Communication Strategies in Multi-Objective Normal-Form Games

Willem Röpke
Artificial Intelligence Lab
Vrije Universiteit Brussel
Belgium
willem.ropke@vub.be

Roxana Rădulescu
Artificial Intelligence Lab
Vrije Universiteit Brussel
Belgium
roxana.radulescu@vub.be

Diederik M. Roijers
Vrije Universiteit Brussel (BE)
HU Univ. of Appl. Sci. Utrecht (NL)
diederik.roijers@vub.be

Ann Nowé
Artificial Intelligence Lab
Vrije Universiteit Brussel
Belgium
ann.nowe@vub.be

ABSTRACT

Multi-objective multi-agent systems are prevalent in many real-world scenarios. We are interested in such systems where multiple learning agents act in an environment and receive a vectorial payoff relating to the range of objectives, rather than a single scalar reward. In this work, we investigate the question of whether communication in multi-objective normal-form games (MONFGs) can alter possible equilibria and lead to changes in action selection probabilities or learning curves. We carry out a series of experiments on five MONFGs that immerse the agents in a range of self-interested and cooperative scenarios. We investigate the nuances of communicating in different situations and determine that communication can alter the learning process and lead to the emergence of new solution concepts that have not been previously observed in multi-objective multi-agent settings. We also find that communication is preferred in cases where no Nash equilibria exist. On the other hand, when Nash Equilibria are present, agents are indifferent to communication.

KEYWORDS

multi-agent systems, multi-objective decision making, reinforcement learning, communication, Stackelberg games, solution concepts, Nash equilibrium

1 INTRODUCTION

In the past, multi-agent decision making problems have been successfully modeled using multi-agent systems (MAS) with results in challenging domains such as traffic management [8, 19], robot teams [7] and smart electric grids [14]. One fundamental problem that still remains however, is the fact that these systems have mostly been developed to focus on optimising for a single objective, even though the environments in which they operate often present a range of conflicting objectives. In fact, many real-life scenarios have multiple objectives and as such require a multi-objective approach. For example, a firm providing logistics has the (possibly conflicting) objectives to deliver all goods as fast as possible to their destination, with the lowest associated cost as possible, while also minimising carbon emissions to minimise its strain on the environment. The problem that remains for the firm in question is what specific policy to use to get the best overall utility. In recent years, more research is being done in this area with the development of multi-objective multi-agent systems (MOMASs) [20] as a modeling tool for decision making in the presence of multiple agents and objectives. In this paper we take a utility-based approach towards this problem, and assume that there exists a utility function to scalarise the vectorial payoff and derive the final utility of each agent [16]. We focus on a subset of MOMASs, namely multi-objective normal-form games (MONFGs) [2, 4, 21, 23]. This paper proposes novel solutions towards finding Nash equilibria or otherwise stable solutions by introducing communication between the different agents. The communication strategies we develop are modeled as Stackelberg games. In each game we pick a leader who commits to either an action or mixed strategy and a follower that is able to use this information to select their response. We test these communication strategies in both self-interested and cooperative settings, by influencing how the follower can respond to the message. This is then evaluated on a set of environments, to determine the impact on the learning process and learned strategies. Concretely, we contribute the following[1]:

1. We develop and empirically evaluate a range of different communication techniques, both under cooperative and self-interested dynamics.
2. We find that communicating in settings where agents are looking to cooperate can moderately boost the learning process.
3. We show that when agents are in a game with self-interested dynamics, agents cycle through a range of stationary policies, thereby playing a cyclic equilibrium. We note that this is the first time this solution concept has emerged in a MONFG.
4. We show that when agents are left to choose whether they should communicate or not, agents that are in a game where no Nash equilibria exist will communicate approximately half the time. Agents that are in a game where there is a Nash equilibrium are indifferent to communication, since they are capable of finding that Nash equilibrium reliably without it.

[1] The implementation of all experiments performed in this work can be found at the following link: https://github.com/wilrop/communication_monfg

2 BACKGROUND

In this section, we discuss the necessary background. We start by defining MONFGs and utility functions. We then move to explain multi-objective optimisation criteria and the impact on possible solutions. After this, we briefly explain the actor-critic framework, which is the learning algorithm that we have used for our agents. Lastly, we discuss the concept of Stackelberg games, as it is the model on which we base our communication framework for this work.

2.1 Multi-Objective Normal-Form Games

A MONFG can be intuitively understood as a regular normal-form game where the payoff received by the agents is not a scalar value, but rather a vector of payoffs. Each element in this vector then corresponds to the value of a different objective. We can formally define this as follows:

Definition 2.1 (Multi-objective normal-form game). A (finite, n-person) multi-objective normal-form game is a tuple $(N, \mathcal{A}, \boldsymbol{p})$ with $n \geq 2$ and $d \geq 2$ objectives, where:

- N is a finite set of n players, indexed by i;
- $\mathcal{A} = A_1 \times \cdots \times A_n$, where A_i is a finite set of actions available to player i. Each vector $a = (a_1, \ldots, a_n) \in \mathcal{A}$ is called an action profile;
- $\boldsymbol{p} = (\boldsymbol{p_1}, \ldots, \boldsymbol{p_n})$ where $\boldsymbol{p_i} : A \to \mathbb{R}^d$ is the vectorial payoff of player i, given an action profile.

In this paper, we follow a utility-based approach [16]. In this approach, each agent i derives a utility from the payoff vector by applying their own scalarisation function, called the utility function $u_i : \mathbb{R}^d \to \mathbb{R}$, to this vector. An example of this is a linear utility-function which simply assigns a weight $w \in [0, 1]$ to each objective o. It is also possible for utility functions to be non-linear. In general, we assume that each utility function is at least monotonically increasing, which intuitively means that an agent will always prefer more of any objective over less (given equal rewards for the other objectives). Formally:

$$(\forall o, p_o^{\pi} \geq p_o^{\pi'}) \implies u(\boldsymbol{p}^{\pi}) \geq u(\boldsymbol{p}^{\pi'}) \tag{1}$$

with two policies π and π'. This definition of a utility function instantly raises the question about what the agent should optimise for. In some cases, it would be favourable to optimise for the utility of each individual policy execution, resulting in what is called the Expected Scalarised Returns (ESR) criterion [9, 15, 21]:

$$p_{u,i} = \mathbb{E}\left[u\left(\boldsymbol{p}_i^{\boldsymbol{\pi}}\right)\right] \tag{2}$$

with $p_{u,i}$ the expected utility for agent i with utility function u and $\boldsymbol{p}_i^{\boldsymbol{\pi}}$ the vectorial reward for agent i under joint policy $\boldsymbol{\pi}$.

Alternatively, it is possible that the agent cares about optimising for the utility it can derive from several executions of the same policy, in which case we first calculate the expectation over the returns before scalarising this vector. This is called the Scalarised Expected Returns (SER) criterion[16, 21]:

$$p_{u,i} = u\left(\mathbb{E}\left[\boldsymbol{p}_i^{\boldsymbol{\pi}}\right]\right) \tag{3}$$

where $p_{u,i}$ is now the utility of the expected returns.

As a concrete illustration to show the differences between the ESR and SER optimisation criterion, consider the logistics example that we previously mentioned. Choosing whether to optimise for the average utility of each individual package delivery or attempting to optimise the utility for the average delivery can result in two very different policies. The choice between these optimisation criteria is thus important to consider, as it has also been shown that ESR and SER are not equivalent under non-linear utility functions [21]. This work further showed that in stateless settings ESR can be reduced to a single-objective problem when the utility functions are known, implying that regular reinforcement learning (RL) techniques can be used to solve such problems. The SER criterion on the other hand can not easily be solved by traditional RL techniques and has been understudied thus far. For these reasons, we concern ourselves with the SER optimisation criterion in this work.

2.2 Solution Concepts

To understand the fundamental dynamics of multi-objective games we use game-theoretic equilibria as solution concepts. A well-known solution concept is the Nash equilibrium [12]. The original formulation for Nash equilibria (NE) used in single-objective games is adapted to multi-objective games as follow [20]:

Definition 2.2 (Nash equilibrium for scalarised expected returns). A joint policy $\boldsymbol{\pi}^{NE}$ leads to a Nash equilibrium under the scalarised expected returns criterion if for each agent $i \in 1, \cdots, n$ and for any alternative policy π_i :

$$u_i\left(\mathbb{E}\boldsymbol{p}_i\left(\pi_i^{NE}, \boldsymbol{\pi}_{-i}^{NE}\right)\right) \geq u_i\left(\mathbb{E}\boldsymbol{p}_i\left(\pi_i, \boldsymbol{\pi}_{-i}^{NE}\right)\right)$$

This definition is similar to the definition of a NE in a regular NFG. We must note that although every single-objective NFG has at least one NE, it has been proven that in MONFGs under SER, Nash equilibria need not exist [21]. An open question that remains is in what cases NEs do exist and to what other behaviour agents converge, if there are no NEs.

In the context of this work, it is also important to define the concept of a cyclic equilibrium in multi-objective settings. Cyclic equilibria extend the concept of Nash equilibria to cyclic policies, which are a sequence of stationary policies $\pi = \{\pi_1, \cdots, \pi_k\}$. They were first described in Markov games [32], but due to the way in which our agents communicate can also occur in MONFGs. We can formally define this solution concept as follows:

Definition 2.3 (Cyclic Nash equilibrium for scalarised expected returns). A joint cyclic policy $\boldsymbol{\pi}^{NE}$, with $\pi_i^{NE} = \{\pi_{i,1}^{NE}, \cdots, \pi_{i,k}^{NE}\}$ leads to a cyclic Nash equilibrium under the scalarised expected returns criterion if for each agent $i \in \{1, \cdots, n\}$, each policy $j \in \{1, \cdots, k\}$ and for any alternative cyclic policy π_i:

$$u_i\left(\mathbb{E}\boldsymbol{p}_i\left(\pi_{i,j}^{NE}, \boldsymbol{\pi}_{-i,j}^{NE}\right)\right) \geq u_i\left(\mathbb{E}\boldsymbol{p}_i\left(\pi_{i,j}, \boldsymbol{\pi}_{-i,j}^{NE}\right)\right)$$

2.3 Actor-Critic

To learn policies from interaction, agents in MOMASs often employ reinforcement learning (RL) algorithms. A popular algorithm that can be used for this is policy gradient [27, 30]. In policy gradient, we learn a policy π_θ parameterised by θ that outputs an action simply based on the input as follows:

$$\pi(a|s, \boldsymbol{\theta}) = \Pr\{A_t = a | S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\} \tag{4}$$

The parameters θ in this policy can be learned using the gradient of an objective function $J(\boldsymbol{\theta})$ by performing gradient ascent on this function. In practice, using regular policy gradient often results in high variance. Actor-critic methods combine both policy gradient and value-based approaches, by using learned Q-values (the critic) as a baseline for updating the policy (the actor) [27]. By doing this, we can use these Q-values as a baseline in order to reduce variance and increase stability. The objective function that is used in this work comes naturally from the SER optimisation criterion that we follow.

$$J(\boldsymbol{\theta}) = u\left(\sum_{a \in A} \pi(a|\boldsymbol{\theta})\boldsymbol{Q}(a)\right) \tag{5}$$

In this formula, $a \in A$ is an action available to the agent and $\boldsymbol{Q}(a)$ is the vectorial Q-value associated with this action. This makes it so that $\sum_{a \in A} \pi(a|\boldsymbol{\theta})\boldsymbol{Q}(a)$ represents the expected multi-objective return of the current policy. To optimise for the SER, we have to perform two consecutive steps. Firstly, we need to update the Q-table by following an adaptation of the Q-learning update rule to account for vectorial state-action values [21]:

$$\boldsymbol{Q}(a_t) \leftarrow \boldsymbol{Q}(a_t) + \alpha_Q \left[\boldsymbol{p}_t - \boldsymbol{Q}(a_t)\right] \tag{6}$$

Where α_Q is the learning rate for the Q-values. We can then update the parameters $\boldsymbol{\theta}$ by calculating the gradient of the objective function with the new Q-values as follows:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha_\theta \nabla J(\boldsymbol{\theta}_t) \tag{7}$$

2.4 Stackelberg Game

A Stackelberg game is a game-theoretic concept that attempts to model leader-follower dynamics in games. In a Stackelberg game, multiple agents are in a scenario where one or more agents are classified as leaders and other players as followers. In each round, the leaders will publicly commit to a certain action, or a mixed strategy in some cases, after which the followers are able to react to the committed action by selecting their best response [13]. Announcing ones action before actually taking it might present itself as a bad idea, since other agents could simply take advantage of it. This is certainly the case in single-objective games. As an example take the case of the rock-paper-scissors game. By announcing what you will play, the other player can always win. However, in a multi-objective setting, an agent's utility function is generally unknown to other agents. By committing to playing a specific action, this might lead agents to learn about each others utility functions faster, and in turn resulting in acceptable compromises faster. We use Stackelberg games as a basis for the communication process of agents in a MONFG.

3 COMMUNICATING IN MONFGS

To study whether communication can benefit learning in MONFGs, we propose four different communication settings, meant to evaluate different aspects and dynamics that can emerge. As a baseline, we also consider a fifth setting in which no communication takes place.

Stackelberg-like Communication. In all communication experiments, the agents play several rounds of the same game. Each round is conducted taking inspiration from Stackelberg games, in that one agent is the *leader* and communicates with the other agent about their intentions. The other agent is the *follower* who is then able to react to this communication. Agents switch after each round between being the leader and being the follower. Also note that after a round is played, agents continue with the regular approach of updating the Q-values and policies, just as it would happen when no communication was involved.

Communicating to Learn to Cooperate. The first experiment that uses communication places the agents in a *cooperative setting*. We thus assume that the agents are aiming to learn one optimal *joint policy*. Each episode, the leader commits to playing an action that was selected by their policy. This action is then communicated to the follower, who first gets a chance to update their policy, knowing what the other agent will play. In this setting, agents thus learn joint-action Q-values $Q(a, a')$ as done in joint-action learners [6]. Because the follower knows what action the leader will play, it knows which Q-values to select from the joint-action table. It then uses the selected Q-values to update its policy using the actor-critic implementation. As an example, say agent 1 commits to playing action 1. Agent 2 can now take this row from its table and update their policy using the Q-values found in this row. We show a visual representation of this algorithm in Figure 1. Contrary to the Stackelberg game literature, the communicated strategy does not arise from the inability of the leading agent to keep its strategy hidden (e.g., airport security may need to assume that their surveillance strategy is known). Rather, the communication in these MONFGs is meant to foster cooperation, by letting the other agent know what the agent likes to play. This closely resembles the iterated best response algorithm, in which each round one player is able to optimise their policy with regards to the other player's current policy [3, 5].

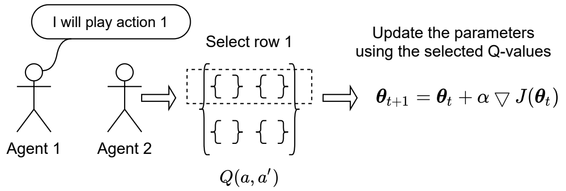

I will play action 1
Select row 1
Update the parameters using the selected Q-values
$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \bigtriangledown J(\boldsymbol{\theta}_t)$
{ } { }
{ } { }
Agent 1
Agent 2
$Q(a, a')$

Figure 1: The setting under cooperative dynamics. In this setting the leader commits to playing a certain action, after which the following agent can optimise their own policy.

Self-interested Communicative Agents. The second experiment concerns a self-interested setting, i.e., MONFGs in which the interests of the agents are not aligned. Again, each episode one agent is selected as the leader and one agent as the follower. Only now, instead of the agents wanting to find a single joint cooperative strategy, agents learn distinct policies for different situations. When leading, agents learn a communication policy that suits them best, while when following agents aim to select their best response with regards to the received message. Out all of our settings, this most closely resembles the Stackelberg game. There are however key differences between single-objective Stackelberg games and communication between self-interested agents in a multi-objective normal form game. In a single-objective setting, playing one's best response would imply simply picking the maximum Q-value for the committed action, resulting in a deterministic strategy. In this multi-objective setting however, using scalarised expected returns, playing a mixed strategy against the committed action can be the best response. Specifically, this means that each agent learns a different policy for each possible action that can be suggested by the other agent. We show this algorithm in Figure 2.

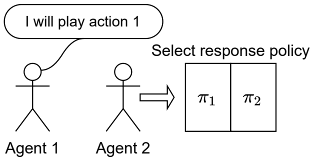


Figure 2: The setting under self-interested dynamics. In this setting the leader again commits to playing a certain action. This time however, the following agent reacts by playing their best response policy.

Policy Communication for Cooperation. We have implemented two more communication settings, this time focusing on the leader communicating their entire policy rather than their next action. We investigate a cooperative setting where the follower uses the communicated policy to update their own policy analogous to the dynamics in cooperative action communication.

We note that, in contrast to the previous experiments though, the policy communication setting cannot easily be adjusted for self-interested dynamics. In fact, this would imply that we move away from the actor-critic implementation we follow now. This is because agents have continuous policies, which makes it so we cannot learn a *separate* best response to each possible communication (i.e., the leader's communicated policy). We thus reserve investigating the self-interested setting for future work.

A Hierarchical Approach to Communication. The last setting that we propose, presents a natural conclusion to the question of how communication can influence learning agents in MONGFs. In this experiment, agents learn a total of three policies. The first policy is at the top of the hierarchy and decides whether an agent should communicate their policy to the other agent or not. The other two policies are explicit policies for when the agent is in a communication setting or when it is not. As an example, say agent 1 decides with its top level policy that it wants to communicate. It will send its current policy under communication as a message to agent 2, who then updates their policy used under communication and responds by playing an action according to this policy. On the other hand, if agent 1 decides not the communicate, it will play according to its no-communication policy, as will agent 2.

4 EXPERIMENTS

All of our experiments were evaluated on five MONFGs that were also used in previous work [21, 22, 31]. We broadly include two types of games: one where no NE under SER and with the given utility functions exist and one where at least one pure NE exists. The games with no NE under SER can be seen in Tables 1, 2 and the games with NE in Tables 3, 4, 5.

	L	M	R
L	(4, 0)	(3, 1)	(2, 2)
M	(3, 1)	(2, 2)	(1, 3)
R	(2, 2)	(1, 3)	(0, 4)

Table 1: Game 1 - The (im)balancing act game. This game has no NE under SER.

	L	R
L	(4, 0)	(2, 2)
R	(2, 2)	(0, 4)

Table 2: Game 2 - The (im)balancing act game without M. This game also has no NE under SER.

	L	M
L	(4, 0)	(3, 1)
M	(3, 1)	(2, 2)

Table 3: Game 3 - The (im)balancing act game without R. This game has one pure NE under SER, (L, M), with an expected utility of 10 for agent 1 and 3 for agent 2.

	L	M
L	(4, 1)	(1, 2)
M	(3, 1)	(3, 2)

Table 4: Game 4 - A 2-action game with pure NE under SER in (L,L) and (M,M) with expected utilities of 17 and 4 under (L, L) and 13 and 6 under (M, M).

	L	M	R
L	(4, 1)	(1, 2)	(2, 1)
M	(3, 1)	(3, 2)	(1, 2)
R	(1, 2)	(2, 1)	(1, 3)

Table 5: Game 5 - A 3-action game with three pure NE under SER. (L,L) with expected utilities of 17 and 4, (M,M) with expected utilities of 13 and 6 and (R,R) with expected utilities of 10 and 3. Note that (R, R) is Pareto-dominated by (L, L) and (M, M).

Throughout the experiments we see that games with no NE all show similar behaviour to each other. The same applies to the games with NE. For this reason, we opt to show only the results for Game 1 and Game 5, as they have the largest action spaces and as such show the most interesting results. For each setting, we show the scalarised expected returns over time for both agents in Figure 3 and the empirical state distribution over the last 10% of episodes in Figure 4 for Game 1. We show the same plots for Game 5 in Figure 5 and Figure 6.

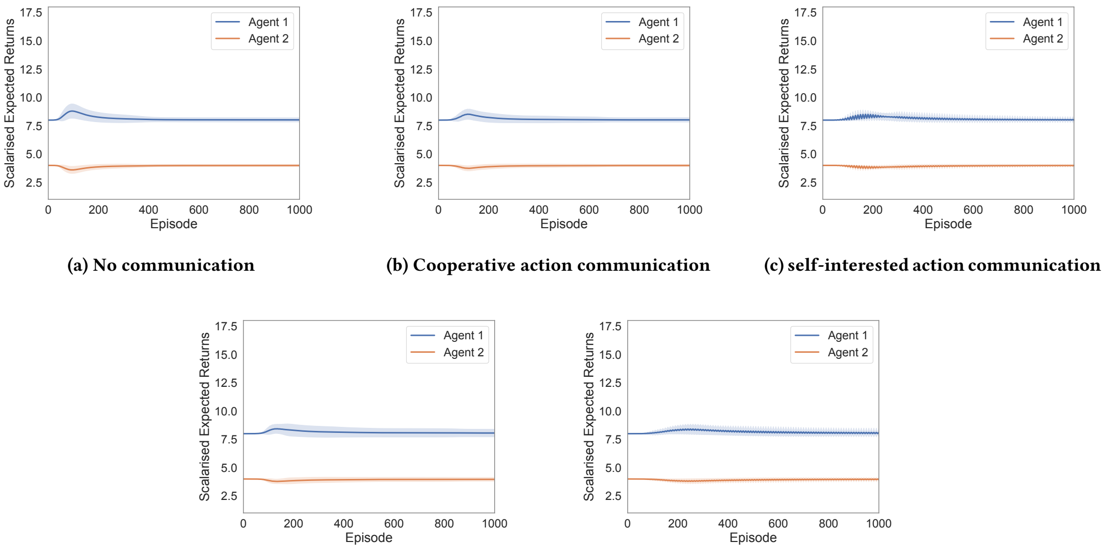
In each game, both agents get the same payoff vector $\boldsymbol{p} = [p^1, p^2]$. The utility function for agent 1 (row player) is:

$$u_1([p^1, p^2]) = p^1 \cdot p^1 + p^2 \cdot p^2 \tag{8}$$

and for agent 2 (column player) is:

$$u_2([p^1, p^2]) = p^1 \cdot p^2 \tag{9}$$

In all experiments, agents learn an explicit policy by using an actor-critic implementation. The actual policy is computed as a simple



(a) No communication

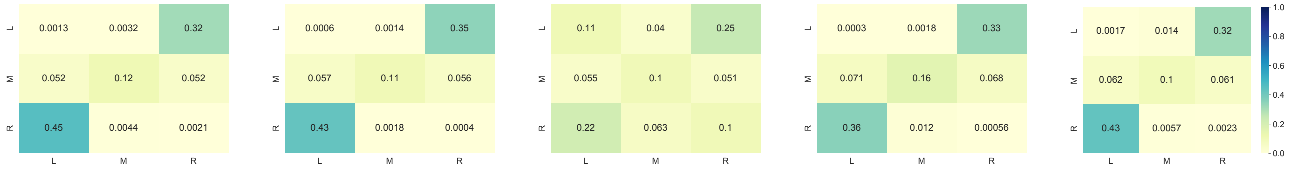
(b) Cooperative action communication

(c) self-interested action communication

(d) Policy communication

(e) Optional policy communication

Figure 3: The scalarised expected returns for both agents when operating in Game 1 under the different communication dynamics. We only show the first 1000 episodes for a clearer visualisation of the learning process. The coloured area around the scalarised expected returns represents the standard deviation of the data.



(a) No communication

(b) Cooperative action communication

(c) Self-interested communication

(d) Policy communication

(e) Optional policy communication

Figure 4: The empirical state distribution for the last 10% of episodes from each run when operating in Game 1 under the different communication dynamics.

softmax function over the parameters $\boldsymbol{\theta}$:

$$\pi(a = a_1|\boldsymbol{\theta}) = \frac{e^{\theta_i}}{\sum_{j=1}^{|A_i|} e^{\theta_j}} \tag{10}$$

Each experiment was ran for 5000 episodes and averaged over 100 trials. Lastly, we have used a learning rate for all Q-values and parameters $\boldsymbol{\theta}$ of 0.05, except when explicitly mentioned otherwise.

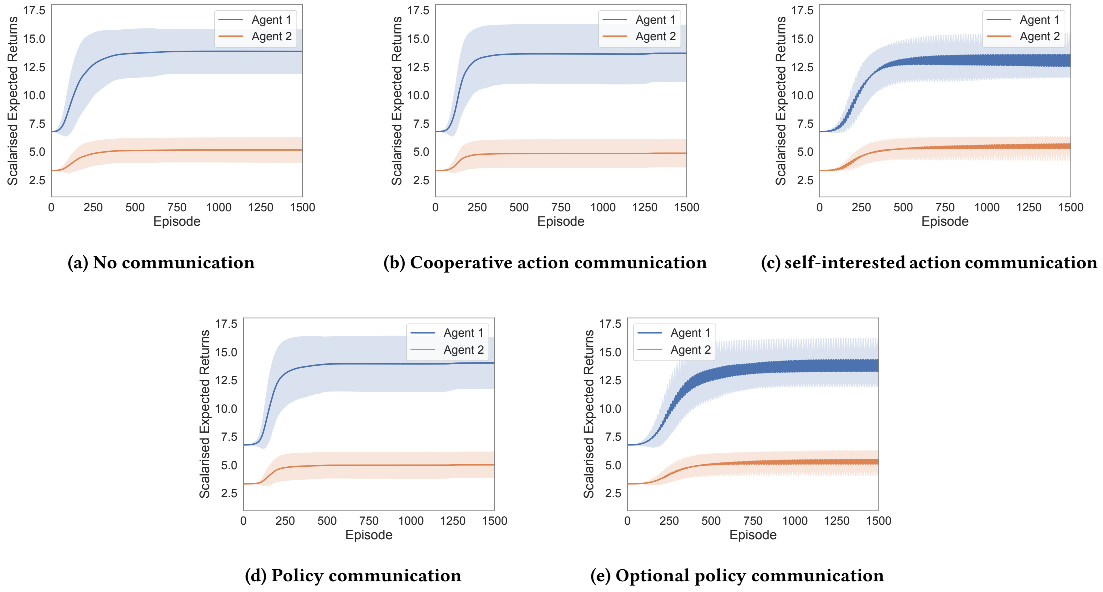
4.1 No Communication

The first setting that we evaluate has the agents playing the MONFG without any form of communication. This experiment serves as a baseline for other experiments, to see the precise impact of the communication later. In Figure 3a we show the obtained SER over time for both agents in Game 1 without communication. After a small deviation in the beginning, we see that agents converge quickly to a SER of 8 for agent 1 and 4 for agent 2. We also show the state distributions for the last 10% of every run in Figure 4a. We can clearly see that agents mostly resort to some sort of middle ground by playing (R, L), (M, M) or (L, R) each with a respective payoff vector of (2, 2).

We show the same plots for Game 5, with the scalarised expected returns in Figure 5a and the state distribution in Figure 6a. We again see that agents converge quite quickly at a utility around 13.5 for agent 1 and 5 for agent 2. This time however, agents are able to converge on a joint strategy by playing mostly (M, M) or (L, L), which are the two undominated Nash equilibria in this game.

4.2 Cooperative Action Communication

The first experiment that uses communication places the agents in a cooperative setting. As we can see in Figure 3b for Game 1 and Figure 5b for Game 5, this leads to the same SER, but with a more directed learning curve. What we mean by this is that the agents



(a) No communication

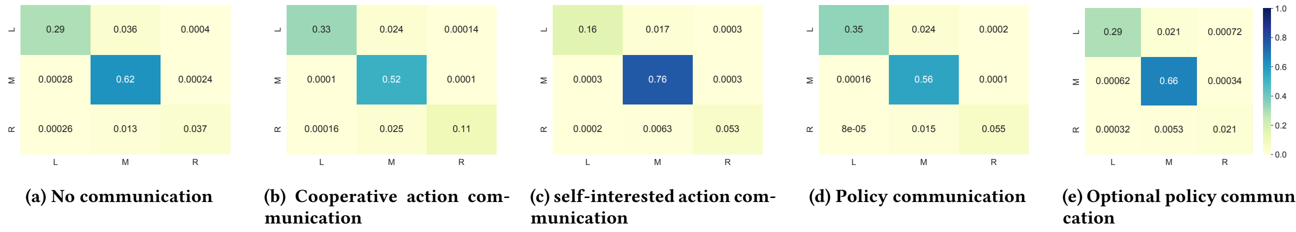
(b) Cooperative action communication

(c) self-interested action communication

(d) Policy communication

(e) Optional policy communication

Figure 5: The scalarised expected returns for both agents when operating in Game 5 under the different communication dynamics. We only show the first 1500 episodes for a clearer visualisation of the learning process. The coloured area around the scalarised expected returns represents the standard deviation of the data.



(a) No communication

(b) Cooperative action communication

(c) self-interested action communication

(d) Policy communication

(e) Optional policy communication

Figure 6: The empirical state distribution for the last 10% of episodes from each run when operating in Game 5 under the different communication dynamics.

appear to converge at around the same time as without communication, but once the agent starts learning the optimal policy, there is less time required for it to converge. In the games without NE, this translates to the agent having a smaller divergence from the final strategy. In the games with NE, we see a steeper learning curve towards the optimal strategy. From the state distribution plots we can substantiate the conclusion that the agents end up playing the same policies as without communication, as the figures are nearly identical. We however do note that the communication can lead agents to play a dominated Nash equilibrium more often than without communication. This can be attributed to the policy update after the action communication. In earlier episodes, agents are mostly communicating random actions, which might lead agents to optimise for dominated NE. Because the follower agent now has two update moments, suboptimal behaviour might be reinforced too quickly and before sufficiently exploring alternatives. This drawback appears to be intrinsic to the immediate policy updates used here, since removing this functionality would render communication useless and reduces this setting to the non-communication setting.

4.3 Self-interested Action Communication

The second experiment involving communication also lets agents commit to an action, but now allows the follower to pick their best response action for a more self-interested setting. This leads to the first surprising results in Figure 3c and Figure 5c. In this setting, it is possible for agents to end up in a cyclic equilibrium, where each agent has a different policy when committing and when following. We also observe the first big difference between games with no NE and games with pure NE. Looking at the state distributions for Game 1 with no NE in Figure 4c, we see that agents do not end up reliably playing a certain strategy. This is explainable due to the very definition of a NE. In a game with pure NE, when one agent commits to playing their action of an equilibrium, the other agent has no choice but to also play this equilibrium, otherwise it would not be a NE by definition. However, since there are no NE in Game 1, agents can take advantage of the knowledge and exploit it by playing an action that is more favourable to them, which then leads to the observed behaviour where agents do not play a certain strategy reliably. In Game 5 where there are NE, the inverse is true. Here, an agent knows a priori what the agent will play and as such is free to select their best response, which is their part of the NE. This leads to the NE being played with a very high probability and all other actions only rarely.

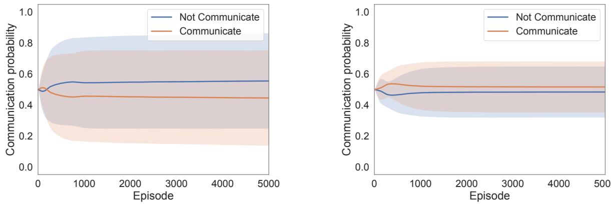
4.4 Policy Communication

The results in this case are very similar to the results obtained in the cooperative action communication setting. Again, we see in the SER plots that agents have a moderately more directed learning curve than in the no communication setting. Also similar to this setting, we see that when no NE are present in the MONFG, agents end up playing the middle ground. This results in a state plot that shows no real preference for any action combination. In the games with NE, agents do play these NE, as the policy communication simply helps to converge more directly to the equilibrium. We also note that this setting appears to avoid ending up in dominated NE, contrary to the cooperative action communication setting. This can be attributed to the fact that agents are now clear about their uncertainty over the actions in the earlier episodes. In the setting where a single action is communicated, the leader selects and communicates actions mostly at random in the beginning. To the follower however, this single action appears as a pure strategy that the leader truly prefers as it has no way of knowing the underlying probability distribution of the leader. This leads the follower to optimise for a suboptimal policy. Because the entire policy is communicated in this setting, agents are able to avoid this trap.

4.5 Hierarchical Approach to Communication

The last experiments that were performed present a natural conclusion to the question of how communication can influence learning agents in MONGFs, by making the communication optional. The top-level policy for whether an agent should communicate or not is learned through the same actor-critic implementation as for the action selection policies, but the learning rate for Q-values and parameters $\boldsymbol{\theta}$ are set to 0.01 in this case.

First, in the games where no NE exist, the results are comparable to the self-interested action communication setting. As we can see in the SER plot for Game 1 in Figure 3e agents can again end up a cyclic equilibrium. However, one interesting result that makes it different from the self-interested action communication setting is that this time, agents are able to reliably converge to the same strategies as seen in Figure 4e. This setting appears to have circumvented the problem by combining the convergence property of the no-communication setting, with the policy communication settings. It is also interesting to look at the actual communication probabilities over time for both agents. We show this for Game 1 in Figure 7. We can clearly see that here, agents will always prefer at least some level of communication, averaging ≈ 45% for agent 1 and ≈ 50% for agent 2.

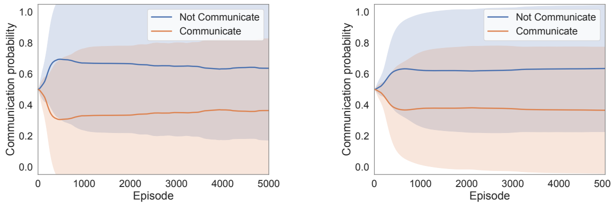


(a) Agent 1

(b) Agent 2

Figure 7: The probability of communication for both agents in Game 1. The coloured area around the communication probabilities represents the standard deviation of the data.

We also see the same pattern of cyclic equilibria in Game 5. However when looking at the communication plots in Figure 8, we can clearly see that here agents are indifferent to communication. Some runs they opt for 100% communication and other for 100% no communication. This can be attributed to the fact that they will reach a NE in either case. This same result is also supported by the state distribution in Figure 6e. We can see here that the results are almost identical to the results obtained under the no-communication setting.



(a) Agent 1

(b) Agent 2

Figure 8: The probability of communication for both agents in Game 5. The coloured area around the communication probabilities represents the standard deviation of the data.

5 RELATED WORK

In this section we will go over related work around multi-objective multi-agent systems and specifically MONGFs. For a more in depth overview, we refer to the recent survey by [20]. MONFGs were first introduced by [2]. At this time, no formal distinction was made between ESR and SER and utility functions were also mostly assumed to be linear [21]. Early work focused mostly on different solution concepts in this area, with for example calculating Pareto equilibria [4, 26] and ideal equilibria [29]. More recently, the formal distinction between between ESR and SER that we followed in this work was created [16, 17]. Later work proved that these two optimisation criteria were not equal in general and that under SER no Nash equilibria necessarily have to exist when using non-linear utility functions [21]. This work also implemented a simple Q-learning approach using ϵ-greedy as its action-selection mechanism for learning agents in these environments.

Further work developed more advanced algorithms that can handle multi-objective and multi-agent settings. Opponent modeling in specific, which has seen significant success in single-objective multi-agent systems [1] and has been identified as a direction for future research [20], is actively being explored. [31] introduced a comparative study of opponent modeling in MONFGs where they first showed an actor-critic implementation that was adapted to work under the SER criterion and further detailed the intricacies of using opponent-modeling in different settings. Following this work, [22] introduced the LOLAM algorithm that was able to perform opponent modeling with opponent learning awareness in this MONFG setting. Most work on this topic has been with regards to MONFGs, with the notable exception of work by [11] which proved that potential-based reward shaping does not alter the Pareto front in multi-objective stochastic games.

In this work we also extensively build on the existing framework of Stackelberg games. Stackelberg games were originally used to model the dynamics of a duopoly by announcing the amount of outputs for a certain firm [28]. They have seen significant adoption in other areas as well, with successful applications in for example scheduling [18], energy management [10] and notably security with the introduction of Stackelberg security games [24]. This work is not the first to realise the potential of Stackelberg games in systems with multiple objectives. An interesting example of this was the use of a multi-objective Stackelberg game between a regulator and a mining company [25]. In this work, the regulator has the conflicting objectives of maximising tax-revenues while minimising pollution and the mining company reacts to the decisions of the regulator to maximise its profit.

6 CONCLUSION AND FUTURE WORK

In this work we investigated the potential effects of different communication strategies in MONFGs. We have taken a utility-based approach and applied scalarised expected returns as the optimisation criterion. We used Stackelberg games as a model for our communication framework, which lets one player commit to an action (or mixed strategy) after which the other player can select their best response.

Five different experimental settings were developed to discover the different dynamics and see the influence of communication. In the first setting, we prohibit any agent from communicating to get a baseline for future behaviour. The next setting saw the agents in cooperative action, by every episode choosing one agent as the leader that commits to an action and letting the other agent update their policy based on this committed action. This resulted in slightly more directed learning curves across all experiments. There were however also drawbacks, as agents that were in a setting with pure NE now more often ended up playing dominated Nash equilibria. This can be attributed to the fact that the additional update after the action communication effectively reduces the time for exploration. The second setting had the same setup, but in this case, each agent was allowed to pick their best response, thereby enabling self-interested behaviour. What resulted in practice was the first occurrence of cyclic equilibria in MONFGs, where agents repeatedly play a sequence of stationary policies. The fourth setting had agents committing to their current policy. Again, the following agent could adjust their policy, based on what they now know about the other agent. This resulted in approximately the same results as the cooperative action communication. However, we note that in this setting, agents appeared less likely to end up in dominated Nash equilibria in games were there were any. The last experiment attempted to answer the overarching question whether communication is actually a benefit in these scenarios. We did this by letting agents choose for themselves whether they wanted to communicate or not. We saw that in games without any Nash equilibria, agents chose to communicate approximately 50% of the time. Lastly, we observed that in games where there were pure NE, agents were indifferent to communicating, since they were already able to find these equilibria reliably without the help of communication.

For future work, we look to three different possible paths. First of all, it would be interesting to create more intricate games, possibly by adding Gaussian noise or by simply adding more states. We could also extend the 2-player game to more general n-player games. This could show even more clearly the exact impact of communication in different settings, but would also require us to adapt the current learning algorithms. Secondly, we aim to explore more communication strategies within the Stackelberg framework. Specifically, we only present one optional communication experiment, namely with policy communication. However, it would also be interesting to see what happens in a situation with optional action communication. Thirdly, we aim to shift our focus from the relatively simplistic setting of MONFGs to more complex and realistic multi-objective stochastic games. As was noted before, research in this area is still in its infancy, but nevertheless very promising.

ACKNOWLEDGMENTS

This research was supported by funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” program.

REFERENCES

[1] Stefano V. Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. , 66–95 pages. https://doi.org/10.1016/j.artint.2018.01.002 arXiv:1709.08071

[2] David Blackwell. 1954. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.* 6, 1 (1954), 1–8. https://doi.org/10.2140/pjm.1956.6.1

[3] Shaunak D Bopardikar, Alberto Speranzon, and Cédric Langbort. 2017. Convergence analysis of Iterated Best Response for a trusted computation game. *Automatica* 78 (2017), 88–96. https://doi.org/10.1016/j.automatica.2016.11.046

[4] Peter Borm, Dries Vermeulen, and Mark Voorneveld. 2003. The structure of the set of equilibria for two person multicriteria games. *European Journal of Operational Research* 148, 3 (2003), 480–493. https://doi.org/10.1016/S0377-2217(02)00406-X

[5] C Chen, S Krishnan, M Laskey, R Fox, and K Goldberg. 2017. An algorithm and user study for teaching bilateral manipulation via iterated best response demonstrations. In *2017 13th IEEE Conference on Automation Science and Engineering (CASE)*. IEEE, Xi’an, China, 151–158. https://doi.org/10.1109/COASE.2017.8256095

[6] Caroline Claus and Craig Boutilier. 1998. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence (AAAI ’98/IAAI ’98)*. American Association for Artificial Intelligence, USA, 746–752.

[7] Rosemary Emery-Montemerlo, Geoff Gordon, Jeff Schneider, and Sebastian Thrun. 2005. Game theoretic control for robot teams. In *Proceedings - IEEE International Conference on Robotics and Automation*, Vol. 2005. IEEE, Barcelona, Spain, 1163–1169. https://doi.org/10.1109/ROBOT.2005.1570273

[8] Hodjat Hamidi and Ali Kamankesh. 2018. An Approach to Intelligent Traffic Management System Using a Multi-agent System. *International Journal of Intelligent Transportation Systems Research* 16, 2 (2018), 112–124. https://doi.org/10.1007/s13177-017-0142-6

[9] Conor F Hayes, Mathieu Reymond, Diederik M Roijers, Enda Howley, and Patrick Mannion. 2021. Distributional Monte Carlo Tree Search for Risk-Aware and Multi-Objective Reinforcement Learning. In *AAMAS 2020*.

[10] Nian Liu, Xinghuo Yu, Cheng Wang, and Jinjian Wang. 2017. Energy Sharing Management for Microgrids with PV Prosumers: A Stackelberg Game Approach. *IEEE Transactions on Industrial Informatics* 13, 3 (2017), 1088–1098. https://doi.org/10.1109/TII.2017.2654302

[11] Patrick Mannion, Sam Devlin, Karl Mason, Jim Duggan, and Enda Howley. 2017. Policy invariance under reward transformations for multi-objective reinforcement learning. *Neurocomputing* 263 (2017), 60–73. https://doi.org/10.1016/j.neucom.2017.05.090

[12] John Nash. 1951. Non-Cooperative Games. *The Annals of Mathematics* 54, 2 (1951), 286. https://doi.org/10.2307/1969529

[13] Martin J Osborne. 2004. *An introduction to game theory*. Vol. 3. New York: Oxford university Press, Oxford, United Kingdom.

[14] Markus Peters, Wolfgang Ketter, Maytal Saar-Tsechansky, and John Collins. 2013. A reinforcement learning approach to autonomous decision-making in smart electricity markets. *Machine Learning* 92, 1 (2013), 5–39. https://doi.org/10.1007/s10994-013-5340-0

[15] Diederik M Roijers, Denis Steckelmacher, and Ann Nowé. 2018. Multi-objective reinforcement learning for the expected utility of the return. In *Proceedings of the Adaptive and Learning Agents workshop at AAMAS (FAIM)*, Vol. 2018.

[16] Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* 48 (2013), 67–113.

[17] Diederik M. Roijers and Shimon Whiteson. 2017. Multi-objective decision making. In *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Vol. 34. Morgan and Claypool, 129. https://doi.org/10.2200/S00765ED1V01Y201704AIM034

[18] Tim Roughgarden. 2004. Stackelberg scheduling strategies. *SIAM J. Comput.* 33, 2 (2004), 332–350. https://doi.org/10.1137/S0097539701397059

[19] Roxana Rădulescu, Manon Legrand, Kyriakos Efthymiadis, Diederik M. Roijers, and Ann Nowé. 2018. Deep Multi-agent Reinforcement Learning in a Homogeneous Open Population. In *BNAIC 2018: Artificial Intelligence*, Martin Atzmueller and Wouter Duivesteijn (Eds.). 90–105.

[20] Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2020. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* 34, 1 (apr 2020), 10. https://doi.org/10.1007/s10458-019-09433-x

[21] Roxana Rădulescu, Patrick Mannion, Yijie Zhang, Diederik M. Roijers, and Ann Nowé. 2020. A utility-based analysis of equilibria in multi-objective normal-form games. *The Knowledge Engineering Review* 35 (2020), e32. https://doi.org/10.1017/S0269888920000351

[22] Roxana Rădulescu, Timothy Verstraeten, Yijie Zhang, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2020. Opponent Learning Awareness and Modelling in Multi-Objective Normal Form Games. arXiv:2011.07290 [cs.MA]

[23] L S Shapley and Fred D Rigby. 1959. Equilibrium points in games with vector payoffs. *Naval Research Logistics Quarterly* 6, 1 (mar 1959), 57–61. https://doi.org/10.1002/nav.3800060107

[24] Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. 2018. Stackelberg Security Games: Looking beyond a Decade of Success. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. AAAI Press, Stockholm, Sweden, 5494–5501.

[25] A Sinha, P Malo, A Frantsev, and K Deb. 2013. Multi-objective Stackelberg game between a regulating authority and a mining company: A case study in environmental economics. In *2013 IEEE Congress on Evolutionary Computation*. IEEE, Cancun, Mexico, 478–485. https://doi.org/10.1109/CEC.2013.6557607

[26] Kiran K. Somasundaram and John S. Baras. 2009. Achieving symmetric pareto nash equilibria using biased replicator dynamics. In *Proceedings of the IEEE Conference on Decision and Control*. IEEE, Shanghai, China, 7000–7005. https://doi.org/10.1109/CDC.2009.5400799

[27] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction* (second ed.). MIT press, Cambridge, MA.

[28] Heinrich Von Stackelberg. 2011. *Market structure and equilibrium*. Springer-Verlag Berlin Heidelberg, Heidelberg, Germany. 1–134 pages. https://doi.org/10.1007/978-3-642-12586-7

[29] Mark Voorneveld, Sofia Grahn, and Martin Dufwenberg. 2000. Ideal equilibria in noncooperative multicriteria games. *Mathematical Methods of Operations Research* 52, 1 (2000), 65–77. https://doi.org/10.1007/s001860000069

[30] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8, 3 (1992), 229–256. https://doi.org/10.1007/BF00992696

[31] Yijie Zhang, Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2020. Opponent Modelling for Reinforcement Learning in Multi-Objective Normal Form Games. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Auckland, New Zealand, 2080–2082.

[32] Martin Zinkevich, Amy Greenwald, and Michael L Littman. 2005. Cyclic Equilibria in Markov Games. In *Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05)*. MIT Press, Vancouver, British Columbia, Canada, 1641–1648.