Commitment and Cyclic Strategies in Multi-Objective 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

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

Diederik M. Roijers
Vrije Universiteit Brussel
HU Univ. of Appl. Sci. Utrecht
Belgium / The Netherlands
diederik.roijers@vub.be

ABSTRACT

Multi-objective games present a natural framework for studying strategic interactions between rational individuals concerned with more than one objective. We explore both the impact of commitment on the equilibria as well as the learning behaviour of agents in such games. It is well known that in single-objective normal-form games, committing to a future strategy can never be worse than the utility from a Nash equilibrium. We show that this property does not hold in multi-objective games. On the other hand, we are able to construct games in which commitment is beneficial for both players, highlighting the nuances that commitment introduces. Furthermore, we find that optimal commitment can induce the same joint-action distribution as a cyclic Nash equilibrium and show that such cyclic Nash equilibria may exist even when no Nash equilibrium exists. We evaluate these characteristics in a learning setting, to explore whether this behaviour can be expected in applications as well. We find that all proposed theorems can empirically be observed from the learning dynamics. In addition, we observe that commitment can lead to the same joint-action distribution as in a cyclic Nash equilibrium, but that it is not guaranteed when there are multiple best-responses for the follower.

KEYWORDS

Game theory, Multi-objective, Commitment, Cyclic equilibrium

1 INTRODUCTION

Leadership games, also referred to as Stackelberg games, feature one or more leaders who are obligated to commit to a strategy *a priori* and followers that react to this commitment. It is well known that optimal commitment to mixed strategies in two-player single-objective games can never hurt the leader [30] and efficient computational methods exist for the calculation of strategies to commit to [7, 11, 13]. We advance this line of research by considering the characteristics of commitment in games where players have multiple (conflicting) objectives and possibly non-linear utility functions. The relevance of such multi-objective games has been argued many times throughout the history of game theory [24, 31, 32], but they often remain under-explored [22]. As a motivating example, consider a natural extension to the well-known Bach or Stravinsky game. In this game, two friends are independently selecting whether to go to a Bach concert or a Stravinsky concert. Just as in the original game, both players favour one composer over the others, but also place a value on spending time together. In addition, they might care about the travel distance to the concert, entry price, seating arrangements, etc. Which of these *objectives* are taken into account, and how the players value the available trade-offs between these objectives, will lead to different (equilibrium) strategies.

The main focus of this paper is to provide similar guarantees as the ones available for single-objective games or show that such guarantees do not hold. To that extent, we first show that the guarantee that optimal commitment in two-player games is lower bounded by the lowest Nash equilibrium [30], is not valid in multi-objective games. On the positive side, we find that commitment may however enable players to coordinate their actions and result in a substantially higher utility for both players than when playing the game simultaneously.

We further identify an important caveat in determining which strategy to commit to in multi-objective games. When calculating or learning such strategies, it does not suffice to restrict players to stationary mixed strategies. Rather, committing to a non-stationary strategy may be optimal for the leader. We find that this non-stationary strategy can be part of a cyclic Nash equilibrium and show that cyclic Nash equilibria may exist even when no stationary Nash equilibrium exists. Complementary to our theoretical analysis of commitment and cyclic equilibria, we report several experiments to empirically support our theorems and show that learning optimal commitment strategies is feasible in multi-objective games. While we find that each property directly related to commitment is clearly visible in the experiments, commitment alone is not sufficient to ensure cyclic Nash equilibria distributions are played.

2 PRELIMINARIES

2.1 Multi-Objective Normal-Form Game

Multi-Objective Normal-Form Games (MONFGs) extend the canonical (single-objective) Normal-Form Game (NFG) with scalar payoffs to vector-valued payoffs [2]. We formally define this below:

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

Proc. of the Adaptive and Learning Agents Workshop (ALA 2022), Cruz, Hayes, da Silva, Santos (eds.), May 9-10, 2022, Online, https://ala2022.github.io/. 2022.

- 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 $a = (a_1, \dots, a_n) \in \mathcal{A}$ is called an action profile;
- $\boldsymbol{p} = (\boldsymbol{p}_1, \dots, \boldsymbol{p}_n)$ where $\boldsymbol{p}_i : \mathcal{A} \to \mathbb{R}^d$ is the vectorial payoff function for player i, given an action profile.

In some cases, we restrict a player to only play pure strategies from their set of actions $a_i \in A_i$. In general however, players are free to play any mixed strategy $s_i \in S_i$, with S_i the set of all probability distributions over A_i.

2.2 Utility-Based Approach

To deal with the vectorial nature of payoffs, we employ a utility-based approach [17, 18]. This approach assumes that each player i has a utility function $u_i : \mathbb{R}^d \to \mathbb{R}$ which maps vectors to scalar utilities.

In most current work, no restrictions are placed on the types of utility function that can be used. While imposing no restrictions on utility functions allows for broad conclusions and generalisation, it also introduces additional subtleties. Specifically, in the presence of mixed strategies, it becomes critical to decide *when* to apply the utility function. On the one hand, we might apply the utility function before calculating the expectation of such a strategy. This is known as the Expected Scalarised Returns (ESR) criterion [8, 16].

$$\mathbb{E}\left[u_i\left(\boldsymbol{p}_i(s)\right)\right] = \sum_{a \in \mathcal{A}} u_i\left(\boldsymbol{p}_i(a)\right) \prod_{j=1}^{n} s_j(a_j) \tag{1}$$

with u_i and $\boldsymbol{p}_i$ the utility function and payoff function for player i and s the joint strategy. On the other hand, a player may derive their utility from the expected vectorial returns obtained from a mixed strategy. This leads to the Scalarised Expected Returns (SER) criterion [27, 33].

$$u_i\left(\mathbb{E}\left[\boldsymbol{p}_i(s)\right]\right) = u_i\left(\sum_{a \in \mathcal{A}} \boldsymbol{p}_i(a) \prod_{j=1}^{n} s_j(a_j)\right) \tag{2}$$

It has been shown that the selection of optimisation criterion influences the (learned) strategies in single-agent [28] as well as multi-agent settings [23]. In addition, applying the ESR criterion to an MONFG effectively reduces the game to an equivalent single-objective *trade-off* game which can be solved using traditional game-theoretical techniques. Therefore, in this work we focus our attention on MONFGs under the SER criterion.

We note that the utility-based approach encompasses the older axiomatic approaches (which we discuss in Section 5). Specifically, by making assumptions about the utility functions different axioms can be derived. For example, if the utility functions are unknown but monotonically increasing, this leads to Pareto-optimality, and more specifically Pareto-Nash equilibria as the desired solution concept.

2.3 Equilibria

2.3.1 *Nash equilibrium.*

Perhaps the most known and widely studied equilibrium concept is the Nash Equilibrium (NE) [14]. A joint strategy is an NE if no agent can unilaterally deviate while still improving their utility. Nash equilibria have also been studied in the context of MONFGs with a utility-based approach. Notably, while it has been shown that every single-objective finite NFG must have a Nash equilibrium, this does not hold in MONFGs under SER [23] and existence can only be guaranteed when imposing restrictions on the utility functions used by players [21]. Below, we present a formal definition of a Nash equilibrium in an MONFG under SER:

Definition 2.2 (Nash equilibrium for scalarised expected returns). A joint strategy s^{NE} is a Nash equilibrium in an MONFG under the scalarised expected returns criterion if for all players $i \in \{1, \cdots, n\}$ and all alternative strategies $s_i \in S_i$:

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

i.e. s^{NE} is a Nash equilibrium under SER if no player can increase the *utility of its expected payoffs* by deviating unilaterally from s^{NE}.

2.3.2 *Cyclic Nash equilibrium.*

Cyclic Nash equilibria (CNE) were first introduced in the context of Markov games [34]. However, recent work has shown that cycling amongst policies may arise in MONFGs as well given a repeated setting and alternating commitment [20]. Moreover, as we demonstrate in Sections 3 and 4, cyclic equilibria might be preferred by both agents over a Nash equilibrium and can be efficiently learned in some cases. A cyclic strategy s_i is a finite sequence of stationary strategies $s_i = \{s_{i,1}, \cdots, s_{i,k}\}$, that is continuously cycled through [34]. If no agent can unilaterally deviate from their cyclic strategy and improve on its utility, the joint strategy is a CNE. The strategy is non-stationary as crucially the actual strategy that is played at each time step depends on the position in the cycle. We formally define CNE below:

Definition 2.3 (Cyclic Nash equilibrium for scalarised expected returns). A joint cyclic strategy s^{NE}, with $s_i^{NE} = \{s_{i,1}^{NE}, \cdots, s_{i,k}^{NE}\}$ is a cyclic Nash equilibrium in an MONFG under the scalarised expected returns criterion if for all players $i \in \{1, \cdots, n\}$ and all alternative cyclic strategies s_i:

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

2.4 Commitment

We study the influence of commitment in multi-objective games using the Stackelberg game model [29]. In a Stackelberg game, one or more players are designated as leaders that commit to playing a strategy. Other players are assumed to be followers and select a strategy conditioned on the commitment in response. We refer to the game that is played without commitment as the simultaneous move game.

A leadership equilibrium, also referred to as a Stackelberg equilibrium, in a multi-objective Stackelberg game is defined as follows and contains the *optimal commitment* for the leader:

Definition 2.4 (Leadership equilibrium). A joint strategy s is a leadership equilibrium in a two-player MONFG if the leader's mixed strategy $s_1 \in S_1$ maximises their utility u_1, given that for each $s_1' \in S_1$ the follower plays a strategy $s_2 \in S_2$ maximising u_2 given s_1'.

Note that while Definition 2.4 does not specify an optimality criterion as was done in Definitions 2.2 and 2.3, the interpretation of the utility depends implicitly on the optimality criterion employed by the players.

	L	R
L	1, 1	3, 0
R	0, 0	2, 1

(a) Better for the leader.

	L	R
L	1, 1	3, 0
R	0, 1	2, 2

(b) Better for both players.

Table 1: Example games for which commitment is better for only the leader (1a) or both players (1b) relative to the highest possible Nash payoff.

Extending the concept of a leadership equilibrium beyond two players is non-trivial as different assumptions have to be made regarding the leader coordination method [4], sequence of commitment [7], etc. In the case of MONFGs, this is even further complicated because of the fact that NE need not exist. As such, we only consider two-player games in this work .

Similar to the single-objective case, it is possible for the set of best responses in Definition 2.4 to contain more than one element. In that case, the definition requires an additional assumption on how the follower decides what strategy to play [4]. We sometimes consider a pessimistic view of the follower in which they select their best-response which is worst for the leader, known as a weak Stackelberg equilibrium [3]. There also exist strong Stackelberg equilibria which break ties in favour of the leader [3].

It has been shown that optimal commitment to a mixed-strategy can never decrease a players payoff relative to a NE in two-player single-objective games and can even be strictly better [30]. As an example of this, consider the NFG from Table 1a [12]. The only Nash equilibrium in this game is (L, L), with a payoff of 1 for both players. However, if the row player would be able to commit to playing R, the column player is forced to play R as well, leading to a payoff of 2 for player 1. Table 1b shows a slight variation of this game so that commitment is strictly better *for both players*. The only Nash equilibrium is again (L, L) with a utility of 1, but commitment by player 1 to R is now better for both players. In fact, any mixed-strategy commitment from player 1 where their probability of playing L is less than $\frac{1}{2}$, is met with a best response of R by player 2 which is still strictly better for both players.

Commitment has been studied as a technique for learning agents in the context of MONFGs as well [20]. It has been demonstrated that alternating commitment in two-player games, i.e. players alternate between being the leader and being the follower in repeated plays of the game, can lead to faster learning. This work focused more on the learning aspect and further showed that cyclic strategies naturally arise in this setting. We extend on this notion and provide a stronger theoretical understanding.

3 THEORETICAL CONSIDERATIONS

In this section, we consider various characteristics from commitment in single-objective games and explore whether they also apply in a multi-objective setting. For simplicity, we assume in each of the following games that the row player is the leader, while the column player is the follower.

In single-objective games it is known that the payoff from committing to an optimal strategy can never be worse for the leader relative to the lowest Nash equilibrium [30]. We show that in MONFGs this does not hold and optimal commitment may in fact be strictly worse than the utility from any Nash equilibrium. We formally state this in Theorem 3.1.

	L	R
L	$(-1,-1);(-1,-1)$	$(-1,1);(-1,1)$
R	$(1,-1);(1,-1)$	$(1,1);(1,1)$

Table 2: A two-player MONFG where committing can be worse than playing a Nash equilibrium.

THEOREM 3.1. *In a finite two-player multi-objective normal-form game where players are optimising for the scalarised expected returns criterion, the utility from optimal commitment may be worse for the leader than the utility from any Nash equilibrium.*

PROOF. Consider the game in Table 2. The row player uses the following utility function for their two objectives:

$$u(p_1, p_2) = p_1 \cdot p_2 - p_2^2$$

and assume the column player has a utility function that is a constant k for every payoff vector, i.e.:

$$u(p_1, p_2) = k$$

As the column player's utility is a constant, every best-response from the row player to a strategy of the column player is a Nash equilibrium *in the simultaneous move game*. We shall now consider the different utilities from Nash equilibria in this game.

The column player plays a strategy $\left(\frac{1}{2} - \varepsilon, \frac{1}{2} + \varepsilon\right)$ with $\varepsilon \in [-\frac{1}{2}, \frac{1}{2}]$. This leads to the following expected payoff for the row player when responding with the pure strategy L:

$$\begin{aligned}
\boldsymbol{p}(L) &= (\frac{1}{2} - \varepsilon) \cdot (-1, -1) + (\frac{1}{2} + \varepsilon) \cdot (-1, 1) \\
&= \left(\left(-(\frac{1}{2} - \varepsilon) - (\frac{1}{2} + \varepsilon)\right), \left(-(\frac{1}{2} - \varepsilon) + (\frac{1}{2} + \varepsilon)\right)\right) \\
&= (-1, 2\varepsilon)
\end{aligned}$$

Similarly for the pure strategy R:

$$\begin{aligned}
\boldsymbol{p}(R) &= (\frac{1}{2} - \varepsilon) \cdot (1, -1) + (\frac{1}{2} + \varepsilon) \cdot (1, 1) \\
&= \left(\left((\frac{1}{2} - \varepsilon) + (\frac{1}{2} + \varepsilon)\right), \left(-(\frac{1}{2} - \varepsilon) + (\frac{1}{2} + \varepsilon)\right)\right) \\
&= (1, 2\varepsilon)
\end{aligned}$$

It is clear that for any ε there is a strategy for the row player that leads to a payoff vector with the same signs, thus leading to a utility greater than or equal to zero. Because the column player never has an incentive to deviate, this joint strategy is a NE. As such, each Nash equilibrium in the simultaneous move game has a utility greater than or equal to zero.

When considering this as a *Stackelberg game*, we show that any strategy commitment made by the row player can result in a utility strictly less than zero.

The leader, i.e. the row player, commits to a strategy $\left(\frac{1}{2} - \delta, \frac{1}{2} + \delta\right)$ with $\delta \in [-\frac{1}{2}, \frac{1}{2}]$. We get the following payoff vector for the leader when the follower plays the pure strategy L:

$$\begin{aligned} \boldsymbol{p}(L) &= (\frac{1}{2} - \delta) \cdot (-1, -1) + (\frac{1}{2} + \delta) \cdot (1, -1) \\ &= \left(\left(-(\frac{1}{2} - \delta) + (\frac{1}{2} + \delta)\right), \left(-(\frac{1}{2} - \delta) - (\frac{1}{2} + \delta)\right)\right) \\ &= (2\delta, -1) \end{aligned}$$

and when the follower players the pure strategy R:

$$\begin{aligned} \boldsymbol{p}(R) &= (\frac{1}{2} - \delta) \cdot (-1, 1) + (\frac{1}{2} + \delta) \cdot (1, 1) \\ &= \left(\left(-(\frac{1}{2} - \delta) + (\frac{1}{2} + \delta)\right), \left((\frac{1}{2} - \delta) + (\frac{1}{2} + \delta)\right)\right) \\ &= (2\delta, 1) \end{aligned}$$

It is clear that whenever the leader commits to a strategy with $\delta \neq 0$, there is an action from the follower which leads to a payoff vector where the payoff for one objective is positive and negative for the other objective. If the follower selects this action, the utility for the leader is strictly less than -1. In addition, if the leader commits to a strategy with $\delta = 0$, i.e., a uniform distribution over their actions, both actions by the follower result in a utility of -1. When considering weak Stackelberg equilibria [3], and thus taking a pessimistic view of the follower such that they select their best-response which minimises the leader's utility, these strategies will always be played. As such, a uniformly mixed strategy is the optimal commitment for the leader. This results in a utility of -1, which is less than that of any Nash equilibrium. □

We note that the utility functions used in this proof are not monotonically increasing, thus violating an assumption that is often made in multi-objective decision making [17, 18]. It is unclear whether introducing the monotonically increasing assumption would shift the result, making this an open question. The construction shown here relies on a similar construction to the one used in [30] to show that commitment can be worse in infinite (single-objective) NFGs. This allows us to provide an alternative intuition for Theorem 3.1. Namely, because we can reduce each finite MONFG to an infinite pure strategy NFG [21], it is not surprising that two-player MONFGs exist showing the adverse results of commitment.

We have shown, by means of a counterexample, that optimal commitment need not be as good for the leader as playing the simultaneous move game *in general*. Of course, Theorem 3.1 does not exclude the possibility of games where commitment is still preferred. Concretely, we demonstrate that even when employing non-linear utility functions, commitment may ensure a higher utility for both players than any NE in the simultaneous move game.

Theorem 3.2. *In a finite two-player multi-objective normal-form game where players are optimising for the scalarised expected returns criterion, with possibly non-linear utility functions, the utility from commitment may be better for both players than the utility from any Nash equilibrium.*

Proof. Consider the game in Table 3. Both players employ the same utility function shown below.

$$u(p_1, p_2) = p_1^2 + p_2^2$$

	L	R
L	$(1, 0); (1, 0)$	$(2, 1); (0, 0)$
R	$(0, 0); (0, 1)$	$(1, 1); (1, 1)$

Table 3: A two-player MONFG where committing can be better for both players than playing a Nash equilibrium.

	L	R
L	$(10, 2); (10, 2)$	$(0, 0); (0, 0)$
R	$(0, 0); (0, 0)$	$(2, 10); (2, 10)$

Table 4: A two-player MONFG where committing to a cyclic strategy is optimal.

Following the algorithm from [21], we find the pure strategy NE (L, L) with a utility of one for both players. This algorithm uses the fact that when employing quasiconvex utility functions, pure strategy NE can be calculated from a scalarisation of the game. We now show that this is the only NE. Observe that the utility function is a strictly convex function. This implies that mixed strategies will never be a best-response to a fixed (possibly mixed) strategy of the opponent, except when both actions return the same expected payoff vector [21, Lemma 4]. Intuitively, this is because for strictly convex functions, a mixture of *different* points is guaranteed to be worse than any single point. As such, mixing over points can only be optimal when the points themselves are equal. There is no strategy from the column player that returns the same expected payoff vector for both actions of the row player. Given that this is also the case for the column player, no mixed-strategy NE exist.

Consider a commitment from the row player to action R. The best-response from the column player would be to play R as well. This leads to a utility of 2 for both players, which is strictly greater than that of any Nash equilibrium. □

We now introduce an additional caveat in determining optimal commitment for two-player MONFGs. Specifically, it does not suffice to restrict attention to stationary strategies and equilibria. Rather, it may be optimal for the leader to commit to a *cyclic strategy*. Moreover, this may be preferred by both players.

Theorem 3.3. *In a finite two-player multi-objective normal-form game where players are optimising for the scalarised expected returns criterion, commitment to a cyclic strategy may be optimal.*

Proof. Consider the game in Table 4. Both players employ the same utility function as shown below:

$$u(p_1, p_2) = p_1 \cdot p_2$$

There are two pure Nash equilibria, namely (L, L) and (R, R) both amounting to a utility of 20. There are also mixed Nash equilibria, however the drawback of those is that due to the independence of strategy selection, players expect to select a joint-action that results in a payoff vector of $(0, 0)$ a certain amount of the time. Given the utility functions, the optimal combination of payoffs would be $(6, 6)$ with a utility of 36. This expected payoff can only be attained by playing both (L, L) and (R, R) with probability $\frac{1}{2}$. It is clear that this joint-action distribution cannot be reached with stationary strategies. However, committing to the cyclic strategy $\{L, R\}$, induces the

	L	R
L	(2, 0); (2, 0)	(0, 1); (1, 1)
R	(1, 0); (1, 1)	(0, 2); (0, 2)

Table 5: An example game where no Nash equilibrium exists, but a cyclic Nash equilibrium does exist.

opponent to respond with the same cyclic strategy. This leads to the optimal expected return of $(6, 6)$ and a utility of 36. □

Note that the above commitment is also a CNE of the game. As a final theoretical contribution, Theorem 3.4 extends this further by formalising that such CNE can exist even when no NE exists.

Theorem 3.4. *In a finite n-player multi-objective normal-form game where players are optimising for the scalarised expected returns criterion, a cyclic Nash equilibrium may exist, even when no stationary Nash equilibrium exists.*

Proof. We show a construction for Theorem 3.4 in Table 5. The row player has the following utility function:

$$u(p_1, p_2) = p_1^2 + p_2^2$$

and the column player's utility function is shown below:

$$u(p_1, p_2) = p_1 \cdot p_2$$

First observe that there is no Nash equilibrium. The row player has a *strictly convex* utility function, implying that a mixed strategy can never be a best-response, unless both actions return the same payoff vector [21, Lemma 4]. There is no strategy from the column player which induces the same payoff for both actions. As such, the row player will always play a pure strategy. Observe next that the column player has a unique best-response to a pure strategy from the row player by playing the opposite pure strategy. Therefore, we can restrict the set of possible NE to the pure strategies. Lastly, any pure strategy by one player is best countered by the opposite pure strategy from the other player. As such, no NE exists.

We now show that $s_1 = \{L, R\}$ and $s_2 = \{L, R\}$ is a cyclic Nash equilibrium. First of all, it is clear that an *expected* payoff of $(1, 1)$ is a global maximum for the column player, given the payoffs of the game. Moreover, the row player is incentivised to play L when their opponent plays L as it dominates R. When their opponent plays R, the row player wants to play R, as it now dominates L. As such, no player can deviate from the joint cyclic strategy while improving their utility and a cyclic Nash equilibrium is reached. □

We highlight that the above cyclic equilibrium results in a joint state distribution where both (L, L) and (R, R) are played 50% of the time. This distribution is equivalent to that of the (multi-signal) correlated equilibrium as defined in [23]. Moreover, in Table 1b the joint-state distribution from commitment is also equal to that of a correlated equilibrium in the game. In that case, it appears as though the commitment by the leader can serve as a correlation signal, enabling agents to coordinate their strategies. As demonstrated by these different games, there are a number of relations and nuances between the different equilibrium concepts. We discuss studying these aspects further as a direction for future work in Section 6.

4 LEARNING LEADERSHIP AND CYCLIC EQUILIBRIA

We empirically investigate whether the theoretical contributions are observable in learning settings. The motivation for this is twofold: firstly, we now know theoretically that commitment and cyclic policies may be preferred over simple stationary independent action selection, but still lack empirical evidence. Secondly, merely knowing that such characteristics exist, does not give an immediate indication of how to actually learn or calculate the optimal strategies. As such, it becomes pertinent to study under what circumstances and learning algorithms we might arrive at optimal policies.

4.1 Learning Setup

We briefly present the learning setup used for our empirical evaluation. We design two extensions on previous work [19, 20] to incorporate the novel theoretical insights directly in the reinforcement learning algorithm. Both algorithms build on the multi-objective actor-critic approach which has been succesfully applied before because of its natural extension to multi-objective games [15, 33].

4.1.1 *Pessimistic Follower.*

The first algorithm intends to simulate a malicious follower aiming to minimise the leaders utility. Concretely, the leader commits to a stationary mixed strategy. The follower learns joint-action Q-values and calculates a strategy that minimises the utility function of the leader. Note that this is only a best-response for the follower in the case of a constant utility function, as is the case for the game in Table 2. To avoid getting stuck in local optima due to insufficient exploration of the joint-action space, we equip the follower with an exploration parameter that forces them to make a random move with probability ε. The initial value for ε is 1, i.e. always explore, but is decayed over time with factor 0.95 and clipped to a minimum of 0.01.

4.1.2 *Non-stationary.*

The second algorithm is an adaptation of the self-interested communication protocol from [20]. In short, this algorithm enables the leader to learn a mixed strategy, but forces them to commit to a pure strategy in each round by sampling from this strategy. The follower in turn learns a best-response to each individual pure strategy commitment using actor-critic as well. Our extension ensures that this ensemble of policies is itself a best-response *as a whole* against the entire commitment strategy from the leader. The critical difference is that this enables players to learn optimal non-stationary strategies, rather than strategies that are only optimal for each specific commitment.

We stress that the learning algorithm does not directly learn commitment to a cyclic strategy. Instead, the leader learns a mixed strategy from which they sample a pure strategy to commit to. This allows the players to learn the same joint-strategy distribution as for a cyclic Nash equilibrium consisting of only pure strategies.

4.1.3 *Code.*

The code is made publicly available at https://github.com/wilrop/Cyclic-Equilibria-MONFG.

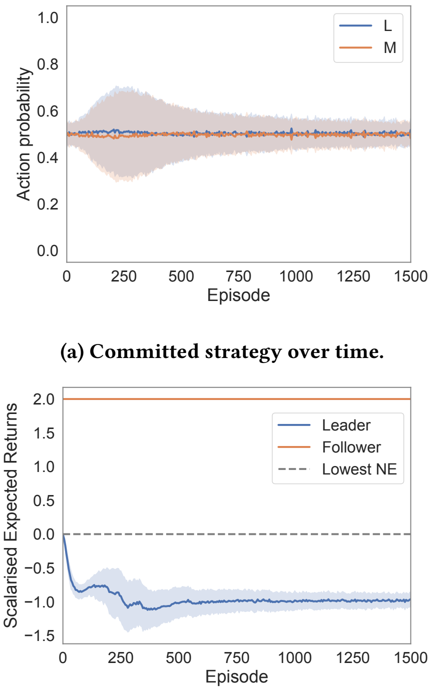
4.1.4 *Parameters.*

We perform 100 runs of the same experiment to report average behaviour. To measure the action probabilities and utility in each episode, we run a Monte-Carlo simulation of 100 rollouts. This is necessary as we study the utility agents obtain from the *expected* payoffs of their strategies. We use a learning rate of 0.2 for the Q-values and 0.005 for the policy gradient parameters.

4.2 Commitment May Be Exploited

In the first experiment, we attempt to empirically validate the theoretical possibility of a malicious follower that selects their strategy as a worst-response to the commitment. To that extent, we use the pessimistic follower algorithm described above on the game given in Table 2 with the same utility functions as used in the proof for Theorem 3.1. The constant utility k for the follower is set to 2 and each experiment is executed for 1500 episodes.

In Figure 1, we show the commitment strategy learned by the leader as well as the utility for both players. Note that we also add the lowest NE to Figure 1b to visualise the difference between the Stackelberg and simultaneous move game.



(a) Committed strategy over time.

(b) Utility over time.

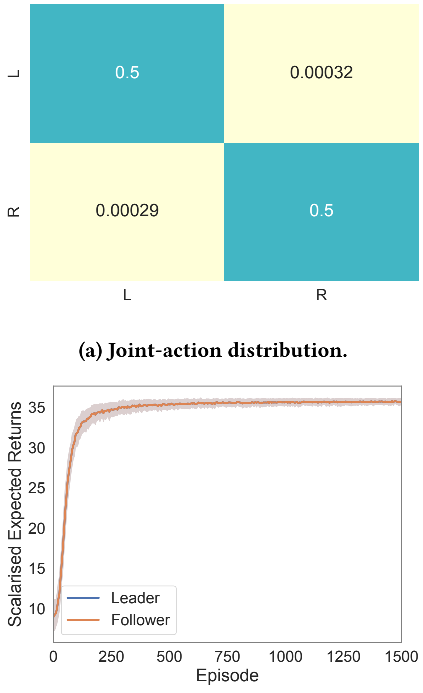
Figure 1: Results for learning commitment in Table 2.

It is clear that the leader explores different commitment strategies in the earlier episodes. However, as shown in their utility, this can always be exploited by the follower. After exploring in the earlier episodes, the leader converges on committing to a mixed-strategy of $(\frac{1}{2}, \frac{1}{2})$, which minimises the level of exploitation that is possible by the follower. We also highlight that the final utility from this leadership equilibrium is considerably less than the lower bound from Nash equilibria. As such, when applying these techniques in practical applications, it becomes extremely important to take potentially malicious behaviour from followers into account.

4.3 Commitment May Be Better

In our second experiment, we verify whether commitment may indeed be better for both players relative to a NE and if such strategies can be learned. For this experiment we use the game from Table 4 and the utility function $u(p_1, p_2) = p_1 \cdot p_2$ for both players. Each run is executed for 1500 episodes. Note that we omit experiments for the game in Table 3 as the optimal pure strategy of (R, R) is relatively simple to learn. Rather, we immediately consider the case where non-stationary behaviour is optimal.

We show the joint-action distribution of the final 10% of episodes, i.e. when players have converged on their strategy, and the utilities for both players in Figure 2.



(a) Joint-action distribution.

(b) Utility over time.

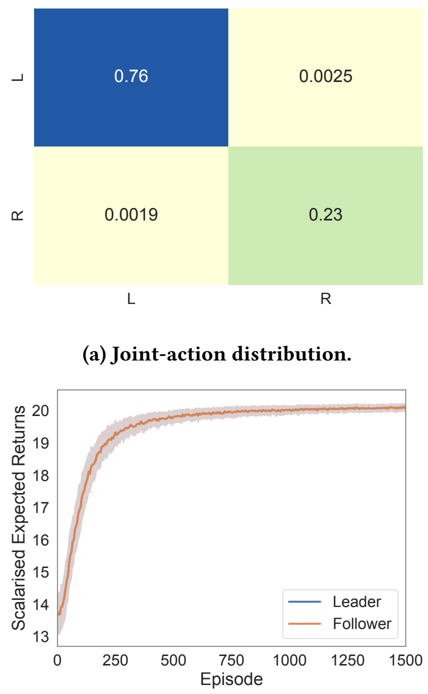
Figure 2: Results for learning commitment in Table 4.

Observe that players have converged to the optimal joint-action distribution through commitment which results in the maximum utility of 36. One critique of this game is that its payoffs are suggestive of an optimal strategy. Indeed, only the joint-actions (L, L) and (R, R) have a utility greater than zero. In addition, because agents start with an initial probability distribution of 50% for both actions, learning might be relatively easy. For this reason, we repeat the same experiment but introduce more complicated payoffs in Table 6.

Note that mixing (L, L) and (R, R) still leads to the optimal utility, this time of $\approx$ 20. This entails playing (L, L) $\frac{3}{4}$th of the time and (R, R) for the remainder. Considering cyclic equilibria directly, this could for example be reached by committing to the strategy $\{L, L, L, R\}$. We show the joint-action distribution and the utility over time in Figure 3.

	L	R
L	(10, 2); (10, 2)	(2, 3); (2, 3)
R	(4, 2); (4, 2)	(6, 3); (6, 3)

Table 6: A similar payoff table to Table 3 which induces more complicated learning dynamics.



(a) Joint-action distribution.

(b) Utility over time.

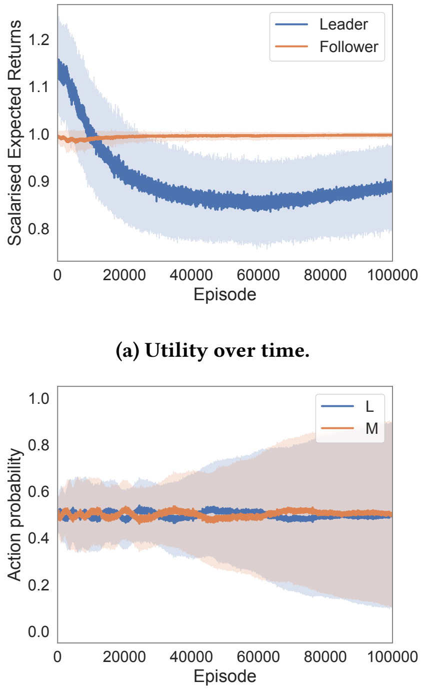
Figure 3: Results for learning commitment in Table 6.

As expected, the leader is again able to learn an optimal strategy to commit to and coordinate an optimal joint-action distribution with the follower. We note however that learning appears slightly slower, which is to be expected given the noisier payoff structure. This shows that learning optimal commitment strategies is possible and can lead to substantially higher utilities for both players.

4.4 Cyclic Nash Equilibria

Recall from Theorem 3.4 that cyclic Nash equilibria may exist when no Nash equilibrium exists. In addition, as demonstrated in Table 3, it is possible that commitment results in the same joint-action distribution as a cyclic Nash equilibrium. For this reason it begs the question: is commitment *enough* to learn CNE distributions? Note that there is already one caveat. The algorithm learns a mixed-strategy, but only enables the leader to commit to a pure strategy. As such, this restricts the number of CNE action distributions that can be represented through commitment. However, as we will show next, even when the cyclic Nash equilibrium comprises solely out of pure strategies, commitment alone is not enough to guarantee the same action distribution as the CNE.

For these experiments, we assume the game in Table 5 and the same utility functions as used in the proof of Theorem 3.4. We first execute the experiment with the row player as the leader, this time for 100.000 episodes.



(a) Utility over time.

(b) Commitment strategy.

Figure 4: Results for when the row player is the leader.

From Figure 4a, the leader's utility varies substantially over time and throughout different runs of the experiment. The leader does not reach a utility of 2, corresponding to playing the CNE distribution. Moreover, the leader's utility drops below a utility of 1, which is possible when exclusively committing to L or R. We hypothesise that this is due to the learning algorithm itself not allowing the leader to escape local maxima. Because both players need to *learn* policies, any change in commitment from the leader is countered by the follower. This then leads to oscillating behaviour where the leader increases their commitment to one action, only to be punished by the follower and update their policy towards the other action. This hypothesis is also supported by the (non-stationary) commitment strategy of the leader shown in Figure 4b. We note that it appears that the leader's utility has a slight upwards trend, implying that in the limit they can escape such local maxima. However, given the relative simplicity of this game compared to real-world situations, this approach does not appear to scale and highlights a limitation of independent learners.

We repeat the same experiment, but this time with the column player as the leader. Each run is executed for 10.000 episodes. We show the results in Figure 5. This time players are able to reliably converge on the action distribution of the CNE, guaranteeing them their optimal utility. The difference between both experiments may

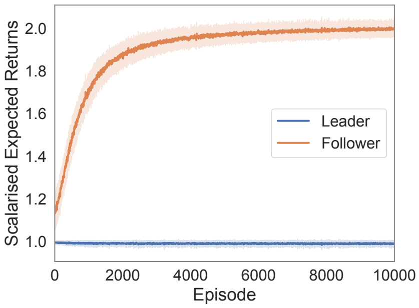


Figure 5: The utility with the column player as leader.

be explained by the best-responses available to the follower. In Figure 4, the follower has multiple strategies to reach their optimal outcome. As such, there is no direct incentive to collaborate with the leader on a strategy which also benefits the leader. Moreover, when considering weak Stackelberg equilibria, exploitative behaviour is to be expected. In Figure 5 however, the only commitment strategy guaranteeing the leader their optimal utility of 1 is to play the cyclic Nash equilibrium distribution, thereby inducing the follower to coordinate on this strategy. As such, these experiments show that while commitment may serve as a tool to coordinate players and result in CNE, careful considerations should be taken due to the many nuances in theoretical and empirical results.

5 RELATED WORK

There are two main research areas that comprise the related work of this paper. First, there are a number of studies discussing commitment in single-objective games. The seminal paper by von Stengel et al. [30] contributes several theoretical results on leadership equilibria and guarantees on their payoffs respective to Nash equilibria. Important to note is that these contributions are not limited to finite games, but extend, in some-cases, to continuous games. The work by Letchford et al. [12] builds on previous work to study exactly how much value there is in commitment. There have been a number of works exploring algorithms for calculating optimal commitment strategies. We highlight the works by Conitzer et al. [6, 7] introducing linear programming methods for this purpose. Computational methods for commitment in more complicated games have also been studied [11, 13]. Recent work also considers settings beyond two-player games. Castiglioni et al. [4] consider Stackelberg games with multiple leaders, while Coniglio et al. [5] consider games with multiple followers. We note that Stackelberg games are a successful contribution out of the game-theoretic community into practical applications, being widely applied in security settings under the name of Stackelberg security games [25].

The second line of research that is directly related is multi-objective decision making [9, 18]. We employ a utility-based approach which makes the utility functions of the players explicit. An older but related approach, sometimes called the axiomatic approach [17], leaves the existence of the utility of the players implicit by assuming the Pareto-front as its solution concept. This corresponds to unknown utility functions that are nonetheless monotonically increasing in all objectives. In the context of games, taking such an approach often implies characterizing or calculating the Pareto-Nash equilibria of a multi-objective game [10, 26]. However, these approach can always be framed under the utility-based framework. In the case of Pareto-Nash equilibria for example, mixed strategies lead to expected payoff vectors. This is equivalent to the SER criterion and a joint strategy is only a Pareto-Nash equilibrium if it is a NE for all possible monotonically increasing utility functions the agents might have. We refer to the survey of Rădulescu et al. [22] for an in-depth overview of multi-objective decision making in multi-agent settings.

6 CONCLUSION AND FUTURE WORK

We explored the value of commitment in multi-objective games and the range of behaviours it might ensure. To that extent, we first showed that while commitment in single-objective games is guaranteed to be at least as good as the worst Nash equilibrium, this guarantee does not hold for multi-objective games. Next, we demonstrated that commitment may also have a positive effect and result in a higher utility for both players relative to any Nash equilibrium. We also noted that commitment to non-stationary strategies may be preferred over stationary strategies and subsequently showed that even when no Nash equilibrium exists, such cyclic Nash equilibria may exist.

Additionally, we empirically evaluated our example games to study whether the theoretical contributions are also present in a learning setting. We first demonstrated that a malicious follower is able to accurately exploit commitment from a leader, thus motivating caution in any practical setting. Next, we also demonstrated that the positive side of commitment, increased utility for both players, is achievable even among more complicated payoff structures. Lastly, we demonstrated that while cyclic Nash equilibria distributions can sometimes be reached through commitment, commitment alone is not enough to guarantee this.

We have already mentioned several interesting directions for future work. First, we aim to further study the relations between different equilibrium concepts. Specifically, we have seen that there is an intimate relation between cyclic Nash equilibria, leadership equilibria and Nash equilibria. Additionally, we observe that in several cases the final joint-action distribution resembles that of a correlated equilibrium [1]. While correlated equilibria have been the subject of extensive work in single-objective games, much less is known about them in the context of multi-objective games [23].

Second, it has been established that there is a connection between finite multi-objective games and single-objective games with infinite pure strategy sets [21]. We aim to explore whether the contributions from multi-objective games could be extrapolated to this setting. As an example, we previously stated that whether commitment may be exploited when employing only monotonically increasing utility functions is an open question. If this question could be resolved positively, it would prove interesting to provide such a guarantee in the context of single-objective games.

ACKNOWLEDGMENTS

The first author is supported by the Research Foundation – Flanders (FWO), grant number 1197622N. This research was supported by funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” program.

REFERENCES

[1] Robert J. Aumann. 1974. Subjectivity and Correlation in Randomized Strategies. *Journal of Mathematical Economics* 1, 1 (1974), 67–96. https://doi.org/10.1016/0304-4068(74)90037-8

[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] M Breton, A Alj, and A Haurie. 1988. Sequential Stackelberg equilibria in two-person games. *Journal of Optimization Theory and Applications* 59, 1 (1988), 71–97. https://doi.org/10.1007/BF00939867

[4] Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. 2021. Committing to correlated strategies with multiple leaders. *Artificial Intelligence* 300 (2021), 103549. https://doi.org/10.1016/j.artint.2021.103549

[5] Stefano Coniglio, Nicola Gatti, and Alberto Marchesi. 2017. Pessimistic Leader-Follower Equilibria with Multiple Followers. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 171–177. https://doi.org/10.24963/ijcai.2017/25

[6] Vincent Conitzer and Dmytro Korzhyk. 2011. Commitment to Correlated Strategies. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence (AAAI'11)*. AAAI Press, San Francisco, California, 632–637.

[7] Vincent Conitzer and Tuomas Sandholm. 2006. Computing the Optimal Strategy to Commit To. In *Proceedings of the 7th ACM Conference on Electronic Commerce (EC '06)*. Association for Computing Machinery, Ann Arbor, Michigan, USA, 82–90. https://doi.org/10.1145/1134707.1134717

[8] 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 *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*. IFAAMAS, Online, 3.

[9] Conor F Hayes, Roxana Rădulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A Irissappane, Patrick Mannion, Ann Nowé, Gabriel Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. 2021. A Practical Guide to Multi-Objective Reinforcement Learning and Planning. arXiv:2103.09568 [cs.AI]

[10] Anisse Ismaili. 2018. On Existence, Mixtures, Computation and Efficiency in Multi-objective Games. In *PRIMA 2018: Principles and Practice of Multi-Agent Systems*, Tim Miller, Nir Oren, Yuko Sakurai, Itsuki Noda, Bastin Tony Roy Savarimuthu, and Tran Cao Son (Eds.). Springer International Publishing, Cham, 210–225.

[11] Joshua Letchford and Vincent Conitzer. 2010. Computing Optimal Strategies to Commit to in Extensive-Form Games. In *Proceedings of the 11th ACM Conference on Electronic Commerce (EC '10)*. Association for Computing Machinery, Cambridge, Massachusetts, USA, 83–92. https://doi.org/10.1145/1807342.1807354

[12] Joshua Letchford, Dmytro Korzhyk, and Vincent Conitzer. 2014. On the Value of Commitment. *Autonomous Agents and Multi-Agent Systems* 28, 6 (nov 2014), 986–1016. https://doi.org/10.1007/s10458-013-9246-9

[13] Joshua Letchford, Liam MacDermed, Vincent Conitzer, Ronald Parr, and Charles L. Isbell. 2012. Computing Optimal Strategies to Commit to in Stochastic Games. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*. AAAI Press, Toronto, Ontario, Canada, 1380–1386.

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

[15] Roxana Rădulescu, Timothy Verstraeten, Yijie Zhang, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2022. Opponent Learning Awareness and Modelling in Multi-Objective Normal Form Games. *Neural Computing and Applications* 34, 3 (Feb. 2022), 1759–1781. https://doi.org/10.1007/s00521-021-06184-3

[16] 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)*.

[17] 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. https://doi.org/10.1613/jair.3987

[18] 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

[19] Willem Röpke, Roxana Radulescu, Diederik M. Roijers, and Ann Nowe. 2021. Communication Strategies in Multi-Objective Normal-Form Games. In *Proceedings of the Adaptive and Learning Agents Workshop 2021 (ALA-21)*.

[20] Willem Röpke, Diederik M. Roijers, Ann Nowé, and Roxana Radulescu. 2021. Preference Communication in Multi-Objective Normal-Form Games. *CoRR* abs/2111.09191 (2021). arXiv:2111.09191 https://arxiv.org/abs/2111.09191

[21] Willem Röpke, Diederik M. Roijers, Ann Nowé, and Roxana Rădulescu. 2021. On Nash Equilibria in Normal-Form Games With Vectorial Payoffs. arXiv:2112.06500 [cs.GT]

[22] 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

[23] 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

[24] 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

[25] 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 Twenty-Seventh International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Stockholm, Sweden, 5494–5501. https://doi.org/10.24963/ijcai.2018/775

[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] Peter Vamplew, Richard Dazeley, Ewan Barker, and Andrei Kelarev. 2009. Constructing Stochastic Mixture Policies for Episodic Multiobjective Reinforcement Learning Tasks. In *AI 2009: Advances in Artificial Intelligence*, Ann Nicholson and Xiaodong Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 340–349.

[28] Peter Vamplew, Cameron Foale, and Richard Dazeley. 2022. The Impact of Environmental Stochasticity on Value-Based Multiobjective Reinforcement Learning. *Neural Computing and Applications* 34, 3 (Feb. 2022), 1783–1799. https://doi.org/10.1007/s00521-021-05859-1

[29] Heinrich von Stackelberg. 2011. *Market Structure and Equilibrium* (first ed.). Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12586-7

[30] Bernhard von Stengel and Shmuel Zamir. 2010. Leadership games with convex strategy sets. *Games and Economic Behavior* 69, 2 (2010), 446–457. https://doi.org/10.1016/j.geb.2009.11.008

[31] A P Wierzbicki. 1995. Multiple criteria games — Theory and applications. *Journal of Systems Engineering and Electronics* 6, 2 (1995), 65–81.

[32] A Zapata, A M Mármol, L Monroy, and M A Caraballo. 2019. A Maxmin Approach for the Equilibria of Vector-Valued Games. *Group Decision and Negotiation* 28 (2019), 415–432. Issue 2. https://doi.org/10.1007/s10726-018-9608-4

[33] 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.

[34] 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.