systems

MDPI

Article

It's All about Reward: Contrasting Joint Rewards and Individual Reward in Centralized Learning Decentralized Execution Algorithms

Peter Atrazhev and Petr Musilek *

Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada
* Correspondence: pmusilek@ualberta.ca

Abstract: This paper addresses the issue of choosing an appropriate reward function in multi-agent reinforcement learning. The traditional approach of using joint rewards for team performance is questioned due to a lack of theoretical backing. The authors explore the impact of changing the reward function from joint to individual on learning centralized decentralized execution algorithms in a Level-Based Foraging environment. Empirical results reveal that individual rewards contain more variance, but may have less bias compared to joint rewards. The findings show that different algorithms are affected differently, with value factorization methods and PPO-based methods taking advantage of the increased variance to achieve better performance. This study sheds light on the importance of considering the choice of a reward function and its impact on multi-agent reinforcement learning systems.

Keywords: agent coordination; multi-agent reinforcement learning; centralized learning decentralized execution

Citation: Atrazhev, P.; Musilek, P. It's All about Reward: Contrasting Joint Rewards and Individual Reward in Centralized Learning Decentralized Execution Algorithms. *Systems* **2023**, *11*, 180. https://doi.org/10.3390/systems11040180

Academic Editors: Philippe Mathieu, Juan M. Corchado, Alfonso González-Briones and Fernando De la Prieta Pintado

Received: 1 February 2023
Revised: 22 March 2023
Accepted: 23 March 2023
Published: 30 March 2023

1. Introduction

Multi-agent reinforcement learning (MARL) is a promising field of artificial intelligence (AI) research, and over the last couple of years, has seen increasingly more pushes to tackle less "toy" problems (full game environments such as ATARI and the Starcraft Multi-Agent Environment (SMAC)) and instead try to solve complex "real-world" problems [1–3]. Coordination of agents across a large state space is a challenging and multifaceted problem, with many approaches that can be used to increase coordination. These include communication between agents, both learned and established, parameter sharing and other methods of imparting additional information to function approximators, and increasing levels of centralization.

One paradigm of MARL that aims to increase coordination is called centralized Learning decentralized Execution (CLDE) [4]. CLDE algorithms train their agents' policies with additional global information using a centralized mechanism. During execution, the centralized element is removed, and the agent's policy is based on conditions only on local observations. This has been shown to increase the coordination of agents [5]. CLDE algorithms separate into two major categories: centralized policy gradient methods [6–8] and value decomposition methods [9,10]. Recently, however, there has been work that has put into question the assumption that centralized mechanisms do indeed increase coordination. Lyu et al. [11] found that in actor–critic systems, the use of a centralized critic led to an increase in variance seen in the final policy learned; however, they noted more coordinated agent behaviour while training and concluded that the use of a centralized critic should be thought of as a choice that carries with it a bias variance trade-off .

One aspect of agent coordination that is similarly often taken at face value is the use of a joint reward in cooperative systems that use centralization. The assumption is that joint rewards are necessary for the coordination of systems that rely on centralization. We

have not been able to find a theoretical basis for this claim. The closest works addressing team rewards in cooperative settings that we could find include works on difference rewards which try to measure the impact of an individual agent's actions on the full system reward [12]. The high learnability, among other nice properties, makes difference rewards attractive but impractical, due to the required knowledge of the total system state [13–15].

We investigate the effects of changing the reward from a joint reward to an individual reward in the Level-Based Foraging (LBF) environment. We investigate how different CLDE algorithm performances change as a result of this change and discuss this performance change. In this work, we study the effect of varying reward functions from joint rewards to individual rewards on Independent Q Learning (IQL) [16], Independent Proximal Policy Optimization (IPPO) [17], independent synchronous actor–critic (IA2C) [6], multi-agent proximal policy optimization (MAPPO) [7], multi agent synchronous actor–critic (MAA2C) [5,6], value decomposition networks (VDN) [10], and QMIX [9] when evaluated on the LBF environment [18]. This environment was chosen as it is a gridworld environment, and therefore simpler to understand when compared to other MARL environments such as those based on the StarCraft environment; however, it is a very challenging environment that requires cooperation to solve and has the ability to include the forcing of cooperative policies and partial observability for study.

We show empirically that using an individual reward in the LBF environment causes an increase in the variance in the reward term in the Temporal Difference (TD) error signal and any derivative of this term. We study the effects that this increase in variance has on the selected algorithms and discuss whether this variance is helpful for the learning of better joint policies in the LBF environment. Our results show that PPO-based algorithms, with and without centralized systems and QMIX, perform better with individual rewards, while actor–critic models based on A2C suffer when using individual rewards.

This work is comprised of multiple sections, starting with the background in Section 2. Section 3 outlines our experimental method, and we report our results in Section 4. We discuss the results and compare them to the previous results in Section 5. All supplementary information pertaining to this work can be found in the Appendices A–C.

2. Background

2.1. Dec-POMDPs

We define a fully cooperative task as a decentralized partially observable Markov decision process (Dec-POMDP) which consists of the tuple $M = < D, S, A, T, O, o, R, h, b_0 >$ [4]. Where D is the set of agents, S is the set that describes the true state of the environment, A is the joint action set over all agents, and T is the transition probability function, mapping the joint actions to state. O is the joint observation set, o represents the observation probability function, and R is the reward function which describes the set of all individual rewards for each agent $R = R^i_t$. The problem horizon, h, is equivalent to the discount factor γ in the RL literature. The initial state distribution is given by b_0. M describes a partially observable scenario in which agents interact with the environment through observations, without ever knowing the true state of the environment. When agents have full access to the state information, the tuple becomes $< D, S, A, T, R, h, b_0 >$ and is defined as *Multi-agent Markov Decision Process (MMDP)* [4].

2.2. Reward Functions

2.2.1. Joint Reward

The entire team receives a joint reward value at each time step taken as the sum of all individual agent rewards $R = R^i = \cdots = R^N = \sum_{i=1}^{N} R^i_t$. The joint reward has an interesting property that is usually left aside: by being the summation of all agents' rewards, if an agent is not participating in a reward event, they still receive a reward. This creates a small but nonzero probability for all agents to receive a reward in any state and for any action. In addition, in partially observable tasks, these reward events can occur with no context for some of the agents. The advantage of the joint reward is a salient signal

across all that can be learned from, as well as additional information about the performance of team members that may or may not be observable.

2.2.2. Individual Reward

Mixed tasks differ from the fully cooperative case only in terms of the reward received by the agents. Mixed tasks attribute individual rewards to each agent rather than a joint reward, making the term R in the tuple M, $R = R_t^i$ for each agent i. During reward events, a reward is only given to agents who participate in reward events. This reduces the saliency of the reward signal during a reward event, and can cause increased variance in the reward signal when different agents achieve a reward.

2.3. Level-Based Foraging

Level-Based Foraging (LBF) is a challenging exploration problem in which multiple agents must work together to collect food items scattered randomly on a gridworld [18]. The environment is highly configurable, allowing for partial observability and the use of cooperative policies only. In LBF, agents and food are assigned random levels, with the maximum food level always being the sum of all agent levels. Agents can take discrete actions, such as moving in a certain direction, loading food, or not taking any action. Agents receive rewards when they successfully load a food item, which is possible only if the sum of all agent levels around the food is equal to or greater than the level of the food item. Agent observations are discrete and include the location and level of all food and agents on the board, including themselves.

The LBF environment is highly configurable, starting with gridworld size, number of agents, and number of food items. The scenarios in the LBF are described using the following nomenclature: *NxM-Ap-Bf*, where *N* and *M* define the size of the gridworld, *A* indicates the number of agents, and *B* indicates the number of food objectives in the world. A 10 by 10 grid world with three agents and three food would be described as *10x10 -3p-3f*. Additionally, partial observability can be configured by adding *Cs-* before the grid size. *C* defines the radius size that agents can observe. For all objects outside the radius, the agent will receive a constant value of -1 in that observation. Finally, the addition of the *-coop* tag after the number of food causes the game to enforce that all agents must be present to collect food, thereby forcing cooperative policies to be the only policies that can be learned. As an example, an eight-by-eight gridworld with two players and two food that forces cooperative policies while subjecting the agents to partial observability with a radius of two would be specified as *2s-8x8-2p-2f-coop* . An example of the LBF gridworld is shown in Figure 1.

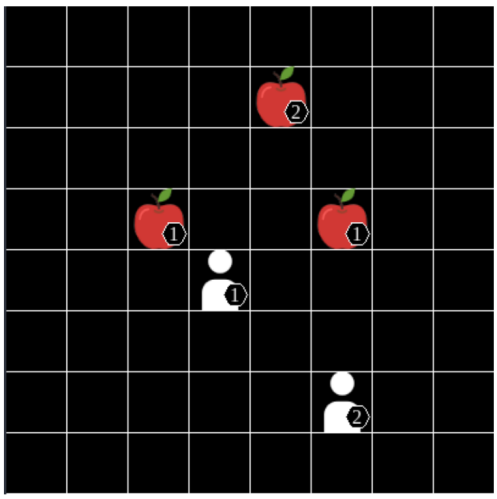

Figure 1. LBF Foraging-8x8-2p-3f example gridworld taken from Papoudakis et al. [5]

3. Method

To compare our results with those of previous publications, we made sure that the scenarios and scenario parameters matched those of Papoudakis et al. [5] and Atrazhev et al. [19], and the results were compared to the results of those previous works.

To remain consistent with previous publications, the LBF scenarios selected for this study are *8x8-2p-2f-coop*, *2s-8x8-2p-2f-coop*, *10x10-3p-3f*, and *2s-10x10-3p-3f*. Algorithms are also selected based on these criteria: IQL [16], IA2C [6], IPPO [17], MAA2C [5], MAPPO [7], VDN [10] and QMIX [9] were selected as they are studied in both Papoudakis et al. [5] and in Atrazhev et al. [19] and represent an acceptable assortment of independent algorithms, centralized critic CLDE algorithms, and value factorization CLDE algorithms.

To evaluate the performance of the algorithm, we calculate the average returns and maximum returns achieved throughout all evaluation windows during training, and the 95% confidence interval across ten seeds.

Our investigation consists of varying two variables, the reward function, and episode length. The length of the episode was varied between the reported value of 25 used by Papoudakis et al. [5] and 50, which is the default length of the episode in the environment. We perform two separate hyperparameter tunings, one for each reward type, adhering to the hyperparameter tuning protocol included in Papoudakis et al. [5].

All other experimental parameters are taken from Papoudakis et al. [5], and we encourage readers to look into this work for further details.

4. Results

We compare IQL, IA2C, IPPO, MAA2C, MAPPO, VDN, and QMIX and report the mean returns and max returns achieved by algorithms using individual rewards in Tables 1 and 2, respectively. The mean returns and maximum returns of algorithms using joint rewards are reported in Tables 3 and 4, respectively. We include tables for the increased episode length (50 timesteps) in the Appendix C.

Table 1. Maximum returns and 95% confidence interval of algorithms using individual rewards in selected scenarios over 10 seeds, after a hyperparameter search was completed. Bolded values indicate the best result in a scenario.

Scenario	IQL	IA2C	IPPO	MAA2C	MAPPO	VDN	QMIX
8x8-2p-2f-c	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.11**	**1.00 ± 0.00**
8x8-2p-2f-2s-c	0.97 ± 0.0	0.94 ± 0.01	0.95 ± 0.01	0.93 ± 0.01	0.93 ± 0.01	0.97 ± 0.0	**0.98 ± 0.00**
10x10-3p-3f	0.94 ± 0.02	0.86 ± 0.01	0.88 ± 0.04	0.85 ± 0.03	0.86 ± 0.02	0.93 ± 0.04	**0.95 ± 0.02**
10x10-3p-3f-2s	0.75 ± 0.01	0.71 ± 0.02	0.76 ± 0.02	0.7 ± 0.02	0.73 ± 0.07	0.74 ± 0.01	**0.77 ± 0.01**

Table 2. Mean return values and 95% confidence interval of algorithms using individual rewards in selected scenarios over 10 seeds, after a hyperparameter search was completed. Bolded values indicate the best result in a scenario.

Scenario	IQL	IA2C	IPPO	MAA2C	MAPPO	VDN	QMIX
8x8-2p-2f-c	0.78 ± 0.08	0.82 ± 0.02	**0.84 ± 0.07**	0.78 ± 0.05	0.77 ± 0.07	0.7 ± 0.08	0.75 ± 0.04
8x8-2p-2f-2s-c	0.83 ± 0.01	0.71 ± 0.01	0.77 ± 0.01	0.68 ± 0.01	0.69 ± 0.03	0.81 ± 0.01	**0.86 ± 0.01**
10x10-3p-3f	0.68 ± 0.02	0.7 ± 0.02	**0.72 ± 0.03**	0.66 ± 0.03	0.69 ± 0.02	0.55 ± 0.04	0.58 ± 0.03
10x10-3p-3f-2s	**0.62 ± 0.0**	0.55 ± 0.02	0.58 ± 0.02	0.51 ± 0.02	0.53 ± 0.06	0.57 ± 0.01	**0.62±0.01**

Table 3. Maximum returns and 95% confidence interval of algorithms using joint rewards in selected scenarios over 10 seeds, after a hyperparameter search was completed. Bolded values indicate the best result in a scenario.

Scenario	IQL	IA2C	IPPO	MAA2C	MAPPO	VDN	QMIX
8x8-2p-2f-c	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**
8x8-2p-2f-2s-c	0.97 ± 0.01	**1.0 ± 0.0**	0.63 ± 0.02	**1.0 ± 0.0**	0.56 ± 0.02	0.98 ± 0.0	0.97 ± 0.0
10x10-3p-3f	0.89 ± 0.08	**0.99 ± 0.01**	0.89 ± 0.02	0.98 ± 0.01	0.9 ± 0.24	0.9 ± 0.03	0.91 ± 0.02
10x10-3p-3f-2s	0.7 ± 0.01	**0.84 ± 0.04**	0.56 ± 0.01	**0.85 ± 0.01**	0.58 ± 0.01	0.77 ± 0.01	0.76 ± 0.04

Table 4. Mean return values and 95% confidence interval of algorithms using joint rewards in selected scenarios over 10 seeds, after a hyperparameter search was completed. Bolded values indicate the best result in a scenario.

Scenario	IQL	IA2C	IPPO	MAA2C	MAPPO	VDN	QMIX
8x8-2p-2f-c	0.77 ± 0.08	0.96 ± 0.01	0.96 ± 0.01	**0.97 ± 0.01**	0.96 ± 0.02	0.78 ± 0.04	0.69 ± 0.04
8x8-2p-2f-2s-c	0.82 ± 0.01	**0.94 ± 0.01**	0.39 ± 0.02	**0.94 ± 0.0**	0.45 ± 0.02	0.84 ± 0.0	0.79 ± 0.01
10x10-3p-3f	0.47 ± 0.07	**0.88 ± 0.02**	0.71 ± 0.03	0.87 ± 0.02	0.59 ± 0.21	0.56 ± 0.03	0.46 ± 0.04
10x10-3p-3f-2s	0.56 ± 0.01	0.67 ± 0.05	0.44 ± 0.0	**0.69 ± 0.02**	0.46 ± 0.0	0.6 ± 0.01	0.56 ± 0.05

We generally observe that in the individual reward case, QMIX is able to consistently achieve the highest maximal return value in all scenarios. In terms of the highest mean returns, QMIX is able to outperform IPPO in the partially observable scenarios. In the joint reward case, the majority of the results are in line with those reported in [5]; however, we note that the average return results for QMIX are much higher with our hyperparameters. We go into more detail regarding these results in Appendix A.

When comparing joint reward performance with individual reward performance, we note that the effects of reward are not easily predictable. Centralized critic algorithms are evenly split in performance, with MAPPO performing better with individual reward, while MAA2C's performance suffers. This is paralleled by the independent versions of MAPPO and MAA2C. The value factorization algorithms are also divided, with QMIX performance becoming the top-performing algorithm across the tested scenarios. VDN, however, sees an incredible drop in performance when using joint rewards. Finally, IQL performance when using individual reward is relatively unaffected in the simpler 8x8 scenarios but decreases in the larger scenarios.

5. Discussion

5.1. Independent Algorithms

5.1.1. IQL

Our results show that IQL achieves increased mean return values and maximum return values when using individual rewards. Our results also show that IQL experienced a reduction in loss variance when using individual rewards. Since IQL is an independent algorithm, the joint reward is the only source of information from other agents. Seeing that IQL does not observe the other agents specifically, our results suggest that the joint reward seems to increase the variance in the loss function by the nonzero probability of agents receiving the reward at any timestep, as discussed earlier. The reduction in variance in the loss function allows for better policies to be learned by each individual agent, and this is further evidenced by the reduction in variance and simultaneous increase in the mean of the absolute TD error that agents have in the CLBF experiments.

5.1.2. IPPO

IPPO is able to use the individual reward signal to achieve higher mean returns and maximum returns in all scenarios except for the 8x8-2p-2f-coop. We believe that this is in large part due to the decrease in variance that is observed in the maximum policy values that are learned. Our results show that the TD error that is generated from multiple different individual rewards appears to be higher and more varied than the TD error that is generated from a joint reward. This variance seems to permeate through the loss function, allowing the algorithm to continue discovering new higher policies through training. It seems that joint rewards cause the TD error to start out strong, and quickly the algorithm finds a policy (or set of policies) that has the maximal chances of achieving rewards at all timesteps. This is a local minimum, but the error is too small for policies to escape the minima.

5.1.3. IA2C

IA2C suffers from the increase in variance in individual rewards. We note evidence of divergent policy behaviour in a number of metrics, most notably the critic and policy gradient loss. The critic is still able to converge; however, the policy gradient loss diverges quite a bit more in the individual reward case. It seems that a joint reward is necessary to help coordinate the agent's behaviour.

5.2. Value Factorization Algorithms

5.2.1. VDN

VDN with individual rewards has a very rapid reduction in loss values. Our data suggest that when using individual rewards, VDN converges prematurely on suboptimal policies, causing the observed reduction in mean and max return. This may be due to the fact that VDN does not incorporate any state information into the creation of the joint value function. The authors seem to have relied on the information contained in the joint rewards to help guide the coordination of agents through the learned joint value function. With individual rewards, the joint action value function simply optimizes for the first policy that serves to maximize returns without regard for agent coordination or guiding agents to find optimal policies.

5.2.2. QMIX

Our results show that when individual rewards are used with qmix, return mean and maximum return values are increased. When comparing joint rewards to independent rewards, independent rewards show signs of faster convergence in loss and gradient norms. Qmix's combination of monotonicity constraints and global state information in its hypernetwork seems to be able to find coordinated policies when using individual rewards that achieve higher returns than those found when using joint rewards. By leveraging the global state information during training, the improvement shows significantly higher in the partially observable scenarios where the increased information builds stronger coordination between agents.

5.3. Centralized Critic Algorithms

Performance in centralized critics is varied and seems to depend on the underlying algorithm used.

5.3.1. MAA2C

The increase in information that is imparted by MAA2C's centralized critic seems to not be enough to counter the increase in variance that is caused by individual rewards. When using joint rewards, the critic is able to converge and is able to guide the actor policies to find optimal values relatively quickly, and is best demonstrated by the convergence of the TD error. When using individual rewards, there seems to be too much variance for the critic to be able to converge quickly. It has been shown that simply adding a centralized critic to an actor–critic MARL algorithm with the hopes of decreasing variance in the agent learning is not necessarily true and will actually increase the variance seen by actors [11]. It seems that in MAA2C, using the joint reward to decrease the variance seen by the critic is a good way to increase performance. We do, however, note that when we increased the episode length, the individual reward mean and max returns continued to increase; however, they do not show any evidence of rapid convergence. It seems that more research is required on the effects of increasing the episode length to determine if the joint reward has a bias component.

5.3.2. MAPPO

Similarly to IPPO, MAPPO performs better when using individual rewards than when using joint rewards. MAPPO's centralized critic does not seem to be able to prevent the critic from converging prematurely. Centralized critics have been shown to increase

variance [11]; however, our results show that the increase in variance in the critic loss is not enough. Just as in IPPO, the critic converges within 100 episodes when using joint rewards. This corresponds to the majority of the gains in return, which seems to indicate that some local minima are found by the algorithm.

6. Conclusions and Future Work

In summary, our results show that different CLDE algorithms respond in different ways when the reward is changed from joint to individual in the LBF environment. MAPPO and QMIX show that they are able to leverage the additional variance present in the individual reward to find improved policies, while VDN and MAA2C suffer from the increase and perform worse. Of the centralized critic algorithms, it seems that it is crucial that the centralized algorithm critic be able to converge slowly enough to find the optimal joint policy, but not fast enough to find a local minima. In addition, if the critic is too sensitive to the increase in variance, it may diverge as in MAA2C and be unable to find the optimal policy. Value decomposition methods also seem to need additional state information to condition the coordination of agents to learn optimal policies. Since much of the emergent behaviour sought in MARL systems is a function of how agents work together, we feel that the choice of reward function may be of even more importance in MARL environments than in a single-agent environment. Our results hint that there may be some greater bias variance-type trade-off between joint and individual rewards; however, more research will need to be performed to confirm this.

As we have outlined in several sections of this work, there are still many questions that need answering before we can definitively say that the choice of using a joint reward or an individual reward when training MARL algorithms comes down to a bias variance trade-off. First, this theory of increased variance would need to be studied in simpler scenarios that can be solved analytically in order to confirm that individual rewards do increase variance. This simpler scenario would need to have the same sparse positive reward as seen in the LBF. Following the establishment of this theoretical underpinning, the next step would be to either relax the sparse constraint or the positive reward constraint and still see if the theory holds true. Once that is performed, a definitive conclusion could be presented about the effects of varying reward functions between joint and individual rewards in cooperative MARL systems.

Author Contributions: Conceptualization, P.A.; methodology, P.A. and P.M.; investigation, P.A.; software, P.A.; resources, P.M.; writing—original draft preparation, P.A.; writing—review and editing, P.M.; supervision, P.M.; funding acquisition, P.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data for the reproduction of the figures as well as validation of this research can be found at: https://github.com/at-peter/System-all-about-rewards-data.

Acknowledgments: The authors gratefully acknowledge the indirect support provided by the Mitacs Accelerate Entrepreneur program and by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Hyperparameter Optimization

CLBF Hyperparameter Optimisation

The appendix of [5] contains the hyperparameter search protocol that they used in order to perform their hyperparameter search. In order to keep the comparison to [5], we propose following the same hyperparameter search protocol, which is outlined in Table A1.

Table A1. Hyperparameter search protocol taken from [5].

Hyperparameters	Values
Hidden dimension	64/128
Learning rate	0.0001/0.0003/0.0005
Reward Standardization	True/False
Network Type	FC/GRU
Evaluation Epsilon	0.0/0.05
Epsilon Anneal	50,000/200,000
Target Update	200 (hard)/0.01 (soft)
Entropy Coefficient	0.01/0.001
n-step	5/10

The hyperparameter search was performed as follows. A search with three seeds was performed on the 10x10-3p-3f scenario to narrow down a short list of candidate hyperparameter configurations. Priority was given to hyperparameter sets that repeat.

Table A2 Shows the difference between previously tested hyperparameters and the hyperparmeters that were discovered during the hyperparameter search on the CLBF environment.

Table A2. IPPO selected hyperparameters.

Hyperparameters	Papoudakis et al. [5]	Our Hyperparameter Search
Hidden dimension	128	64
Learning rate	0.0003	0.0003
Reward Standardization	False	True
Network Type	FC	GRU
Target Update	200 (hard)	200 (hard)
Entropy Coefficient	0.001	0.01
n-step	5	10

Appendix B. Validation of Papoudakis et al. [5] Results

As part of our work on the analysis of algorithmic performance, we replicated the work that was performed as part of [5] on the LBF environment. This section includes the data that were collected from our repeated experiments. We used the hyperparameters that were reported in the appendix section of [5] and ran 10 runs for each hyperparameter configuration. The selected hyperparameters were those for parameter sharing, and parameter sharing was used for the data collection to keep in line with the results in [5].

We found discrepancies between the reported data in [5] for VDN and QMIX, and these discrepancies also seem to explain some of the results we reported in [19]. Notably, we found that the convergence of the value factorization methods was not reported properly in [5], and these convergence values are in line with the increase in convergence rates that we found in [19].

Table A3. Maximum returns and 95% confidence interval of hyperparameter configurations taken from [5]. Bolded values are those that differ significantly from [5].

Tasks/Algs	IQL	IA2C	IPPO	MAA2C	MAPPO	VDN	QMIX
8x8-2p-2f-c	1.0 ± 0.0	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.00	1.0 ± 0.0	1.0 ± 0.00	1.0 ± 0.00
8x8-2p-2f-2s-c	0.97 ± 0.01	1.0 ± 0.0	0.63 ± 0.02	1.0 ± 0.0	0.56 ± 0.02	0.98 ± 0.00	$\mathbf{0.97 \pm 0.0}$
10x10-3p-3f	0.89 ± 0.08	0.99 ± 0.01	0.89 ± 0.02	0.98 ± 0.01	0.9 ± 0.24	0.9 ± 0.03	$\mathbf{0.91 \pm 0.02}$
10x10-3p-3f-2s	0.7 ± 0.01	0.84 ± 0.04	0.56 ± 0.01	0.85 ± 0.01	0.58 ± 0.01	0.77 ± 0.01	$\mathbf{0.76 \pm 0.04}$

Table A4. Average returns and 95% confidence interval of hyperparameter configurations taken from [5]. Bolded values are those that differ significantly from [5].

Tasks/Algs	IQL	IA2C	IPPO	MAA2C	MAPPO	VDN	QMIX
8x8-2p-2f-c	0.77 ± 0.08	0.96 ± 0.01	0.96 ± 0.01	0.97 ± 0.01	0.96 ± 0.02	0.78 ± 0.04	$\mathbf{0.69 \pm 0.04}$
8x8-2p-2f-2s-c	0.82 ± 0.01	0.94 ± 0.01	0.39 ± 0.02	0.94 ± 0.0	0.45 ± 0.02	0.84 ± 0.0	$\mathbf{0.79 \pm 0.01}$
10x10-3p-3f	0.47 ± 0.07	0.88 ± 0.02	0.71 ± 0.03	0.87 ± 0.02	0.59 ± 0.21	0.56 ± 0.03	$\mathbf{0.46 \pm 0.04}$
10x10-3p-3f-2s	0.56 ± 0.01	0.67 ± 0.05	0.44 ± 0.0	0.69 ± 0.02	0.46 ± 0.0	0.6 ± 0.01	$\mathbf{0.56 \pm 0.05}$

Appendix C. Variance Analysis Data

This section of the appendix contains all the statistical data analysis that was used during the empirical variance analysis in Section 4. The statistical analysis used Bartlett's test in order to determine if the variance in two means is the same. The α value used to determine statistical significance is $\alpha = 0.05$. Bartlett's test tests the null hypothesis h_0 that the variances of each data distribution tested are identical. If the p-value is below that of the selected α, then the null hypothesis is rejected, and the variances of the data tested are not the same. In our analysis, the data collected for each run were averaged over, and then the set of 10 replicates was used in Bartlett's test. The nan value indicates that there was no variation at all because the algorithm was able to solve the scenario perfectly in the 25 timestep scenarios for both individual and joint rewards.

Appendix C.1. IQL

Below are the statistics that were gathered on the IQL algorithm. The result aspects of the algorithm that were compared include the following:loss, grad norm, mean of selected q values, means of return, max of returns, and target network mean q values for the selected action. Variances are evaluated between joint reward and independent reward. Bolded p-values reject the null hypothesis, indicating that the variances between the 25-step and 50-step runs are different.

Table A5. p-values of Bartlett's test for homogeneity of variances for gradient norm values of IQL between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.88	0.15	**0.048**	0.35
50	**0.066**	0.63	0.18	**0.044**

Table A6. p-values of Bartlett's test for homogeneity of variances for loss values of IQL between 25 timesteps and 50 timesteps.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.014**	0.095	$\mathbf{1.45 \times 10^{-3}}$	0.71
50	0.21	0.069	$\mathbf{6.42 \times 10^{-4}}$	0.67

Table A7. p-values of Bartlett's test for homogeneity of variances for the mean q value of selected actions of IQL between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.68	0.50	0.074	**0.002**
50	0.56	0.49	0.059	0.64

Table A8. p-values of Bartlett's test for homogeneity of variances for the target value of selected actions of IQL between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.70	0.47	0.080	$\mathbf{2.08 \times 10^{-3}}$
50	0.56	0.47	0.061	0.63

Table A9. *p*-values of Bartlett's test for homogeneity of variances for the mean return values of IQL between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.99	0.36	$\mathbf{2.80 \times 10^{-3}}$	$\mathbf{7.73 \times 10^{-3}}$
50	0.73	0.48	**0.023**	0.57

Table A10. *p*-values of Bartlett's test for homogeneity of variances for the max return values of IQL between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{1.84 \times 10^{-6}}$	0.77	$\mathbf{5.28 \times 10^{-5}}$	0.41
50	nan	$\mathbf{9.22 \times 10^{-3}}$	$\mathbf{1.60 \times 10^{-3}}$	0.47

Appendix C.2. IPPO

Below are the statistics that were gathered on the IPPO algorithm. The statistics that were tested include the following: mean return, max return, agent grad norms, critic grad norms, critic loss, policy gradient loss, maximum Pi values of the actor, and advantage means. Variances are evaluated between joint reward and independent reward. Bolded *p*-values reject the null hypothesis, indicating that the variances between the 25-step and 50-step runs are different.

Table A11. *p*-values of Bartlett's test for homogeneity of variance for mean returns of IPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{4.52 \times 10^{-5}}$	**0.074**	0.54	$\mathbf{1.14 \times 10^{-5}}$
50	0.16	$\mathbf{3.39 \times 10^{-6}}$	0.75	$\mathbf{1.37 \times 10^{-4}}$

Table A12. *p*-values of Bartlett's test for homogeneity of variance for max returns of IPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.0**	**0.013**	**0.049**	$\mathbf{9.04 \times 10^{-4}}$
50	**0.0**	$\mathbf{8.66 \times 10^{-7}}$	$\mathbf{8.26 \times 10^{-6}}$	0.34

Table A13. *p*-values of Bartlett's test for homogeneity of variance for agent grad norms returns of IPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{8.27 \times 10^{-17}}$	$\mathbf{9.97 \times 10^{-4}}$	$\mathbf{1.17 \times 10^{-12}}$	$\mathbf{3.23 \times 10^{-14}}$
50	$\mathbf{9.14 \times 10^{-16}}$	$\mathbf{5.11 \times 10^{-8}}$	$\mathbf{8.82 \times 10^{-13}}$	$\mathbf{5.99 \times 10^{-18}}$

Table A14. *p*-values of Bartlett's test for homogeneity of variance for critic grad norms returns of IPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{1.98 \times 10^{-21}}$	$\mathbf{1.15 \times 10^{-13}}$	$\mathbf{2.98 \times 10^{-22}}$	$\mathbf{2.27 \times 10^{-15}}$
50	$\mathbf{4.39 \times 10^{-23}}$	$\mathbf{2.60 \times 10^{-15}}$	$\mathbf{9.49 \times 10^{-20}}$	$\mathbf{5.75 \times 10^{-18}}$

Table A15. *p*-values of Bartlett's test for homogeneity of variance for critic loss of IPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{1.51 \times 10^{-25}}$	$\mathbf{2.56 \times 10^{-10}}$	$\mathbf{2.33 \times 10^{-22}}$	$\mathbf{2.43 \times 10^{-17}}$
50	$\mathbf{1.69 \times 10^{-28}}$	$\mathbf{1.79 \times 10^{-11}}$	$\mathbf{1.01 \times 10^{-18}}$	$\mathbf{2.90 \times 10^{-18}}$

Table A16. *p*-values of Bartlett's test for homogeneity of variance for policy gradient loss of IPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{1.86 \times 10^{-17}}$	$\mathbf{2.06 \times 10^{-9}}$	$\mathbf{5.25 \times 10^{-14}}$	$\mathbf{4.11 \times 10^{-12}}$
50	$\mathbf{1.22 \times 10^{-22}}$	$\mathbf{9.84 \times 10^{-17}}$	$\mathbf{7.55 \times 10^{-14}}$	$\mathbf{4.60 \times 10^{-14}}$

Table A17. *p*-values of Bartlett's test for homogeneity of variance for maximum policy values of IPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{1.39 \times 10^{-5}}$	$\mathbf{1.28 \times 10^{-3}}$	0.19	0.18
50	0.64	$\mathbf{6.63 \times 10^{-4}}$	$\mathbf{1.88 \times 10^{-4}}$	**0.011**

Table A18. *p*-values of Bartlett's test for homogeneity of variance for advantage means of IPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{2.05 \times 10^{-17}}$	$\mathbf{3.98 \times 10^{-10}}$	$\mathbf{9.86 \times 10^{-15}}$	$\mathbf{2.59 \times 10^{-12}}$
50	$\mathbf{1.12 \times 10^{-21}}$	$\mathbf{1.38 \times 10^{-16}}$	$\mathbf{5.35 \times 10^{-14}}$	$\mathbf{8.63 \times 10^{-14}}$

Appendix C.3. IA2C

Below are the statistics that were gathered on the IA2C algorithm. The statistics that were tested include the following: mean return, max return, agent grad norms, critic grad norms, critic loss, policy gradient loss, maximum Pi values of the actor, and advantage means. Variances are evaluated between joint reward and independent reward. Bolded *p*-values reject the null hypothesis, indicating that the variances between the 25-step and 50-step runs are different.

Table A19. *p*-values of Bartlett's test for homogeneity of variances for mean return values of IA2C between 25 timesteps and 50 timesteps.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.016**	0.35	0.63	$\mathbf{6.69 \times 10^{-3}}$
50	**0.003**	0.38	0.12	0.063

Table A20. *p*-values of Bartlett's test for homogeneity of variances for max return values of IA2C between 25 timesteps and 50 timesteps.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.0**	$\mathbf{4.27 \times 10^{-4}}$	**0.071**	0.29
50	**0.0**	**0.0**	$\mathbf{2.07 \times 10^{-9}}$	**0.016**

Table A21. *p*-values of Bartlett's test for homogeneity of variances for critic grad norm of IA2C between 25 timesteps and 50 timesteps.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.25	0.24	0.33	**0.005**
50	0.13	0.31	**0.019**	**0.010**

Table A22. *p*-values of Bartlett's test for homogeneity of variances for critic loss of IA2C between 25 timesteps and 50 timesteps.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.60	0.19	**0.011**	$\mathbf{4.12 \times 10^{-4}}$
50	0.81	0.33	**0.045**	0.17

Table A23. *p*-values of Bartlett's test for homogeneity of variances for PG loss of IA2C between 25 timesteps and 50 timesteps.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.25	0.24	0.33	$\mathbf{4.89 \times 10^{-3}}$
50	0.13	0.31	**0.019**	$\mathbf{9.87 \times 10^{-3}}$

Table A24. *p*-values of Bartlett's test for homogeneity of variances for advantage mean of IA2C between 25 timesteps and 50 timesteps.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.20	0.17	$\mathbf{3.13 \times 10^{-3}}$	**0.033**
50	**0.029**	0.13	0.15	$\mathbf{3.15 \times 10^{-3}}$

Appendix C.4. VDN

Below are the statistics that were gathered on the VDN algorithm. The results aspects of the algorithm that were compared include the following: loss, grad norm, mean of selected q values, means of return, max of returns, and target network mean q values for the selected action. Variances are evaluated between joint reward and independent reward. Bolded *p*-values reject the null hypothesis, indicating that the variances between the 25-step and 50-step runs are different.

Table A25. *p*-values of Bartlett's test for homogeneity of variances for gradient norm values of VDN between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.30	0.41	1.00	0.13
50	0.45	**0.011**	0.55	**0.005**

Table A26. *p*-values of Bartlett's test for homogeneity of variances for loss values of VDN between 25 timesteps and 50 timesteps.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.17	0.40	**0.016**	0.10
50	0.87	0.58	0.33	0.076

Table A27. *p*-values of Bartlett's test for homogeneity of variances for the mean q value of selected actions of VDN between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.034**	0.099	0.77	0.052
50	**0.021**	0.87	0.83	0.20

Table A28. *p*-values of Bartlett's test for homogeneity of variances for the target network mean q values of selected actions of VDN between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.034**	0.11	0.78	**0.038**
50	**0.020**	0.86	0.81	0.18

Table A29. *p*-values of Bartlett's test for homogeneity of variances for the mean return values VDN between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.038**	**0.002**	0.27	0.11
50	0.36	0.75	0.71	0.37

Table A30. *p*-values of Bartlett's test for homogeneity of variances for the max return values VDN between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.0**	0.50	0.31	0.089
50	**0.0**	0.26	0.26	**0.003**

Appendix C.5. QMIX

Below are the statistics that were gathered on the QMIX algorithm. The results aspects of the algorithm that were compared include the following: loss, grad norm, mean of selected q values, means of return, max of returns, and target network mean q values for the selected action. Variances are evaluated between joint reward and independent reward. Bolded *p*-values reject the null hypothesis, indicating that the variances between the 25-step and 50-step runs are different.

Table A31. *p*-values of Bartlett's test for homogeneity of variances for loss values of Qmix between 25 timesteps and 50 timesteps.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{8.18 \times 10^{-6}}$	0.13	$\mathbf{7.12 \times 10^{-5}}$	$\mathbf{5.66 \times 10^{-8}}$
50	$\mathbf{9.17 \times 10^{-32}}$	$\mathbf{1.84 \times 10^{-10}}$	0.25	$\mathbf{3.55 \times 10^{-4}}$

Table A32. *p*-values of Bartlett's test for homogeneity of variances for gradient norm values of Qmix between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{2.92 \times 10^{-8}}$	0.20	$\mathbf{6.36 \times 10^{-7}}$	$\mathbf{7.19 \times 10^{-10}}$
50	$\mathbf{1.09 \times 10^{-17}}$	$\mathbf{1.53 \times 10^{-11}}$	$\mathbf{1.06 \times 10^{-3}}$	$\mathbf{9.73 \times 10^{-7}}$

Table A33. *p*-values of Bartlett's test for homogeneity of variances for the mean q value of selected actions of Qmix between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.04**	0.55	**0.01**	$\mathbf{1.65 \times 10^{-8}}$
50	**0.04**	$\mathbf{2.86 \times 10^{-8}}$	**0.03**	0.08

Table A34. *p*-values of Bartlett's test for homogeneity of variances for the target network mean q values of selected actions of Qmix between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.04**	0.52	$\mathbf{6.18 \times 10^{-3}}$	$\mathbf{1.21 \times 10^{-8}}$
50	**0.03**	$\mathbf{2.66 \times 10^{-8}}$	**0.03**	0.07

Table A35. *p*-values of Bartlett's test for homogeneity of variances for the max return values Qmix between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	*Nan*	0.90	0.76	$\mathbf{2.73 \times 10^{-4}}$
50	$\mathbf{4.65 \times 10^{-6}}$	0.83	$\mathbf{3.90 \times 10^{-4}}$	0.21

Table A36. *p*-values of Bartlett's test for homogeneity of variances for the mean return values Qmix between 25 timesteps and 50 timesteps grouped by scenario.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.87	0.61	0.40	$\mathbf{1.24 \times 10^{-5}}$
50	$\mathbf{7.55 \times 10^{-3}}$	$\mathbf{2.10 \times 10^{-2}}$	0.51	0.25

Appendix C.6. MAA2C

Below are the statistics that were gathered on the MAA2C algorithm. The statistics that were tested include the following: mean return, max return, agent grad norms,critic grad norms, critic loss, policy gradient loss, maximum Pi values of the actor, and advantage means. Bolded *p*-values reject the null hypothesis, indicating that the variances between the 25-step and 50-step runs are different.

Table A37. *p*-values of Bartlett's test for homogeneity of variance for mean returns of MAA2C varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{1.46 \times 10^{-7}}$	$\mathbf{9.93 \times 10^{-5}}$	0.12	0.22
50	$\mathbf{9.03 \times 10^{-8}}$	$\mathbf{7.12 \times 10^{-12}}$	$\mathbf{1.59 \times 10^{-7}}$	0.89

Table A38. *p*-values of Bartlett's test for homogeneity of variance for Max Returns of MAA2C varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.0**	$\mathbf{9.83 \times 10^{-10}}$	$\mathbf{1.09 \times 10^{-4}}$	**0.011**
50	**0.0**	**0.0**	$\mathbf{9.83 \times 10^{-2}}$	0.28

Table A39. *p*-values of Bartlett's test for homogeneity of variance for agent grad norms of MAA2C varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.90	$\mathbf{7.12 \times 10^{-2}}$	0.43	$\mathbf{4.62 \times 10^{-3}}$
50	$\mathbf{1.76\times 10^{-2}}$	0.50	0.80	$\mathbf{4.54 \times 10^{-2}}$

Table A40. *p*-values of Bartlett's test for homogeneity of variance for critic grad norms of MAA2C varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	0.33	**0.020**	$\mathbf{2.81 \times 10^{-5}}$	0.74
50	0.13	**0.005**	0.13	**0.012**

Table A41. *p*-values of Bartlett's test for homogeneity of variance for critic loss of MAA2C varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	**0.027**	$\mathbf{2.52 \times 10^{-4}}$	**0.004**	0.69
50	**0.029**	0.11	$\mathbf{1.20 \times 10^{-5}}$	0.59

Table A42. *p*-values of Bartlett's test for homogeneity of variance for pg loss of MAA2C varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{3.20 \times 10^{-5}}$	$\mathbf{1.32 \times 10^{-5}}$	$\mathbf{2.22 \times 10^{-3}}$	**0.029**
50	**0.034**	$\mathbf{1.17 \times 10^{-6}}$	**0.014**	**0.025**

Table A43. *p*-values of Bartlett's test for homogeneity of variance for max policy values of MAA2C varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{3.28 \times 10^{-7}}$	$\mathbf{6.16 \times 10^{-4}}$	0.18	**0.019**
50	$\mathbf{3.31 \times 10^{-4}}$	$\mathbf{1.64 \times 10^{-3}}$	$\mathbf{1.94 \times 10^{-8}}$	0.061

Table A44. *p*-values of Bartlett's test for homogeneity of variance for advantage mean values of MAA2C varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{1.51 \times 10^{-3}}$	**0.025**	$\mathbf{1.11 \times 10^{-3}}$	0.071
50	0.38	$\mathbf{3.53 \times 10^{-4}}$	0.90	$\mathbf{1.07 \times 10^{-4}}$

Appendix C.7. MAPPO

Below are the statistics that were gathered on the MAPPO algorithm. The statistics that were tested include the following:mean return , max return, agent grad norms, critic grad norms, critic loss, policy gradient loss, maximum Pi values of the actor, and advantage means. Bolded *p*-values reject the null hypothesis, indicating that the variances between the 25-step and 50-step runs are different.

Table A45. *p*-values of Bartlett's test for homogeneity of variance for return means of MAPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{3.21 \times 10^{-4}}$	0.12	$\mathbf{1.08 \times 10^{-6}}$	$\mathbf{2.85 \times 10^{-6}}$
50	$\mathbf{2.76 \times 10^{-6}}$	0.84	0.92	$\mathbf{4.05 \times 10^{-6}}$

Table A46. *p*-values of Bartlett's test for homogeneity of variance for return maxes of MAPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{3.28 \times 10^{-5}}$	0.90	$\mathbf{3.71 \times 10^{-8}}$	$\mathbf{3.70 \times 10^{-4}}$
50	**0.00**	$\mathbf{1.86 \times 10^{-5}}$	$\mathbf{1.66 \times 10^{-6}}$	$\mathbf{1.74 \times 10^{-3}}$

Table A47. *p*-values of Bartlett's test for homogeneity of variance for agent grad norms of MAPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{1.73 \times 10^{-12}}$	$\mathbf{1.85 \times 10^{-5}}$	$\mathbf{2.49 \times 10^{-10}}$	$\mathbf{3.24 \times 10^{-9}}$
50	$\mathbf{6.17 \times 10^{-16}}$	$\mathbf{2.60 \times 10^{-8}}$	$\mathbf{3.29 \times 10^{-13}}$	$\mathbf{4.52 \times 10^{-18}}$

Table A48. *p*-values of Bartlett's test for homogeneity of variance for critic grad norm of MAPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{1.29 \times 10^{-17}}$	$\mathbf{2.35 \times 10^{-11}}$	$\mathbf{6.32 \times 10^{-11}}$	$\mathbf{4.66 \times 10^{-12}}$
50	$\mathbf{1.53 \times 10^{-20}}$	$\mathbf{1.51 \times 10^{-19}}$	$\mathbf{9.26 \times 10^{-22}}$	$\mathbf{8.30 \times 10^{-20}}$

Table A49. *p*-values of Bartlett's test for homogeneity of variance for policy gradient loss of MAPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{3.46 \times 10^{-18}}$	$\mathbf{6.95 \times 10^{-7}}$	$\mathbf{1.15 \times 10^{-7}}$	$\mathbf{6.13 \times 10^{-8}}$
50	$\mathbf{3.70 \times 10^{-22}}$	$\mathbf{1.11 \times 10^{-17}}$	$\mathbf{1.14 \times 10^{-14}}$	$\mathbf{3.87 \times 10^{-16}}$

Table A50. *p*-values of Bartlett's test for homogeneity of variance for max policy values of MAPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{6.89 \times 10^{-3}}$	0.83	$\mathbf{1.44 \times 10^{-6}}$	0.22
50	**0.015**	0.15	0.40	0.10

Table A51. *p*-values of Bartlett's test for homogeneity of variance for advantage mean values of MAPPO varying episode length between 25 timesteps and 50 timesteps and comparing reward functions.

	Foraging-8x8-2p-2f-coop/	Foraging-2s-8x8-2p-2f-coop/	Foraging-10x10-3p-3f/	Foraging-2s-10x10-3p-3f/
25	$\mathbf{2.81 \times 10^{-18}}$	$\mathbf{4.19 \times 10^{-7}}$	$\mathbf{3.13 \times 10^{-7}}$	$\mathbf{5.83 \times 10^{-8}}$
50	$\mathbf{6.95 \times 10^{-22}}$	$\mathbf{1.27 \times 10^{-17}}$	$\mathbf{1.62 \times 10^{-14}}$	$\mathbf{4.02 \times 10^{-16}}$

References

1. 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.; Whiteson, S. The StarCraft Multi-Agent Challenge. In Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montréal, Canada, 13–17 May 2019.
2. Bellemare, M.G.; Naddaf, Y.; Veness, J.; Bowling, M. The arcade learning environment: An evaluation platform for general agents. *J. Artif. Intell. Res.* **2013**, *47*, 253–279. [CrossRef]
3. Ellis, B.; Moalla, S.; Samvelyan, M.; Sun, M.; Mahajan, A.; Foerster, J.N.; Whiteson, S. SMACv2: An Improved Benchmark for Cooperative Multi-Agent Reinforcement Learning. *arXiv* **2022**, arXiv:2212.07489.
4. Oliehoek, F.A.; Amato, C. The Decentralized POMDP Framework. In *A Concise Introduction to Decentralized POMDPs*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 11–32.
5. Papoudakis, G.; Christianos, F.; Schäfer, L.; Albrecht, S.V. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. In Proceedings of the RLEM'20: Proceedings of the 1st International Workshop on Reinforcement Learning for Energy Management in Buildings & Cities, Virtual, 17 November 2020.
6. Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous Methods for Deep Reinforcement Learning. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 22 June 2016; Balcan, M.F., Weinberger, K.Q., Eds.; PMLR: New York, NY, USA, 2016; Volume 48, pp. 1928–1937.
7. Yu, C.; Velu, A.; Vinitsky, E.; Wang, Y.; Bayen, A.M.; Wu, Y. The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24611–24624 .
8. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments, In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
9. Rashid, T.; Samvelyan, M.; de Witt, C.S.; Farquhar, G.; Foerster, J.N.; Whiteson, S. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
10. Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W.M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J.Z.; Tuyls, K.; et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv* **2017**, arXiv:1706.05296.
11. Lyu, X.; Xiao, Y.; Daley, B.; Amato, C. Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning, In Proceedings of the AAMAS '21: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, Virtual, 3–7 May 2021.
12. Tumer, K.; Wolpert, D. A survey of collectives. *Collectives and the Design of Complex Systems*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 1–42.
13. Colby, M.; Duchow-Pressley, T.; Chung, J.J.; Tumer, K. Local approximation of difference evaluation functions. In Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, 9–13 May 2016; pp. 521–529.
14. Agogino, A.K.; Tumer, K. Analyzing and visualizing multiagent rewards in dynamic and stochastic domains. *Auton. Agents Multi-Agent Syst.* **2008**, *17*, 320–338. [CrossRef]
15. Proper, S.; Tumer, K. Modeling difference rewards for multiagent learning. In Proceedings of the AAMAS, Valencia, Spain, 4–8 June 2012; pp. 1397–1398.
16. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
17. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.
18. Christianos, F.; Schäfer, L.; Albrecht, S.V. Shared Experience Actor-Critic for Multi-Agent Reinforcement Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020.
19. Atrazhev, P.; Musilek, P. Investigating Effects of Centralized Learning Decentralized Execution on Team Coordination in the Level Based Foraging Environment as a Sequential Social Dilemma. In Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systemsm, L'Aquila, Italy, 13–15 July 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 15–23.