Opponent Modelling using Policy Reconstruction for Multi-Objective Normal Form Games

Yijie Zhang
Informatics Institute
Universiteit van Amsterdam
The Netherlands
yijie.zhang@student.uva.nl

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

Patrick Mannion
School of Computer Science
National University of Ireland Galway
Ireland
patrick.mannion@nuigalway.ie

Diederik M. Roijers
Microsystems Technology
HU Univ. of Appl. Sci. Utrecht
The Netherlands
diederik.yamamoto-roijers@hu.nl

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

ABSTRACT

In many multi-agent interactions in the real world, agents receive payoffs over multiple distinct criteria; i.e. the payoffs are multi-objective in nature. However, the same multi-objective payoff vector may lead to different utilities for each participant. Therefore, it is essential for an agent to learn about the behaviour of other agents in the system. In this work, we present the first study of the effects of such opponent modelling on multi-objective multi-agent interactions with non-linear utilities. Specifically, we consider multi-objective normal form games with non-linear utility functions under the scalarised expected returns optimisation criterion. We contribute a novel actor-critic formulation to allow reinforcement learning of mixed strategies in this setting, along with an extension that incorporates opponent policy reconstruction using conditional action frequencies. Empirical results in five different MONFGs demonstrate that opponent modelling can drastically alter the learning dynamics in this setting. When equilibria are present opponent modelling can confer significant benefits on agents that implement it. However, when there are no Nash equilibria, opponent modelling can have adverse effects on utility, and has a neutral effect at best (after extensive hyper-parameter optimisation).

KEYWORDS

Multi-agent systems; multi-objective decision making; reinforcement learning; opponent modelling; game theory; solution concepts; Nash equilibrium

1 INTRODUCTION

Game theory classically studies multi-agent decision making with one-dimensional payoffs [15]. Many real-life decision problems however, are much more intricate. For example, while hammering out a contract for building a new piece of software, the different agents may care about price, delivery time, functionality, and so on. In other words, many multi-agent decision problems are inherently multi-objective [21].

In such multi-objective settings, the utility derived from the payoffs may differ from agent to agent. For example, imagine a multi-player online game where a team of players does a quest together. The quest will lead to the same expected amount of experience points, loot and currency for each player in the team. However, depending on their level, class, and playing style, different agents may care about these objectives differently, leading to different individual utilities. While the expected payoffs may be common knowledge, the utility each agent would derive from these payoffs may be private information. Furthermore, it may even be non-trivial for the individual agents to quantify these utilities for themselves [27]. In such cases, it is critical to study the emergent behaviour after multiple interactions as the agents learn more about each other. In other words, it is key to look at it from a reinforcement learning perspective.

An elegant model to study agent interaction in multi-objective settings is the *multi-objective normal form game (MONFG)* [2, 22]. To date, most papers studying MONFGs have considered different – specifically multi-objective – equilibria, which are often agnostic about the utility functions of the individual agents [3, 25]. Furthermore, most research implicitly assumes that the agents are interested in the expected utility of the payoff vector of a single play. This is called the *expected scalarised returns (ESR)* optimality criterion [17]. However, in many games, especially when the game is played multiple times, agents may instead be interested in the utility of the expected payoff (over multiple plays), which is called the *scalarised expected returns (SER)* optimality criterion. As we will study repeated interaction and long-term rewards, befitting the reinforcement learning setting, we are interested in the SER criterion. Recent work by Rădulescu et al. [20] demonstrated that the difference between ESR and SER in MONFGs can drastically alter the equilibria, and that, under SER, Nash equilibria (NE) need not exist at all.

The payoffs in MONFGs are common knowledge, but the utilities the agents derive from these are not. It therefore is important to learn about the opponents, i.e., other agents, through interaction. In this paper, we investigate whether opponent modelling (OM) can be of benefit in reinforcement learning for multi-objective multi-agent decision making problems under SER. Although opponent modelling techniques have a long history of use within the MAS community [1], to date their potential applications to multi-objective multi-agent systems (MOMAS) have not been comprehensively explored.

The contributions of this paper are:

(1) Using an actor-critic framework, we develop the first reinforcement learning methods that can learn stochastic best response strategies for MONFGs under SER.
(2) We contribute a novel algorithm developed specifically for opponent modelling in MONFGs under SER with non-linear utility functions.
(3) We provide the first empirical evidence that opponent modelling can confer significant advantages in MONFGs under SER with non-linear utility functions when Nash equilibria are present. When both agents implement opponent modelling, opponent modelling can increase the probability of converging to (better) Nash equilibria.
(4) When NE are present, we demonstrate that when only a single agent implements opponent modelling, there is an increased likelihood of converging to the best Nash equilibrium for that agent.
(5) Our experimental results show that when no NE are present, opponent modelling can in fact have adverse effects on the utility of agents that implement it. This is the first time such a phenomenon has been noticed. We note that the adverse effects can be mitigated by careful hyperparameter tuning, after which opponent modelling will have a neutral effect on utility at best.

We note that contribution 4 and 5 are surprising, but consistent with the single-objective literature. In the single-objective literature, opponent modelling has been shown to usually improve utility [1], which is not the case under SER without NE. However, in the equivalent single objective settings NE always exist [14], so the situation that there are no NE never occurs.

The next section of the paper introduces the necessary background material. In Section 3 we introduce our novel actor-critic algorithm along with an extension for opponent modelling. Section 4 presents an experimental evaluation of our proposed algorithms in several different MONFGs. Section 5 surveys related prior work on opponent modelling. Finally, Section 6 concludes the paper with some closing remarks and a discussion of promising directions for future research.

2 BACKGROUND

In this section, we discuss the necessary background material on MONFGs, multi-objective optimisation criteria, utility functions, solution concepts, opponent modelling and actor-critic algorithms.

2.1 Multi-Objective Normal Form Games

We are interested in a setting where multiple agents, each having different preferences with respect to the objectives, are interacting and learning in order to optimise the utility they receive. We use the framework of multi-objective normal form games (MONFG) to model the agents' interactions.

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

- $N = \{1, \ldots, n\}$ is a finite set of agents.
- $\mathcal{A} = A_1 \times \cdots \times A_n$, where A_i is the finite action set of agent i (i.e., the pure strategies of i). An *action (pure strategy) profile* is a vector $\mathbf{a} = (a_1, \ldots, a_n) \in \mathcal{A}$.
- $\mathbf{p} = (\mathbf{p_1}, \ldots, \mathbf{p_n})$, where $\mathbf{p_i} \colon \mathcal{A} \to \mathbb{R}^d$ is the vectorial payoff of agent i, given an action profile.

We adopt a utility-based perspective [18], by assuming that for each agent there exists a utility function that maps its vectorial payoffs to a scalar utility.

2.1.1 Utility Functions. In multi-objective normal-form games, the term payoff is used to denote the numeric vector received by agents after each interaction. As mentioned above, we also assume each agent i has a utility function that maps this payoff to a scalar value: $u_i \colon \mathbb{R}^d \to \mathbb{R}$, where d is the number of objectives.

In general, we only require that the utility functions u_i belong to the class of monotonically increasing functions, i.e., given two joint strategies $\boldsymbol{\pi}$ and $\boldsymbol{\pi'}$: $(\forall o,\ p^{\boldsymbol{\pi}}_{i,o} \geq p^{\boldsymbol{\pi'}}_{i,o}) \Rightarrow u_i(\mathbf{p}^{\boldsymbol{\pi}}_i) \geq u_i(\mathbf{p}^{\boldsymbol{\pi'}}_i)$, where $p^{\boldsymbol{\pi}}_{i,o}$ is the payoff in objective o for agent i when the agents follow a joint strategy $\boldsymbol{\pi}$. In other words, if the value of one strategy is superior in at least one objective, we expect to maintain the same ranking after applying the utility function.

We are interested in the setting of repeated interactions, while going beyond the widely used class of linear utility functions, i.e., $u_i(\mathbf{p}) = \sum_{o=1}^{d} w_{i,o} \cdot p_{i,o}$, and considering more general function classes. Furthermore, while the payoffs in MONFGs are known to the players, the utility that each agent derives from it remains hidden from the other agents. Learning about other agents through repeated interactions then becomes an essential component for allowing one to reach favourable outcomes.

2.1.2 Optimisation Criteria. In MONFGs each agent aims to optimise its utility. The utility of an agent can be derived by applying its utility function to its received payoffs. Contrary to single-objective games however, it matters when the utility function is applied. We can distinguish between two options [18, 19]: i) first computing the expectation over the payoffs obtained according to a strategy $\boldsymbol{\pi}$ and only after applying the utility function is denoted as the *scalarised expected returns* (SER) approach:

$$p_{u,i} = u(\mathbb{E}[\mathbf{p}^{\boldsymbol{\pi}}_i]); \tag{1}$$

ii) first applying the utility function before computing the expectation leads to the expected scalarised returns (ESR) approach:

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

The distinction between these options only appears when considering non-linear utility functions [19]. The choice between these criteria depends on what an agent is interested in optimising. ESR should be chosen when what matters is the utility of the payoff vector after every single interaction. Most previous research on MONFGs implicitly assumes ESR [4, 13]. Contrary, SER is more natural in the case of repeated interactions, as in SER the average payoff over multiple interactions determines the utility. SER is the most common choice in the reinforcement learning (RL) literature [18], and has recently been analysed in MONFGs [20, 21]. As we are interested in learning over repeated interactions, we focus on SER.

2.1.3 Solution concepts for MONFGs. In a MONFG under SER, a Nash equilibrium (NE) [14] is defined as a set of strategies for each agent, such that no agent can increase her SER by deviating from the equilibrium joint strategy [20].

Definition 2.2 (Nash equilibrium in a MONFG under SER). A mixed strategy profile π^{NE} is a Nash equilibrium in a MONFG under SER if for all $i \in \{1, ..., N\}$ and all $\pi_i \in \Pi_i$, with Π_i the set of mixed strategies for agent i:

$$u_i\left[\,\mathbb{E}\,\mathbf{p}_i(\pi_i^{NE}, \pi_{-i}^{NE})\right] \geq u_i\left[\,\mathbb{E}\,\mathbf{p}_i(\pi_i, \pi_{-i}^{NE})\right] \tag{3}$$

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

Recent work [20] has demonstrated that NE need not exist in MONFGs under SER with non-linear utility functions; whether any NE exist in this setting depends on the payoff scheme of the MONFG and the utility functions of the agents. Given the lack of theoretical results for the behaviour or learning dynamics for these cases, it is interesting to experimentally determine and characterise the output in these settings.

2.2 Opponent Modelling

As the agents do not know each other's utility functions, it becomes increasingly important to explicitly learn about the other agents. For such opponent modelling, we consider here the approach of policy reconstruction using conditional action frequencies [1]. This implies that an agent will maintain a set of beliefs regarding the strategy of the opponent. Similar to the idea introduced for Opponent Modelling Q-learning [24], joint-action learners [8] and fictitious play [9], we consider empirical distributions derived from observing the actions of the opponent over a window of size w.

Let $\kappa_t^i(a)$ be a timestep dependent counter for $\{t - w + 1, \ldots, t\}$ kept by agent i for every action $a \in A_{-i}$ taken by the other agent. The probability that agent i assigns to agent $-i$ playing a at timestep t is defined as:

$$Pr_t^i(a) = \frac{\kappa_t^i(a)}{\sum_{a' \in A_{-i}} \kappa_t^i(a')} \tag{4}$$

These probabilities can then be used by agent i to represent the policy π_{-i} of her opponent and to derive the valuation of her actions by marginalising out the opponent's strategy.

2.3 Policy Gradient and Actor-Critic

Policy gradient [23, 26] is a family of reinforcement learning algorithms that directly learns a policy $\pi_{\boldsymbol{\theta}}$ parameterised by $\boldsymbol{\theta}$ instead of indirectly inferring a policy based on value functions in value-based methods. After defining an objective function $J(\boldsymbol{\theta})$, policy gradient methods calculate the gradients of the objective w.r.t to $\boldsymbol{\theta}$ using the agent's actual experiences $(s_t, a_t \sim \pi(\cdot|s_t; \boldsymbol{\theta}_t), r_t)_{t=0}^T$ from interacting with the environment and update the parameters $\boldsymbol{\theta}$ to improve the return from the objective function. Although there are many different ways to approximate the true gradients, all policy gradient methods share the same form of updates:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha \nabla J(\boldsymbol{\theta}_t) \tag{5}$$

In addition to policy gradient methods, there is another powerful class of learning methods, called actor-critic methods, which learn a policy, referred to as the actor, and a value function, referred to as the critic [23]. Compared to vanilla policy gradient methods that only use obtained rewards to compute gradients, the critic can usually reduces the variance in gradients and achieves a more stable policy update. In this way, actor-critic methods can be seen as combining both the policy- and value-based methods.

3 OPPONENT MODELLING IN MONFGS

In this paper, we investigate the effects of opponent modelling in the setting of MONFGs under SER with non-linear utility functions. We focus on understanding if opponent modelling can speed up learning or confer a significant advantage for agents who implement it in this setting. Furthermore, when considering MONFGs under SER, we also investigate whether there is a difference in the observed effect of opponent modelling in games with Nash equilibria, compared to games without Nash equilibria.

To investigate the effects of opponent modelling in MONFGs under SER, we design an actor-critic algorithm specially adapted for this framework to optimise SER. There are key benefits to choosing an actor-critic method. Compared to vanilla policy-based methods, actor-critic methods use a learned value function to reduce variance and ensure a stable policy update. And, more importantly, compared to value-based methods actor-critic methods allow the agents to learn an explicitly stochastic policy. Like in policy-based methods, this enables effective exploration and exploitation strategies that are signigicantly better than the often-used hard-coded epsilon-greedy exploration in value-based methods. More importantly however, stochastic policies are essential for the SER optimality criterion, as even if the opponents policy is fixed, the best response may still necessarily be stochastic. Therefore, enabling such explicitly stochastic policies is a significant improvement over recent work on reinforcement learning in MONFGs [20], which used Q-learning with ϵ-greedy to learn best responses based on pure strategies only.

Given an agent's own utility function, when considering maximising its SER, a natural representation of the expected returns is the expectation of action values over the agent's action distribution, i.e, its stochastic policy $\pi_{\boldsymbol{\theta}}$. Let us represent the multi-objective action value vector by $\boldsymbol{Q}(a)$ and the stochastic policy by $\pi(a|\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$. Then, for this agent, we have its SER objective defined as:

$$J(\boldsymbol{\theta}) = u\left(\sum_{a \in \mathcal{A}} \pi(a|\boldsymbol{\theta}) \boldsymbol{Q}(a)\right) \tag{6}$$

where u is the non-linear utility function, $a \in \mathcal{A}$ is an action available to the agent, π the policy of the agent parameterised by $\boldsymbol{\theta}$ and $\boldsymbol{Q}(a) \in \mathbb{R}^d$ is the multi-objective action value vector that can be learned by different means (e.g. Eqn 7). More specifically, $\sum_a \pi(a|\boldsymbol{\theta}) \boldsymbol{Q}(a)$ is the expected multi-objective return vector; by applying the utility function and optimising this quantity, the agent is able to learn a best response mixed strategy under SER.

Next, we propose a base algorithm without opponent modelling as well as an algorithm with opponent modelling within the actor-critic framework.

3.1 Actor-Critic for MONFGS

To optimise SER, we have to take the gradients of $J(\boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$. We divide this into 2 iterative steps. Note that in the SER objective, the action values $\boldsymbol{Q}(a)$ are initialised arbitrarily in the beginning and need to be learned. So the first step is to learn the multi-objective action value vector $\boldsymbol{Q}(a)$. After an action a is chosen using its own policy $\pi(a|\boldsymbol{\theta})$, the agent observes a vectorial payoff $\boldsymbol{p}$ and we use a simple stateless Q-learning update rule (as per [20]) to learn $\boldsymbol{Q}(a)$:

$$\boldsymbol{Q}(a_t) \leftarrow \boldsymbol{Q}(a_t) + \alpha_Q[\boldsymbol{p}_t - \boldsymbol{Q}(a_t)] \tag{7}$$

where α_Q is the learning rate for Q-learning. After the action values have been updated, the objective J can be calculated and analytically derived and we perform the second step to update $\boldsymbol{\theta}$ in the direction of maximising SER:

$$\boldsymbol{\theta_{t+1}} \leftarrow \boldsymbol{\theta_t} + \alpha_\theta \nabla J(\boldsymbol{\theta}_t) \tag{8}$$

where α_θ is the learning rate for policy update.

In this way, as we iterate the above 2 steps, both the action values $\boldsymbol{Q}(a)$ and policy $\pi_{\boldsymbol{\theta}}$ are learned.

3.2 Actor-Critic with Opponent Modelling

When considering opponent modelling, an intuitive approach is to model the opponent's policy π' directly; the simplest way is to represent the opponent's policy as an empirical distribution of action frequencies $\pi'(a')$. By using this modelling approach, the agent is able to aggregate information about the opponent's decision patterns and hence use it to improve its own policy.

In order to combine opponent modelling with our actor-critic algorithm, some extensions have to be made. Firstly, instead of learning $\boldsymbol{Q}(a)$, a new joint action value $\boldsymbol{Q}(a, a')$ will be learned to estimate the expected vectorial payoff for each possible joint action. After each episode, combining the updated $\boldsymbol{Q}(a, a')$ and estimate of the opponent's policy π', the agent will be able to evaluate the expected utility of its next action. However, since stochastic policies are used by both the agent and its opponent, the uncertainty from both strategies should be taken into account. As a result, we can naturally extend the SER objective by marginalising over the opponent's actions a', with the new SER objective $J(\boldsymbol{\theta})$ defined as $J(\boldsymbol{\theta}) = u\left(\mathbb{E}_{\pi(a|\theta)}\mathbb{E}_{\pi(a')}\left[Q\left(a, a'\right)\right]\right)$. By maximising this SER objective, the agent is able to average over all the uncertainty to compute the best mixed strategy in terms of scalarised expected return. More specifically, the SER objective can be expressed as:

$$J(\boldsymbol{\theta}) = u\left(\sum_{a_t \in A} \pi(a_t|\boldsymbol{\theta}) \sum_{a'_t \in A'} \pi'\left(a'_t\right) \boldsymbol{Q}\left(a_t, a'_t\right)\right) \tag{9}$$

We now introduce the algorithm with opponent modelling based on the vanilla actor-critic algorithm that we developed in the previous section in Algorithm 1.

Algorithm 1: Actor-Critic for MONFGS with OM

```
Input: MONFG $G$, number of episodes $M$, learning rates
       $\alpha_Q$, $\alpha_{\boldsymbol{\theta}}$, window size $w$, opponent's action history
       buffer $h$, utility function $u$ for each player.
Output: For each player: the policies $\pi(a|\boldsymbol{\theta})$, the joint
        Q-functions $\boldsymbol{Q}(a, a')$.
For each player, initialize $\boldsymbol{\theta}$ as a zero-vector, initialize
  softmax policy $\pi\left(a = a_i|\boldsymbol{\theta}\right) = \frac{e^{\theta_i}}{\sum_{j=1}^{|A_i|} e^{\theta_j}}$ and initialize joint
  Q-functions $\boldsymbol{Q}(a, a')$ as a zero-tensor.
for each episode $t$ do
    for each player $i$ do
        Play sampled action $a \sim \pi(a|\boldsymbol{\theta})$ and observe
          opponent's action $a'$.
        Observe multi-objective payoff $\boldsymbol{p}$.
        Append opponent's last action $a'$ to the buffer $h$.
        Estimate opponent's action distribution $\pi'(a')$ from
          buffer $h$.
        Update own joint Q-function:
          $\boldsymbol{Q}\left(a, a'\right) \leftarrow \boldsymbol{Q}\left(a, a'\right) + \alpha_Q\left[\boldsymbol{p} - \boldsymbol{Q}\left(a, a'\right)\right]$.
        Calculate own objective:
          $J(\boldsymbol{\theta}) = u(\sum_{a \in A} \pi(a|\boldsymbol{\theta}) \sum_{a' \in \mathcal{A}'} \pi'(a')\boldsymbol{Q}\left(a, a'\right))$.
        Update own policy parameters: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha_\theta \nabla J(\boldsymbol{\theta})$.
```

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

Table 1: Game 1 - A MONFG which has one pure strategy NE in (L,M) under SER, with expected payoffs of 10 and 3.

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

Table 2: Game 2 - A MONFG which has pure strategy NE in (L,L) – payoffs (17, 4), and (M,M) – payoffs (13, 6), under SER. Note that (L,L) offers the highest utility for the row player, whereas (M,M) offers the highest utility for the column player.

4 EXPERIMENTS

To evaluate the impact of opponent modelling, we use multiple 2-player 2-objective MONFGs with different properties. In all these MONFGs, we consider the utility functions as defined in [20]; the row player's utility function is:

$$u_1([p^1, p^2]) = p^1 \cdot p^1 + p^2 \cdot p^2, \tag{10}$$

while the column player's utility function is:

$$u_2([p^1, p^2]) = p^1 \cdot p^2. \tag{11}$$

We first introduce Game 1 (Table 1) that has one NE in pure strategies under SER: (L,M). Secondly, we create a MONFG with multiple NE, referred to as Game 2 (Table 2). There are two equilibria in this case: (L,L) and (M,M). (L,L) offers the highest utility for the row player, while (M,M) is the preferred outcome for the

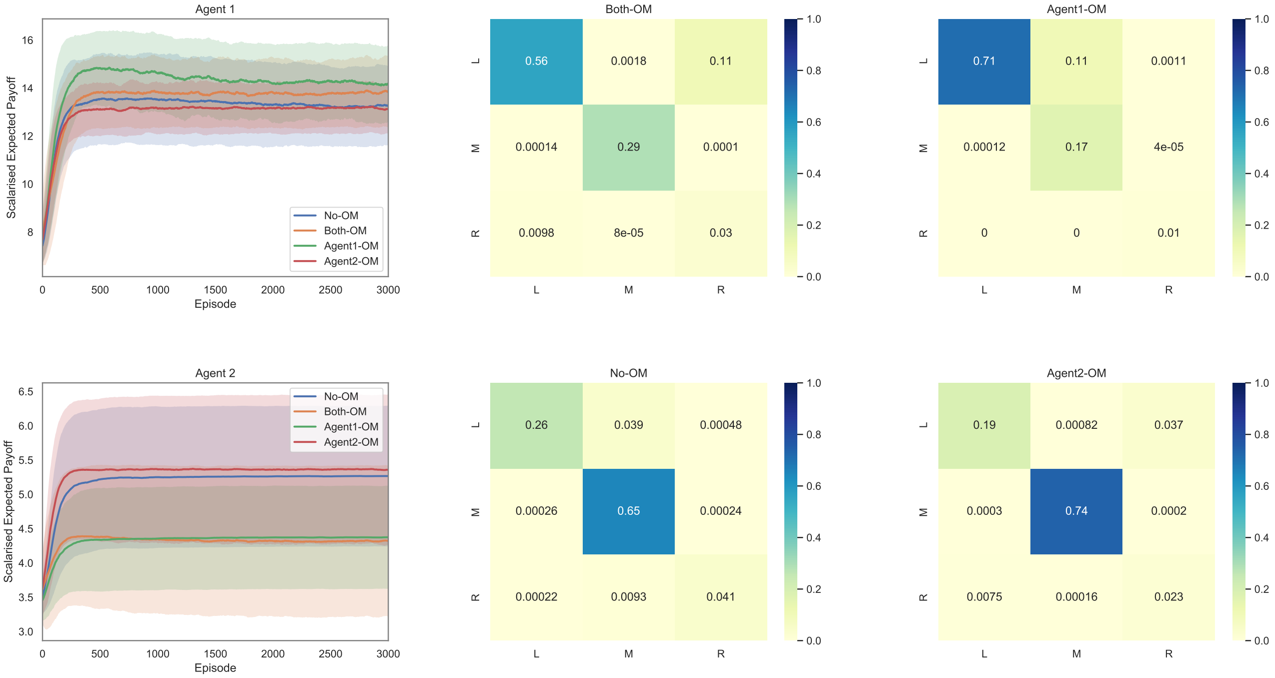

Figure 1: Results on Game 3. The left column shows the estimated SER for Agent 1 (top) and Agent 2 (bottom) under the 4 experiment settings. The middle and right columns show the empirical outcome distributions.

	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 3: Game 3 - A MONFG which has pure strategy NE in (L,L) – payoffs (17, 4), (M,M) – (13, 6), and (R,R) – (10, 3), under SER. Note that (L,L) and (M,M) Pareto-dominate (R,R), and that (L,L) offers the highest utility for the row player, whereas (M,M) offers the highest utility for the column player.

	L	M
L	$(4, 0)$	$(2, 2)$
M	$(2, 2)$	$(0, 4)$

Table 4: Game 4 - There are no NE in this game under SER.

column player. This will allow us to focus closely on the competition between the agents for reaching their preferred equilibrium. For the third MONFG with NE, we extend the previous game to 3-actions, in Game 3 (Table 3), having 3 pure Nash equilibria (i.e., (L,L), (M,M), (R,R)) under SER with the specified utility functions. The (R,R) NE is Pareto-dominated by the other equilibria. Again, (L,L) is the best outcome for the row player in terms of utility, while (M,M) is preferred by the column player.

	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 5: Game 5 - The (Im)balancing act MONFG from Rădulescu et al. [20]. There are no NE in this game under SER.

We also conduct experiments using two MONFGs without any NE under SER. We introduce Game 4 (Table 4), and we also use the (Im)balancing Act MONFG, which we will refer to as Game 5 (Table 5, originally introduced in [20]). Both of these games exhibit similar dynamics when the players use the utility functions in Eqns. 10 and 11. To get the highest utility, agent 1 (row) wishes to make the objectives as unbalanced as possible, whereas agent 2 (column) wishes to make the objectives as balanced as possible. Because of the structure of the payoffs, it is never possible to reach a stable equilibrium in pure or mixed strategies, as one or other of the agents always has an incentive to deviate to its preferred pure strategy to gain extra utility [20].

For each game, we consider four different settings:

(1) Setting 1: neither agent performs opponent modelling.
(2) Setting 2: both agents perform opponent modelling.
(3) Setting 3: only agent 1 performs opponent modelling.

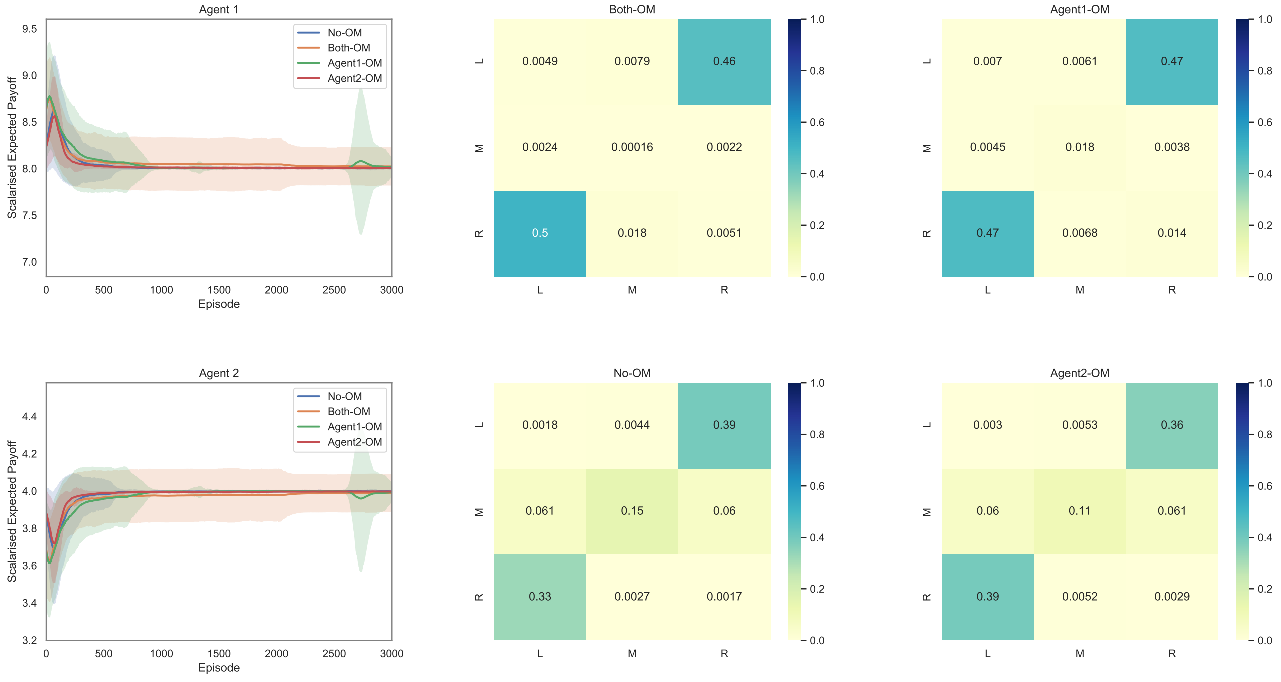

Figure 2: Results on Game 5. The left column shows the estimated SER for Agent 1 (top) and Agent 2 (bottom) under the 4 experiment settings. The middle and right columns show the empirical outcome distributions.

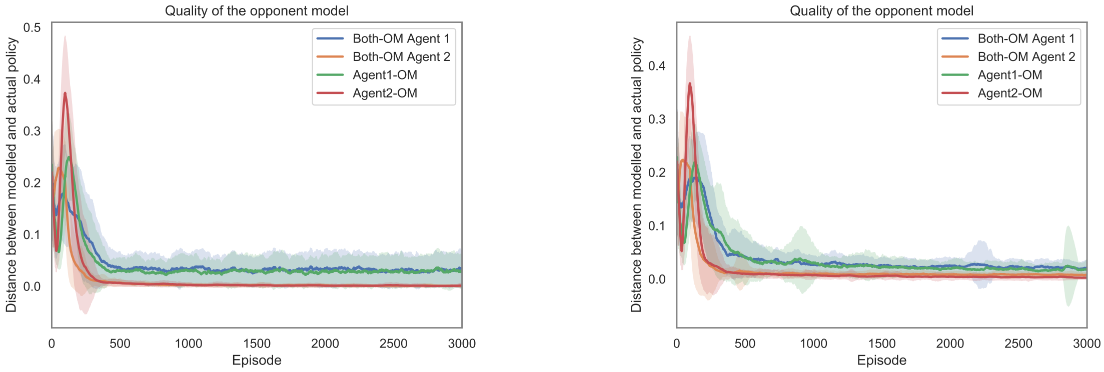

Figure 3: Quality of the opponent models for Game 3 (left) and Game 5 (right).

(4) Setting 4: only agent 2 performs opponent modelling.

For each setting, agents interact for 3000 episodes, averaged over 100 trials. Furthermore, in this experiment, the gradient ∇_θ is computed analytically w.r.t $J(\theta)$. An agent's strategy $\pi(a|\theta)$ is represented using a simple softmax function:

$$\pi(a = a_i|\theta) = \frac{e^{\theta_i}}{\sum_{j=1}^{|A_i|} e^{\theta_j}} \tag{12}$$

The actor learning rate for the presented experimental results is $\alpha_\theta = 0.05$, while h, the opponent modelling window size, is set to 100. For the setting without opponent modelling we used a critic learning rate $\alpha_Q = 0.05$. For the Opponent Modelling Actor-Critic approach, because the agents are learning the Q-function for the join-action space in a deterministic setting, we used $\alpha_Q = 1$. However, we note that we carried out an extensive analysis with respect to all these parameters and we present all the results and observations below.

Note that with any kind of opponent modelling, the agent is generally expected to make better decisions than without opponent modelling because more information is used. However, the performance depends on the quality of the model, as making decisions based on wrong or obsolete information can lead to detrimental results. To quantify which of the four settings leads to a better performance in the framework of MONFGs under SER, we consider three representative criteria, namely the estimated SER, the empirical distribution over the possible outcomes of the game for the last 500 rounds, and a measure for assessing the quality of the opponent model (i.e., the maximum absolute value of the difference between the predicted and the actual policy). The first criterion is easy to understand from an optimisation perspective; since we are trying to maximise SER, the setting with a higher SER is deemed better. The second criterion could be understood from a game theory perspective, where a setting is evaluated by whether it increases the frequency of favourable outcomes for an agent (e.g., reaching preferred Nash equilibria under SER in Game 3). Finally, with our last criterion we try to capture the difficulty of estimating the opponent's policy (e.g., an agent's policy might never converge and always cycle between actions) and understand whether agents are basing their decisions on accurate information about the opponent's strategy.

The reported SER of an agent's mixed strategy is calculated for each trial of 3000 episodes (where an episode consists of a single interaction), on a rolling basis, using a window of 100. The payoffs $\boldsymbol{p}$ the agent received are combined and averaged to obtain the expected return vector. Then the corresponding utility function u is applied on top of this expected return in order to obtain an empirical estimate of the SER.

We present here only the experimental results figures for Game 3 and Game 5. However, we note that we observe the same trend for the rest of our results, depending on whether equilibria are present or not in the considered MONFGs.

4.1 MONFGs with Nash Equilibria

Games 1 and 2. For these settings the results for the SER are highly similar to Game 3. Regarding the empirical outcome distribution for Game 1, the agents manage to reach the NE with higher probability (i.e., ≈ 99%) when both agents are using opponent modelling (OM), compared to the no OM setting (i.e., ≈ 81%). For Game 2 when only agent 1 is performing the OM, the probability of converging to the (L,L) equilibrium is ≈ 79%, while when only agent 2 is performing the OM the probability of converging to the (M,M) equilibrium is ≈ 77%. These results show that using OM can increase the probability of converging to an agent's preferred equilibrium point (when NEs exist).

Game 3. Figure 1 shows the empirical SER and empirical outcome distributions for Game 3. The left columns show both agents' estimated SER for the 4 experimental settings, where the mean and 1 standard deviation confidence regions are shown over 100 statistical trials. We can already notice here that the single-sided use of OM does confer each agent with an advantage, in terms of obtained payoff under SER. Analysing further the results, the middle and right columns show the empirical outcome frequencies averaged over 100 trials for the 4 experimental settings. Without OM, we observe that (M,M) is the most common outcome (≈ 65%) and that (L,L) is the next most common outcome (≈ 26%). When only agent 1 implements OM, the probability of reaching her preferred outcome (L,L) increases (≈ 71%). When only agent 2 implements OM, the probability of her preferred outcome (M,M) increases (≈ 74%). When both agents implement OM, agent 1 gains an advantage as the probability of reaching its preferred outcome increases (≈ 56% vs. ≈ 26% without OM). However, the combined probability of reaching one of the NE actually decreases (≈ 91% without OM vs. ≈ 85% with OM). This somewhat surprising result can be explained by considering the prediction quality plot in Figure 3; we observe that agent 1 consistently has difficulty in computing an accurate estimate of the strategy of agent 2. This highlights the fact that while policy reconstruction using conditional action frequencies is generally useful to agents that implement it in MONFGs with NE under SER, it may actually increase the risk of miscoordination between the agents if the opponent policy estimates are inaccurate.

4.2 MONFGs without Nash Equilibria

Game 4. For this setting opponent modelling did not improve the SER or learning speed for agents. On the contrary, in most settings, the agent performing OM seemed to be unable to accurately capture information regarding the opponent's strategy, and thus made decisions on the basis of incorrect or outdated information. These outcomes were obtained particularly when setting α_Q for the OM Actor-Critic approach to a lower value (e.g., 0.05). At best, we found (after extensive hyperparameter optimisation) that OM can have a neutral contribution to the obtained SER, similar to the results presented for Game 5, and agents are converging to either (L,M) or (M,L) with almost equal probability, whether OM is used or not.

Game 5. From the left column of Figure 2, we can see that the estimated SER for both agent 1 and agent 2 is very similar under all four experimental settings. From the middle and right columns on Figure 2, we observe that the empirical distributions of outcomes share similar structures, i.e., most of the probability density is concentrated in outcomes (R,L) and (L,R). In this game, agent 1 wants the objectives to be as unbalanced as possible, whereas agent 2 wants the objectives to be as balanced as possible. Implementing OM does not confer a significant advantage in terms of outcomes, especially for agent 1 who could gain a large amount of additional utility by increasing the frequency of the unbalanced outcomes (L,L) and (R,R). As with Game 4 above, using a lower value of α_Q can cause OM to decrease the obtained utility for both agents; OM has at best a neutral effect in settings with no NE. This is likely because the opponent model based on action history is always inaccurate due to the opponent rapidly changing its policy, as no stable equilibrium point can be reached when both agents are performing policy updates after each episode.

5 RELATED WORK

Here we present a brief overview of prior work on MONFGs and opponent modelling, with a specific focus on works which are closely related to the contributions of this paper. A comprehensive survey of opponent modelling techniques is presented by Albrecht and Stone [1].

Since their introduction in [2], MONFGs have mostly been considered under linear utility functions or implicitly assuming the ESR criterion [3, 4, 13]. Rădulescu et al. [20] revisit MONFGs and explicitly distinguish between ESR and SER, under non-linear utility functions. They demonstrate the effect of using different criteria on the set of Nash and correlated equilibria. We note, however, that their experimental framework only incorporates a simplistic Q-learning approach, using ϵ-greedy as an action selection mechanism, thus restricting their agents to only learning deterministic strategies.

A straightforward method for opponent modelling in reinforcement learning is building a model of the other agents' policy. Opponent Modelling Q-learning [24] extends Q-learning in a similar manner to our approach for extending the critic: it calculates the probability distribution of the opponent actions from the observed behaviour, and then derives the best action for the agent by marginalising out the opponent's actions from the state – joint-action Q-table.

Opponent modelling has also been incorporated in RL methods based on neural function approximators, by augmenting the model with a module that is able to predict the action of the other agent [11, 12]. Finally, goal prediction is another approach for opponent modelling, presented in Self Other-Modeling [16], where the agent uses his own policy in order to learn to predict the goal of the other agent.

In multi-objective settings, another choice for modelling other agents is to build a model of their utility functions. However, this task is far from trivial, and an important idea is to use, anytime possible, knowledge regarding the structure of the utility space. For example, Chajewska and Koller [5] build a probabilistic model for utilities elicited from a population of users and show how one can find a factorisation of the utility function. More recently, one can notice the use of Bayesian frameworks in the form of Gaussian Processes to model utility functions [7, 10, 27]. Using an active learning approach, one can also intertwine the decision making process (based on partial utility information), with a querying process in order to elicit additional utility information [6, 27], however revealing information regarding one's preferences will not always be in the best interest of agents, especially in competitive settings. Hence, the agent would need to extract preference information from interactions, which is far from trivial.

6 CONCLUSION AND FUTURE WORK

In this work, we presented the first study on the effects of opponent modelling in multi-objective multi-agent settings under the SER optimisation criterion. In contrast to much prior work on opponent modelling in multi-criteria problems, we considered opponents with non-linear utility functions. We adopted the MONFG model for our experimental evaluations of the effects of opponent modelling under SER with non-linear utility functions. A novel formulation of actor-critic for this setting was introduced, along with an extension that incorporates opponent modelling via policy reconstruction using conditional action frequencies. Empirical results in five different MONFGS (three with Nash equilibria, and two without under the SER criterion) demonstrated that opponent modelling can significantly alter the learning dynamics of a MONFG. In cases where NE are present, opponent modelling can confer significant benefits on agents that implement it. However, when there are no NE, we observe that an agent that implements opponent modelling, while its opponent does not, can experience adverse effects on its utility. These adverse effects could be (mostly) mitigated after careful hyper-parameter optimisation of the learning algorithm, but did not contribute to the utility of the agent implementing the opponent modelling. This is highly surprising, and does not occur in the single-objective setting – where there are always NE in mixed strategies.

This study has a number of limitations, leaving much scope for future research to build upon the present work. As we adopted the MONFG model, our analysis considered stateless decision making problems only; therefore this line of work should be extended to sequential settings such as multi-objective stochastic games (MOSGs) [21]. Although estimating the opponent's strategy using empirical action frequencies in MOMAS provided promising initial results, there is scope to develop more complex opponent modelling techniques, such as estimating the opponent's utility function directly from observed behaviour. Provided that the opponent's preferences are fixed, we aim to learn sufficiently accurate utility-based opponent models that are less likely to suffer from the effects of outdated information that we observed with policy reconstruction (e.g. in Games 4 and 5). Gaussian processes are a promising candidate for such utility function estimations; e.g. recent work has adopted Gaussian processes to estimate non-linear utility functions in single agent decision support settings [27].

Furthermore, our experimental evaluations were limited to games with two agents only, so there is much work to be done on opponent modelling in larger MOMAS. In many real world settings (e.g. online games such as MMORPGs, or political negotiations between multiple states), the utility functions of agents in the environment often have varying degrees of alignment to one another. Therefore an agent that can effectively model opponent utility could make predictions about the intentions (i.e. cooperative vs. competitive) of other agents, based on the degree of alignment of an estimated opponent utility function with her own private utility function.

As multi-objective multi-agent decision making is a relatively under-explored area of MAS research, many significant and interesting open questions remain within the field. The choice of optimisation criterion (ESR versus SER) can have drastic effects on the set of equilibria in MOMAS. We already made the highly surprising observation that opponent modelling can have adverse effects under SER, when there are no NE. We want to further investigate this phenomenon. Firstly, in games under SER without NE, we aim to investigate whether explicitly modelling the (properties of) the utility function of the opponent can make opponent modelling effective again. Larger MOMAS may contain agents that choose different optimisation criteria; this could add further complications when determining the conditions for a stable outcome to be reached. While we have proposed several new MONFGs in this paper, in future work it would be worthwhile to develop a larger set of standardised benchmarks that could be used to evaluate the performance of algorithms in a variety of multi-objective multi-agent decision making settings, e.g., cooperative and competitive games, negotiations, and sequential settings.

ACKNOWLEDGMENTS

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaandere" program.

REFERENCES

[1] Stefano V. Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (2018), 66 – 95.

[2] David Blackwell et al. 1956. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.* 6, 1 (1956), 1–8.

[3] PEM Borm, SH Tijs, and JCM Van Den Aarssen. 1988. Pareto equilibria in multiobjective games. *Methods of Operations Research* 60 (1988), 302–312.

[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.

[5] Urszula Chajewska and Daphne Koller. 2000. Utilities as random variables: Density estimation and structure discovery. In *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 63–71.

[6] Urszula Chajewska, Daphne Koller, and Ronald Parr. 2000. Making rational decisions using adaptive utility elicitation. In *AAAI/IAAI*. 363–369.

[7] Wei Chu and Zoubin Ghahramani. 2005. Preference learning with Gaussian processes. In *Proceedings of the 22nd international conference on Machine learning*. ACM, 137–144.

[8] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* 1998, 746-752 (1998), 2.

[9] Drew Fudenberg, Fudenberg Drew, David K Levine, and David K Levine. 1998. *The theory of learning in games*. Vol. 2. MIT press.

[10] Shengbo Guo, Scott Sanner, and Edwin V Bonilla. 2010. Gaussian process preference elicitation. In *Advances in neural information processing systems*. 262–270.

[11] He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. In *International Conference on Machine Learning*. 1804–1813.

[12] Stefan JL Knegt, Madalina M Drugan, and Marco A Wiering. 2018. Opponent Modelling in the Game of Tron using Reinforcement Learning.. In *ICAART (2)*. 29–40.

[13] Dmitrii Lozovanu, D Solomon, and A Zelikovsky. 2005. Multiobjective games and determining pareto-nash equilibria. *Buletinul Academiei de Ştiinţe a Republicii Moldova. Matematica* 3 (2005), 115–122.

[14] John Nash. 1951. Non-Cooperative Games. *Annals of Mathematics* 54, 2 (1951), 286–295.

[15] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. 2007. *Algorithmic game theory*. Cambridge university press.

[16] Roberta Raileanu, Emily Denton, Arthur Szlam, and Rob Fergus. [n.d.]. Modeling Others using Oneself in Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning (ICML)*.

[17] 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 FAIM*.

[18] 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.

[19] Diederik M Roijers and Shimon Whiteson. 2017. Multi-objective decision making. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11, 1 (2017), 1–129.

[20] Roxana Rădulescu, Patrick Mannion, Diederik M. Roijers, and Ann Nowé. 2019. Equilibria in Multi-Objective Games: a Utility-Based Perspective. In *Proceedings of the Adaptive and Learning Agents Workshop (ALA-19) at AAMAS*.

[21] 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, 10 (2020). https://doi.org/10.1007/s10458-019-09433-x

[22] Lloyd S Shapley and Fred D Rigby. 1959. Equilibrium points in games with vector payoffs. *Naval Research Logistics Quarterly* 6, 1 (1959), 57–61.

[23] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press. http://incompleteideas.net/book/the-book-2nd.html

[24] William Uther and Manuela Veloso. 1997. *Adversarial reinforcement learning*. Technical Report. Technical report, Carnegie Mellon University, 1997. Unpublished.

[25] Mark Voorneveld, Dries Vermeulen, and Peter Borm. 1999. Axiomatizations of Pareto equilibria in multicriteria games. *Games and economic behavior* 28, 1 (1999), 146–154.

[26] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.* 8, 3-4 (May 1992), 229–256. https://doi.org/10.1007/BF00992696

[27] Luisa M Zintgraf, Diederik M Roijers, Sjoerd Linders, Catholijn M Jonker, and Ann Nowé. 2018. Ordered preference elicitation strategies for supporting multi-objective decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1477–1485.