et al. 2020; Cobbe et al. 2020; Samvelyan et al. 2021; Chevalier-Boisvert et al. 2023). However, although SMAC-v2 does resolve the limitations of SMAC-v1 w.r.t. generalisation and overfitting, it can still be computationally expensive to conform to the required number of runs outlined in works like (Agarwal et al. 2022; Gorsane et al. 2022) as it relies on a full-fledged video game as a back end. Fortunately there have been some promising developments in using JAX (Bradbury et al. 2018) for RL applications. Most notably JAX-based RL settings (Freeman et al. 2021; Lange 2022; Bonnet et al. 2023), which can be run on computational-accelerators like TPUs and GPUs. This

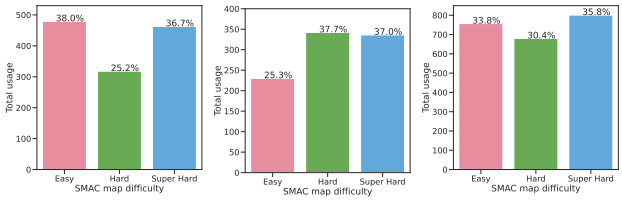

Figure 6: *Historical SMAC environment difficulty usage since release* **(Left)** Historical SMAC scenario usage (2018-2021). **(Middle)** Recent SMAC scenario usage (2022). **(Right)** All SMAC scenario usage (2018-2022).

allows the entire RL training loop to be massively accelerated. Growth of these types of settings have great potential to increase the clarity of RL by making experiments replicable for researchers without access to massive amounts of compute and allow more democratised scrutiny of RL research as a whole.

Emerging XAI in MARL. Although there has been an increase in XAI literature for SARL (Juozapaitis et al. 2019; Madumal et al. 2020; Puiutta and Veith 2020; Glanois et al. 2021; Heuillet, Couthouis, and Díaz-Rodríguez 2021; Vouros 2022; Dazeley, Vamplew, and Cruz 2023) and there are now many methods, frameworks and environment suites to make RL human-interpretable, MARL has been more historically limited. However, recent works have attempted to use XAI to explain not only agent behaviors but also what features are present in different settings. Notably (Hu et al. 2022) make use of algorithm-agnostic metrics to determine which common implementation features in MARL effect the performance of algorithms in popular settings. Further development of methods like this will greatly improve the clarity of MARL algorithms and improve trust in the field when attempting to apply these methods to real-world settings. Additionally, they are able to provide us with clearer paths for which areas in MARL require attention when compared to raw performance metrics.

Conclusion

We compiled 29 papers from the latest MARL publications to extend the MARL publication database to take into account recent development trends.

Despite a push for increased rigour, certain worrying historical trends are still prevalent in MARL w.r.t. replicability. Qmix, despite it's longstanding use as a baseline, still shows large variance across publications. The details required to replicate evaluation methodologies are often unreported and IL baselines are absent. Together, these issues make it difficult to get a true measurement of algorithmic development against historic algorithms.

Additionally environment usage also seems to have not yet diversified and research is focused on SMAC and MPE. However, authors do seem to be moving towards a more streamlined set of scenarios which will reduce computational overhead. Given these trends it is important that replicability is approached proactively by the MARL community to ensure trust in the field is maintained as it grows.

References

Agarwal, R.; Schwarzer, M.; Castro, P. S.; Courville, A.; and Bellemare, M. G. 2022. Deep Reinforcement Learning at the Edge of the Statistical Precipice. arXiv:2108.13264.

Agogino, A. K.; and Tumer, K. 2008. Analyzing and visualizing multiagent rewards in dynamic and stochastic domains. *Autonomous Agents and Multi-Agent Systems*, 17: 320–338.

Albrecht, S. V.; and Stone, P. 2019. Reasoning about Hypothetical Agent Behaviours and their Parameters. arXiv:1906.11064.

Bakhtin, A.; Brown, N.; Dinan, E.; Farina, G.; Flaherty, C.; Fried, D.; Goff, A.; Gray, J.; Hu, H.; Jacob, A. P.; Komeili, M.; Konath, K.; Kwon, M.; Lerer, A.; Lewis, M.; Miller, A. H.; Mitts, S.; Renduchintala, A.; Roller, S.; Rowe, D.; Shi, W.; Spisak, J.; Wei, A.; Wu, D. J.; Zhang, H.; and Zijlstra, M. 2022. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378: 1067 – 1074.

Boggess, K.; Kraus, S.; and Feng, L. 2022. Toward Policy Explanations for Multi-Agent Reinforcement Learning. In *International Joint Conference on Artificial Intelligence*.

Bonnet, C.; Luo, D.; Byrne, D.; Surana, S.; Coyette, V.; Duckworth, P.; Midgley, L. I.; Kalloniatis, T.; Abramowitz, S.; Waters, C. N.; Smit, A. P.; Grinsztajn, N.; Sob, U. A. M.; Mahjoub, O.; Tegegn, E.; Mimouni, M. A.; Boige, R.; de Kock, R.; Furelos-Blanco, D.; Le, V.; Pretorius, A.; and Laterre, A. 2023. Jumanji: a Diverse Suite of Scalable Reinforcement Learning Environments in JAX. arXiv:2306.09884.

Bradbury, J.; Frostig, R.; Hawkins, P.; Johnson, M. J.; Leary, C.; Maclaurin, D.; Necula, G.; Paszke, A.; VanderPlas, J.; Wanderman-Milne, S.; and Zhang, Q. 2018. JAX: composable transformations of Python+NumPy programs.

Brittain, M.; and Wei, P. 2019. Autonomous Air Traffic Controller: A Deep Multi-Agent Reinforcement Learning Approach. arXiv:1905.01303.

Chevalier-Boisvert, M.; Dai, B.; Towers, M.; de Lazcano, R.; Willems, L.; Lahlou, S.; Pal, S.; Castro, P. S.; and Terry, J. 2023. Minigrid & Miniworld: Modular & Customizable Reinforcement Learning Environments for Goal-Oriented Tasks. *CoRR*, abs/2306.13831.

Cobbe, K.; Hesse, C.; Hilton, J.; and Schulman, J. 2020. Leveraging Procedural Generation to Benchmark Reinforcement Learning. arXiv:1912.01588.

Colas, C.; Sigaud, O.; and Oudeyer, P.-Y. 2018. How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments. arXiv:1806.08295.

Colas, C.; Sigaud, O.; and Oudeyer, P.-Y. 2019. A Hitchhiker's Guide to Statistical Comparisons of Reinforcement Learning Algorithms. arXiv:1904.06979.

Dazeley, R.; Vamplew, P.; and Cruz, F. 2023. Explainable reinforcement learning for broad-xai: a conceptual framework and survey. *Neural Computing and Applications*, 1–24.

Ellis, B.; Moalla, S.; Samvelyan, M.; Sun, M.; Mahajan, A.; Foerster, J. N.; and Whiteson, S. 2022. SMACv2: An Improved Benchmark for Cooperative Multi-Agent Reinforcement Learning.

Engstrom, L.; Ilyas, A.; Santurkar, S.; Tsipras, D.; Janoos, F.; Rudolph, L.; and Madry, A. 2020. Implementation Matters in Deep Policy Gradients: A Case Study on PPO and TRPO. arXiv:2005.12729.

Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Foerster, J. N.; Song, F.; Hughes, E.; Burch, N.; Dunning, I.; Whiteson, S.; Botvinick, M.; and Bowling, M. 2019. Bayesian Action Decoder for Deep Multi-Agent Reinforcement Learning. arXiv:1811.01458.

Freeman, C. D.; Frey, E.; Raichuk, A.; Girgin, S.; Mordatch, I.; and Bachem, O. 2021. Brax - A Differentiable Physics Engine for Large Scale Rigid Body Simulation.

Glanois, C.; Weng, P.; Zimmer, M.; Li, D.; Yang, T.; Hao, J.; and Liu, W. 2021. A survey on interpretable reinforcement learning. *arXiv preprint arXiv:2112.13112*.

Gorsane, R.; Mahjoub, O.; de Kock, R.; Dubb, R.; Singh, S.; and Pretorius, A. 2022. Towards a Standardised Performance Evaluation Protocol for Cooperative MARL. arXiv:2209.10485.

Gu, B.; Zhang, X.; Lin, Z.; and Alazab, M. 2021. Deep Multiagent Reinforcement-Learning-Based Resource Allocation for Internet of Controllable Things. *IEEE Internet of Things Journal*, 8(5): 3066–3074.

Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; and Meger, D. 2018. Deep Reinforcement Learning that Matters. arXiv:1709.06560.

Heuillet, A.; Couthouis, F.; and Díaz-Rodríguez, N. 2021. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214: 106685.

Hu, H.; and Foerster, J. N. 2021. Simplified Action Decoder for Deep Multi-Agent Reinforcement Learning. arXiv:1912.02288.

Hu, J.; Jiang, S.; Harding, S. A.; Wu, H.; and Liao, S.-w. 2021. Rethinking the Implementation Tricks and Monotonicity Constraint in Cooperative Multi-Agent Reinforcement Learning.

Hu, S.; Xie, C.; Liang, X.; and Chang, X. 2022. We have only. arXiv:2207.05683.

Jordan, S. M.; Chandak, Y.; Cohen, D.; Zhang, M.; and Thomas, P. S. 2020. Evaluating the Performance of Reinforcement Learning Algorithms. arXiv:2006.16958.

Juliani, A.; Khalifa, A.; Berges, V.-P.; Harper, J.; Teng, E.; Henry, H.; Crespi, A.; Togelius, J.; and Lange, D. 2019. Obstacle Tower: A Generalization Challenge in Vision, Control, and Planning. arXiv:1902.01378.

Juozapaitis, Z.; Koul, A.; Fern, A.; Erwig, M.; and Doshi-Velez, F. 2019. Explainable reinforcement learning via reward decomposition. In *IJCAI/ECAI Workshop on explainable artificial intelligence*.

Kitamura, T.; and Yonetani, R. 2021. ShinRL: A Library for Evaluating RL Algorithms from Theoretical and Practical Perspectives.

Küttler, H.; Nardelli, N.; Miller, A. H.; Raileanu, R.; Selvatici, M.; Grefenstette, E.; and Rocktäschel, T. 2020. The NetHack Learning Environment. arXiv:2006.13760.

Lange, R. T. 2022. gymnax: A JAX-based Reinforcement Learning Environment Library.

Liu, X.; Yu, J.; Feng, Z.; and Gao, Y. 2020. Multi-agent reinforcement learning for resource allocation in IoT networks with edge computing. *China Communications*, 17(9): 220–236.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *Neural Information Processing Systems (NIPS)*.

Madumal, P.; Miller, T.; Sonenberg, L.; and Vetere, F. 2020. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 2493–2500.

Mahajan, A.; Rashid, T.; Samvelyan, M.; and Whiteson, S. 2020. MAVEN: Multi-Agent Variational Exploration. arXiv:1910.07483.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T. P.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous Methods for Deep Reinforcement Learning. arXiv:1602.01783.

Nasir, Y. S.; and Guo, D. 2019. Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks. *IEEE Journal on Selected Areas in Communications*, 37(10): 2239–2250.

Papoudakis, G.; Christianos, F.; Schäfer, L.; and Albrecht, S. V. 2021. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. arXiv:2006.07869.

Pretorius, A.; Cameron, S.; Van Biljon, E.; Makkink, T.; Mawjee, S.; du Plessis, J.; Shock, J.; Laterre, A.; and Beguir, K. 2020. A game-theoretic analysis of networked system control for common-pool resource management using multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 9983–9994.

Puiutta, E.; and Veith, E. M. 2020. Explainable reinforcement learning: A survey. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, 77–95. Springer.

Rashid, T.; Farquhar, G.; Peng, B.; and Whiteson, S. 2020. Weighted QMIX: Expanding Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. arXiv:2006.10800.

Rashid, T.; Samvelyan, M.; de Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. arXiv:1803.11485.

Samvelyan, M.; Kirk, R.; Kurin, V.; Parker-Holder, J.; Jiang, M.; Hambro, E.; Petroni, F.; Küttler, H.; Grefenstette, E.; and Rocktäschel, T. 2021. MiniHack the Planet: A Sandbox for Open-Ended Reinforcement Learning Research. *ArXiv*, abs/2109.13202.

Samvelyan, M.; Rashid, T.; de Witt, C. S.; Farquhar, G.; Nardelli, N.; Rudner, T. G. J.; Hung, C.-M.; Torr, P. H. S.; Foerster, J.; and Whiteson, S. 2019. The StarCraft Multi-Agent Challenge.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Spatharis, C.; Blekas, K.; Bastas, A.; Kravaris, T.; and Vouros, G. A. 2019. Collaborative multiagent reinforcement learning schemes for air traffic management. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–8.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; and Graepel, T. 2017. Value-Decomposition Networks For Cooperative Multi-Agent Learning. arXiv:1706.05296.

Tan, M. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, 330–337.

Tan, M. 1997. Multi-Agent Reinforcement Learning: Independent versus Cooperative Agents. In *International Conference on Machine Learning*.

Vidhate, D. A.; and Kulkarni, P. 2017. Cooperative multi-agent reinforcement learning models (CMRLM) for intelligent traffic control. In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, 325–331.

Vouros, G. A. 2022. Explainable deep reinforcement learning: state of the art and challenges. *ACM Computing Surveys*, 55(5): 1–39.

Wang, J.; Ren, Z.; Liu, T.; Yu, Y.; and Zhang, C. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. arXiv:2008.01062.

Wang, T.; Gupta, T.; Mahajan, A.; Peng, B.; Whiteson, S.; and Zhang, C. 2020. RODE: Learning Roles to Decompose Multi-Agent Tasks. arXiv:2010.01523.

Yu, C.; Velu, A.; Vinitsky, E.; Wang, Y.; Bayen, A.; and Wu, Y. 2021. The Surprising Effectiveness of PPO in Cooperative, Multi-Agent Games.

Zhao, N.; Liu, Z.; and Cheng, Y. 2020. Multi-Agent Deep Reinforcement Learning for Trajectory Design and Power Allocation in Multi-UAV Networks. *IEEE Access*, 8: 139670–139679.