The Knowledge Engineering Review, Vol. 35, e32, 1 of 21.
doi:10.1017/S0269888920000351

A utility-based analysis of equilibria in multi-objective normal-form games

ROXANA RĂDULESCU[1], PATRICK MANNION[2], YIJIE ZHANG[3], DIEDERIK M. ROIJERS[4,1], and ANN NOWÉ[1]

[1]*Artificial Intelligence Lab, Vrije Universiteit Brussel, Pleinlaan 2, Brussels 1050, Belgium*
e-mails: roxana.radulescu@vub.be, ann.nowe@vub.be
[2]*School of Computer Science, National University of Ireland Galway, Galway H91 TK33, Ireland*
e-mail: patrick.mannion@nuigalway.ie
[3]*Universiteit van Amsterdam, Amsterdam, The Netherlands*
e-mail: yijie.zhang@student.uva.nl
[4]*Microsystems Technology, HU University of Applied Sciences Utrecht, Heidelberglaan 15, 3584CS Utrecht, The Netherlands*
e-mail: diederik.yamamoto-roijers@hu.nl

Abstract

In multi-objective multi-agent systems (MOMASs), agents explicitly consider the possible trade-offs between conflicting objective functions. We argue that compromises between competing objectives in MOMAS should be analyzed on the basis of the utility that these compromises have for the users of a system, where an agent's utility function maps their payoff vectors to scalar utility values. This utility-based approach naturally leads to two different optimization criteria for agents in a MOMAS: expected scalarized returns (ESRs) and scalarized expected returns (SERs). In this article, we explore the differences between these two criteria using the framework of multi-objective normal-form games (MONFGs). We demonstrate that the choice of optimization criterion (ESR or SER) can radically alter the set of equilibria in a MONFG when nonlinear utility functions are used.

1 Introduction

Multi-agent systems (MASs) are ideally suited to model a wide range of real-world problems where autonomous actors participate in distributed decision-making. Example application domains include urban and air traffic control (Yliniemi *et al*., 2015; Mannion *et al*., 2016a), autonomous vehicles (Rădulescu *et al*., 2018; Talpert *et al*., 2019), and energy systems (Walraven & Spaan, 2016; Mannion *et al*., 2016b; Reymond *et al*., 2018). Although many such problems feature multiple conflicting objectives to optimize, most MAS research focuses on agents maximizing their return w.r.t. a single objective. By contrast, in multi-objective multi-agent systems (MOMASs), agents explicitly consider the possible trade-offs between conflicting objective functions. Agents in a MOMAS receive vector-valued payoffs for their actions, where each component of a payoff vector represents the performance on a different objective. Following the utility-based approach (Roijers *et al*., 2013), we assume that each agent has a utility function which maps vector-valued payoffs to scalar utility values. Compromises between competing objectives are then considered on the basis of the utility that these trade-offs have for the users of a MOMAS.

The utility-based approach naturally leads to two different optimization criteria for agents in a MOMAS: expected scalarized returns (ESRs) and scalarized expected returns (SERs). To date, the

[*] This article extends an earlier unpublished paper (Rădulescu *et al*., 2019) that was originally presented at the Adaptive and Learning Agents Workshop 2019.

differences between the SER and ESR approaches have received little attention in multi-agent settings, despite having received some attention in single-agent settings (see, e.g., Roijers *et al.*, 2013, 2018). Consequently, the implications of choosing either ESR or SER as the optimization criterion for a MOMAS are currently not well understood. In this work, we use the framework of multi-objective normal-form games (MONFGs) to explore the differences between ESR and SER in multi-agent settings.

In MASs, solution concepts such as Nash equilibria (Nash, 1950, 1951) and correlated equilibria (Aumann, 1974, 1987) specify conditions under which each agent cannot increase its expected payoff by deviating unilaterally from an equilibrium strategy. Such solution concepts are well studied in single-objective settings, to capture stable multi-agent behavior. However, in utility-based MOMAS, the notion of an equilibrium must be redefined, as incentives to deviate from equilibrium strategies are now computed based on the relative utilities of vector-valued payoffs, rather than the relative values of scalar payoffs. Furthermore, the choice of optimization criterion (ESR or SER) influences how equilibria are computed, as agents' incentives to deviate from an equilibrium strategy may be measured in terms of either differences in ESR or differences in SER.

The contributions of this work are

1. We provide the first comprehensive analysis of the differences between the ESR and SER optimization criteria in multi-agent settings.
2. We provide formal definitions of the criteria for Nash equilibria and correlated equilibria under ESR and SER.
3. We prove that the ESR and SER criteria are equivalent in cases where linear utility functions are used.
4. We demonstrate that the choice of optimization criterion radically alters the set of equilibria in a MONFG.
5. We propose two versions of correlated equilibria for MONFGs—single-signal and multi-signal—corresponding to different use cases.
6. We prove that in MONFGs under SER with nonlinear utility functions, Nash equilibria and multi-signal correlated equilibria need not exist, whereas single-signal correlated equilibria can exist. We find that whether these equilibria exist in a specific MONFG depends on the multi-objective payoff structure and the utility functions used. These examples are supported by empirical results.
7. We demonstrate empirically that the well-known previous findings that CE can provide better payoffs than NE in single-objective games (see e.g., Aumann, 1974) can also apply in the more general class of multi-objective games, that is, that in a MOMAS where a coordination signal can be established CE can potentially lead to higher utility for both agents than NE.

The next section of this paper introduces and discusses normal-form games (NFGs), relevant solution concepts, and optimization criteria for multi-objective decision-making. Section 3 provides an overview of prior work on multi-objective games. Section 4 formally defines Nash equilibria and correlated equilibria in MONFGs under ESR and SER and discusses some important theoretical considerations arising from these definitions. Section 5 presents empirical results in support of the conclusions reached in Section 4. Finally, Section 6 concludes with a summary of our findings a discussion of important open questions and promising directions for future work.

2 Background

2.1 NFGs and equilibria

Normal-form (strategic) games (NFGs) constitute a fundamental representation of interactions between players in game theory. Players are seen as rational decision-makers seeking to maximize their payoff. When multiple players are interacting, their strategies are interrelated, each decision depending on the choices of the others. For this reason, we usually try to determine interesting groups of outcomes, called solution concepts. Below, we offer a formal definition for NFG and discuss two well-known solution concepts considered in this work: Nash equilibria and correlated equilibria.

DEFINITION 1 (Normal-form game) *An n-person finite normal-form game G is a tuple* $(N, \mathcal{A}, \mathbf{p})$, *with* $n \geq 2$, *where:*

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

2.1.1 Mixed strategy profile

Let us denote by $P(X)$ the set of all probability distributions over X. We can then define the set of mixed strategies of player i as $\Pi_i = P(A_i)$. The set of *mixed strategy profiles* is then the Cartesian product of all the individual mixed strategy sets $\Pi = \Pi_1 \times \ldots \times \Pi_n$.

We define $\pi_{-i} = (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$ to be a strategy profile without player's i strategy. We can thus write $\pi = (\pi_i, \pi_{-i})$.

A NE (Nash, 1951) can be defined based on a pure or mixed strategy profile, such that each player has selected her best response to the other players' strategies. We offer a more formal definition below.

DEFINITION 2 (Nash equilibrium) *A mixed strategy profile π^{NE} of a game G is a NE if for each player $i \in \{1, \ldots, N\}$ and for any alternative strategy $\pi_i \in \Pi_i$:*

$$\mathbb{E}p_i\left(\pi_i^{NE}, \pi_{-i}^{NE}\right) \geq \mathbb{E}p_i\left(\pi_i, \pi_{-i}^{NE}\right) \tag{1}$$

Thus, under a NE, no player i can improve her payoff by unilaterally changing her strategy. The same definition applies for pure strategy profiles. Nash (1951) has proven that, allowing the use of mixed strategies, any finite NFG has at least one NE.

A CE is a game-theoretic solution concept proposed by Aumann (1974) in order to capture correlation options available to the players when some form of communication can be established prior to the action selection phase (i.e., the players receive signals from an external device, according to a known distribution, allowing them to correlate their strategies). For the current work, we look at correlation signals taking the form of action recommendations.

2.1.2 Correlated strategy

A correlated strategy represents a probability vector σ on $\mathcal{A}$ that assigns probabilities for each possible action profile, that is, $\sigma : \mathcal{A} \to [0, 1]$. The expected payoff of player i given a correlated strategy σ is calculated as:

$$\mathbb{E}p_i(\sigma) = \sum_{\mathbf{a} \in \mathcal{A}} \sigma(\mathbf{a}) p_i(\mathbf{a})$$

2.1.3 Strategy modification

A strategy modification for player i is a function $\delta_i : A_i \to A_i$, such that given a recommendation a_i, player i will play action $\delta_i(a_i)$ instead. The expected payoff of player i given a correlated strategy σ and a strategy modification δ_i is calculated as:

$$\mathbb{E}p_i(\delta_i(\sigma)) = \sum_{\mathbf{a} \in \mathcal{A}} \sigma(\mathbf{a}) p_i(\delta_i(a_i), a_{-i})$$

DEFINITION 3 (Correlated equilibrium) *A correlated strategy σ^{CE} of a game G is a CE if for each player $i \in \{1, \ldots, N\}$ and for any possible strategy modification δ_i:*

$$\mathbb{E}p_i(\sigma^{CE}) \geq \mathbb{E}p_i(\delta_i(\sigma^{CE})) \tag{2}$$

Thus, a CE ensures that no player can gain additional payoff by deviating from the suggestions, given that the other players follow them as well. Although this definition strongly resembles the one of NE, there is one important aspect we emphasize here, namely the distinction between a mixed strategy profile and a correlated strategy. Mixed strategy profiles are composed of independent probability factors, while the action probabilities in correlated strategies are jointly defined.

Table 1. Payoff matrix for the game of Chicken

	S	D
S	6, 6	2, 7
D	7, 2	0, 0

Table 2. A possible CE for the game of Chicken

	S	D
S	0.5	0.25
D	0.25	0

Correlated equilibria can be computed via linear programming in polynomial time (Papadimitriou & Roughgarden, 2008). It has been also shown that no-regret algorithms converge to CE (Foster & Vohra, 1999). Furthermore, CE has the same existence guarantees in finite NFGs (Hart & Schmeidler, 1989) as NE, and any NE is an instance of a CE (Aumann, 1987).

2.1.4 *Example*

Consider the game of Chicken with the payoffs described in Table 1. Each player has two actions: to continue driving toward the other player (D) or to swerve the car (S).

There are three well-known Nash equilibria for this game with expected payoffs $(7, 2)$, $(2, 7)$—pure strategy NE—and $(4\frac{2}{3}, 4\frac{2}{3})$—mixed strategy NE where each player selects S and D with probabilities $\frac{2}{3}$ and $\frac{1}{3}$, respectively.

A possible CE is represented in Table 2, by assigning 0.5 probability for the joint action (S, S), 0.25 for (D, S), and finally 0.25 for (S, D). The expected payoff for this CE is $(5\frac{1}{4}, 5\frac{1}{4})$, values higher than the ones obtained under any NE. Thus, the notion of CE not only extends NE, but it also offers the potential for obtaining higher expected payoffs when players are able to receive a correlation signal (e.g., a recommended action).

2.2 *Multi-objective normal-form games*

DEFINITION 4 (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 players.
- $\mathcal{A} = A_1 \times \cdots \times A_n$, where A_i is the finite action set of player 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 $\mathbf{p_i} : \mathcal{A} \to \mathbb{R}^d$ is the vectorial payoff of player i, given an action profile.

In this work, we adopt a utility-based perspective (Roijers *et al.*, 2013) and assume that each agent has a utility function that maps her vectorial payoff to a scalar utility value. A more detailed discussion of utility functions can be found in Section 2.4.

2.3 *Multi-objective optimization criteria*

When agents consider multiple conflicting objectives, they should balance these in such a way that the user utility derived from the outcome of a decision problem (such as a MONFG) is maximized. This is known as the utility-based approach (Roijers *et al.*, 2013). Following this approach, we assume that there exists a utility function that maps a vector with a value for each objective to a scalar utility:

$$p_{u,i} = u_i(\mathbf{p}_i) \tag{3}$$

where $p_{u,i}$ is the utility that agent i derives from the vector $\mathbf{p}_i$. When deciding what to optimize in a MONFG, we thus need to apply this function to the vector-valued outcomes of the decision problem in some way. There are two choices for how to do this (Roijers *et al.*, 2013; Roijers & Whiteson, 2017). Computing the expected value of the payoffs of a joint strategy first and then applying the utility function leads to the SERs optimization criterion, that is,

$$p_{u,i} = u(\mathbb{E}[\mathbf{p}_i \mid \boldsymbol{\pi}]) \tag{4}$$

where π is the joint strategy for all the agents in a MONFG and $\mathbf{p}_i$ is the payoff received by agent i. SER is employed in most of the multi-objective planning and reinforcement learning literature. Alternatively, the utility function can be applied before computing the expectation, leading to the ESRs optimization criterion (Roijers *et al.*, 2018), that is,

$$p_{u,i} = \mathbb{E}[u(\mathbf{p}_i) \mid \boldsymbol{\pi}] \tag{5}$$

Which of these criteria should be considered best depends on how the games are used in practice. SER is the correct criterion if a game is played multiple times, and it is the average payoff over multiple plays that determines the user's utility. ESR is the correct formulation if the payoff of a single play is what is important to the user.

2.4 Utility functions

From a single-objective game-theoretic perspective, the notions of utility and payoff functions are generally used interchangeably. When transitioning to the multi-objective domain, we usually denote by payoff function the vectorial return (containing a real-valued payoff for each objective) received by a player, given an action profile. The utility (scalarization) function is then used to denote the mapping from this vectorial return to a scalar utility value for a player i: $u_i : \mathbb{R}^d \to \mathbb{R}$.

Linear combinations are a widely used canonical example of a scalarization function:

$$u_i(\mathbf{p}_i) = \sum_{d \in D} w_d p_{i,d} \tag{6}$$

where D is the set of objectives, $\mathbf{w}$ is a weight vector[1], $w_d \in [0, 1]$ is the weight for objective d, and $p_{i,d}$ is the payoff for objective d received by agent i. Nonlinear, discontinuous utility functions may arise in the case where it is important for an agent to achieve a minimum payoff on one of the objectives; such a utility function may look like the following:

$$u_i(\mathbf{p}_i) = \begin{cases} p_{i,t_d} & \text{if } p_{i,d} \geq t_d \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $p_{i,d}$ represents the expected payoff for agent i on objective d, t_d is the required threshold value for d, and p_{i,t_d} is the utility to agent i of reaching the threshold value on d.

Utility functions may not always be known *a priori* and/or may not be easy to define depending on the setting. For example, in the *decision support scenario* (Roijers *et al.*, 2013), it may not be possible for users to specify utility functions directly, instead users may be asked to provide their preferences by scoring or ranking different possible outcomes. After the preference elicitation process is complete, users' responses may then be used to model their utility functions (Zintgraf *et al.*, 2018).

3 Related work

Since their introduction in Blackwell *et al.* (1956), multi-objective (multicriteria) games have been discussed extensively in the literature. Below, we present a non-exhaustive overview of this work, highlighting a few differences with the current considered perspective.

[1] A vector whose coordinates are all non-negative and sum up to 1.

Most previous work in multi-objective games considers utility-function agnostic equilibria, that is, the agents do not know their preferences. For this case, Shapley and Rigby (1959) extend and characterize the set of mixed strategy agnostic Nash equilibria for multicriteria two-person zero-sum games for linear utility functions: joint strategies that are undominated w.r.t. unilateral changes by either agent. They also note that if the preference functions differ, the scalarized game (implicitly assuming ESR) can possibly be no longer zero-sum. While the idea that utility functions could also be nonlinear is discussed by Bergstresser and Yu (1977), for analysis purposes they only consider linear utility functions and derive solution points from the resulting trade-off games. This is important because, as we will discuss in Section 4.2, there is no in-practice difference between ESR and SER in the linear case. The existence of Pareto[2] equilibria for two-person multi-objective games under linear utility functions is proven by Borm *et al.* (1990). A further characterization of Pareto equilibria can be found in Voorneveld *et al.* (1999).

Considering non-cooperative games, Wierzbicki (1995) states that, in realistic scenarios, how to aggregate criteria might not be known; however, some form of scalarization function is necessary in order to compute possible solutions. This corresponds to explicitly taking the user utility into account, and we therefore fully agree with this approach. Conflict escalation and solution selection are discussed when considering linear or order-consistent scalarization functions. Lozovanu *et al.* (2005) formulate an algorithm for finding Pareto–Nash equilibria in multi-objective non-cooperative games (i.e., for every linear utility function for which the weights sum to one, compute the trade-off game, then find its NE). Finally, Lozan and Ungureanu, (2013) propose a method for computing Pareto–Nash equilibrium sets, also under linear utility functions. A third approach is to elicit preferences, that is, information about the utility function, while determining equilibria (Igarashi & Roijers, 2017). As far as we know, however, this also has only been done for linear utility functions.

Notice that, despite the fact that many works admit that it might not always be desirable for a player to share full information about her utility function or that utility functions could take any form (including nonlinear), most analysis and theoretical contributions use linear utility functions only. Furthermore, the utility function is directly applied on the original game in order to derive and analyze the corresponding trade-off game, which corresponds to the ESR case. However, due to the use of linear utility functions, there is no distinction to be made between the ESR and SER optimization criteria, as we will show in Section 4.2. Interestingly enough, the field of multi-objective (single-agent) reinforcement learning typically focuses on the SER case (Vamplew *et al*., 2011; Van Moffaert & Nowé, 2014; Zintgraf *et al*., 2015; Mossalam *et al*., 2016), while in either field this vital choice is typically not made explicitly or explained in the individual papers. In this paper, we aim to make the choice between an ESR and SER perspective explicit and show that this choice has profound consequences in MOMAs.

For a comprehensive overview of prior work on multi-objective multi-agent decision-making, the interested reader is referred to a recent survey article by Rădulescu *et al*. (2020).

4 Computing equilibria in MONFGS

Now that we have covered the necessary background, we begin our exploration of the differences between the ESR and SER optimization criteria in MOMAS. In Section 4.1, we formally define Nash equilibria and correlated equilibria in MONFGs under either ESR or SER. In Section 4.2, we discuss several important theoretical considerations arising from these definitions and introduce a new MONFG for this purpose. Section 4.3 introduces some additional games, which we analyze from the SER perspective.

4.1 Definitions

As agents in MOMAS seek to optimize the utility of their vector-valued payoffs, rather than the value of scalar payoffs in single-objective settings, the standard solution concepts must be redefined based on the agents' utilities. Incentives to deviate from an equilibrium strategy may be defined based on utility,

[2] While the original paper refers to this type of equilibrium as 'Pareto', we note that Pareto is a too loose domination concept when considering only linear utility functions and would prefer 'Convex' in this case. For consistency, however, we keep the original term.

specifically the difference between the utility of an equilibrium action and the utilities of other possible actions. Here, we reformulate the conditions for Nash equilibria (Equation (1)) and correlated equilibria (Equation (2)) under the ESR optimization criterion (Equation (5)) and the SER optimization criterion (Equation (4)).

DEFINITION 1 (NE in a MONFG under ESR) *A mixed strategy strategy profile π^{NE} is a NE in a MONFG under ESR if for all $i \in \{1, ..., N\}$ and all $\pi_i \in \Pi_i$:*

$$\mathbb{E}u_i\big(\mathbf{p}_i(\pi_i^{NE}, \pi_{-i}^{NE})\big) \geq \mathbb{E}u_i\big(\mathbf{p}_i(\pi_i, \pi_{-i}^{NE})\big) \tag{8}$$

that is, π^{NE} is a NE under ESR if no agent can increase the expected utility of her payoffs by deviating unilaterally from π^{NE}.

DEFINITION 2 (NE in a MONFG under SER) *A mixed strategy profile π^{NE} is a NE in a MONFG under SER if for all $i \in \{1, ..., N\}$ and all $\pi_i \in \Pi_i$:*

$$u_i\big(\mathbb{E}\mathbf{p}_i(\pi_i^{NE}, \pi_{-i}^{NE})\big) \geq u_i\big(\mathbb{E}\mathbf{p}_i(\pi_i, \pi_{-i}^{NE})\big) \tag{9}$$

that is, π^{NE} is a NE under SER if no agent can increase the utility of her expected payoffs by deviating unilaterally from π^{NE}.

DEFINITION 3 (CE in a MONFG under ESR) *A probability vector σ^{CE} on $\mathcal{A}$ is a CE in a MONFG under ESR if for all players $i \in \{1, ..., N\}$ and for all strategy modifications δ_i:*

$$\mathbb{E}u_i\big(\mathbf{p}_i(\sigma^{CE})\big) \geq \mathbb{E}u_i\big(\mathbf{p}_i(\delta_i(\sigma^{CE}))\big) \tag{10}$$

that is, σ^{CE} is a CE under ESR if no agent can increase the expected utility of her payoffs by deviating unilaterally from the action recommendations in σ^{CE}.

When applying the SER optimization criterion for CE, there are two cases we can distinguish between, due to the two expectations that CE incorporates for every player i. First, we can define the expected payoff given a signal a_i^r due to the uncertainty about the other players' actions. Second, we can define the expected payoff given the correlated strategy (i.e., a certain probability distribution over the joint action space). Depending on where we place the utility function for taking the scalarized expectation, we distinguish between the *single-signal* and *multi-signal* cases.

4.1.1 Single-signal CE under SER

In the case of a single-signal CE, we assume that the signal is only given once, and that the expected payoffs over which the utility must be computed is conditioned on the signal. Even if the MONFG is played multiple times, the signal does not change. An example of a single persistent signal in a multi-agent decision problem can be a smart grid in which the correlation signal corresponds to the price of electricity in a longer interval (e.g., one or more hours), and the actions of the agents are whether to perform a given task or not within a small interval (e.g., 10 min). In such cases, the utility of the other signals that might have been possible do not matter; they did not occur. Hence, the agent must maximize the utility of its expected vector-valued payoff given a single signal. Or, if the signal is not known at plan time, for each signal separately.

DEFINITION 4 (Single-signal CE in a MONFG under SER) *A probability vector σ^{CE} on $\mathcal{A}$ is a single-signal CE in a MONFG under SER if for all players $i \in \{1, ..., N\}$, given a recommended action a_i^r, and for any strategy modification δ_i on a_i^r (i.e., for any alternative action $a_i \neq a_i^r$):*

$$u_i\big(\mathbb{E}\big[\mathbf{p}_i(\sigma^{CE}) \mid a_i^r\big]\big) \geq u_i\big(\mathbb{E}\big[\mathbf{p}_i(\delta_i(\sigma^{CE})) \mid a_i^r\big]\big) \tag{11}$$

that is, σ^{CE} is a single-signal CE under SER if no agent can increase the *utility of her expected payoffs* by deviating unilaterally from the given action recommendation in σ^{CE}.

In order to emphasize and clarify how the expected payoff is conditioned on a given recommended action a_i^r, we can re-write this definition in a more explicit form (by expanding the conditional expected payoffs):

$$u_i\left(\frac{\sum_{a_{-i}\in\mathcal{A}_{-i}}\sigma^{CE}(a_i^r,a_{-i})\mathbf{p}_i(a_i^r,a_{-i})}{\sum_{a_{-i}\in\mathcal{A}_{-i}}\sigma^{CE}(a_i^r,a_{-i})}\right)\geq u_i\left(\frac{\sum_{a_{-i}\in\mathcal{A}_{-i}}\sigma^{CE}(a_i^r,a_{-i})\mathbf{p}_i(a_i,a_{-i})}{\sum_{a_{-i}\in\mathcal{A}_{-i}}\sigma^{CE}(a_i^r,a_{-i})}\right) \tag{12}$$

4.1.2 Multi-signal CE under SER

The single-signal CE for MONFGs assumes that even if the MONFG is played multiple times, there will be one possible signal. Alternatively, the signal may change every time the game is played, that is, the scalarization is performed after marginalizing over the entire correlated strategy probability distribution.

DEFINITION 5 (Multi-signal CE in a MONFG under SER) *A probability vector σ^{CE} on $\mathcal{A}$ is a multi-signal CE in a MONFG under SER if for all players $i \in \{1, ..., N\}$ and for any strategy modification δ_i:*

$$u_i\big(\mathbb{E}\mathbf{p}_i(\sigma^{CE})\big)\geq u_i\big(\mathbb{E}\mathbf{p}_i(\delta_i(\sigma^{CE}))\big) \tag{13}$$

that is, σ^{CE} is a multi-signal CE under SER if no agent can increase the utility of her expected payoffs by deviating unilaterally from the given action recommendations in σ^{CE}.

Notice that while the ESR case is equivalent to solving the CE for the corresponding single-objective trade-off game, the SER case leads to a much more complicated situation. In a general case, when no restriction is imposed on the form of the utility function, we may end up having to solve a nonlinear optimization problem.

4.2 Theoretical considerations

THEOREM 1 *Every finite MONFG where each agent seeks to maximize the expected utility of its payoff vectors (ESR) has at least one NE.*

Proof In the ESR case, any MONFG can be reduced to its corresponding single-objective trade-off game G', as players will apply the utility function on their payoff vectors after every interaction. We proceed with showing how one can construct G'.

Consider the following finite normal-form game $G' = (N, \mathcal{A}, f)$, where N and $\mathcal{A}$ are the same as in the original MONFG. According to Definition 1, the payoff function for G': $f = (f_1, \ldots, f_n)$.

We define each component $f_i : \mathcal{A} \to \mathbb{R}$ as the composition between player's i utility function $u_i : \mathbb{R}^d \to \mathbb{R}$ and her vectorial payoff function $\mathbf{p}_i : \mathcal{A} \to \mathbb{R}^d$:

$$f_i(a) = (u_i \circ \mathbf{p}_i)(a) = u_i(\mathbf{p}_i(a)), \forall a \in \mathcal{A}$$

Thus, in the ESR case, any MONFG is reduced to a corresponding single-objective trade-off finite NFG that can be constructed as shown above. According to the NE existence theorem (Nash, 1951), the resulting finite NFG G' has at least one NE. □

THEOREM 2 *In finite MONFGs, when linear utility functions are used, the ESR and SER optimization criteria are equivalent*[3].

Proof Let π^{NE} be the NE strategy profile under the ESR optimization criteria, according to Definition 1 and for each player i let u_i be a linear scalarization function, according to Equation (6).

Due to the fact that u_i is a linear function, Jensen's inequality (Jensen *et al.*, 1906) allows us to rewrite each term of Equation (8) as follows:

$$\mathbb{E}u_i\big(\mathbf{p}_i(\pi_i^{NE}\cup\pi_{-i}^{NE})\big)=u_i\big(\mathbb{E}\mathbf{p}_i(\pi_i^{NE}\cup\pi_{-i}^{NE})\big) \tag{14}$$

$$\mathbb{E}u_i\big(\mathbf{p}_i(\pi_i\cup\pi_{-i}^{NE})\big)=u_i\big(\mathbb{E}\mathbf{p}_i(\pi_i\cup\pi_{-i}^{NE})\big) \tag{15}$$

[3] As is the case for single-agent decision problems (Roijers *et al.* 2013; Roijers, 2016).

Table 3. The (im)balancing act game

	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 4. The (Im)balancing act game under ESR with utility functions $u_1(\mathbf{p}) = p^1 \cdot p^1 + p^2 \cdot p^2$ and $u_2(\mathbf{p}) = p^1 \cdot p^2$ applied

	L	M	R
L	(16, 0)	(10, 3)	(8, 4)
M	(10, 3)	(8, 4)	(10, 3)
R	(8, 4)	(10, 3)	(16, 0)

Notice that by replacing the terms from Equation (8) according to Equations (14) and (15), we obtain the definition of the NE under SER (Equation (9)). The same procedure can be applied for CE, to transition from Equations (10)–(13) and prove that, under a linear utility function, the ESR and SER criteria are also equivalent for CE.

When considering a more general case, with u_i being a nonlinear function, despite the fact that Jensen's inequality (Jensen *et al.*, 1906) would allow us to define inequality relations between the terms in Equations (14) and (15) (when constraining u_i to be convex or concave), we have no guarantee that the set of NE and CE remains the same under the two optimization criteria ESR and SER. Thus, no clear conclusions can be drawn when generalizing the form of the utility function. Furthermore, as we show below using a concrete example, in the general case, the ESR and SER criteria are not equivalent.

THEOREM 3 *In finite MONFGs, where each agent seeks to maximize the utility of its expected payoff vectors (SER), Nash equilibria need not exist.*

Proof Consider the following game. There are two agents that can each choose from three actions: *left*, *middle*, or *right*. The payoff vectors are identical for both agents and are specified by the payoff matrix in Table 3.

The utility functions of the agents are given by $u_1([p^1, p^2]) = p^1 \cdot p^1 + p^2 \cdot p^2$ for agent 1, and $u_2([p^1, p^2]) = p^1 \cdot p^2$ for agent 2[4]. In this game, it is easy to see that agent 1 will always want to move toward an as imbalanced payoff vector as possible, that is, concentrate as much of the value in one objective, while agent 2 will always want to move to a balanced solution, that is, spread out the value across the objectives equally. Under SER, the expectation is taken before the utility function is applied. Therefore, a mixed strategy will lead to an expected payoff vector for both agents. If the expected payoff vector is balanced, that is, $[2, 2]$, agent 1 will have an incentive to deterministically take action L or R, irrespective of its current strategy. If the payoff vector is imbalanced, for example, $[2 - x, 2 + x]$, agent 2 will have an incentive to compensate for this imbalance and play *left* more often to compensate if x is positive, and *right* more often if x is negative, and he is always able to do so. Hence, at least one of the agents will always have an incentive to deviate from its strategy, and therefore there is no NE under SER. □

We also note that under ESR there is a mixed NE the game in Table 3, that is, agent 2 plays *middle* deterministically, and agent 1 plays *left* with a probability 0.5 and *right* with a probability 0.5, leading to an expected utility of $3^2 + 1^2 = 10$ for agent 1, and $3 \cdot 1 = 3$ for agent 2. This is not a NE under SER, as the expected payoff vector is $[2, 2]$ for this strategy, and agent 1 has an incentive to play either *left* or *right* deterministically, which would lead to an expected payoff vector of $[3, 1]$ or $[1, 3]$, yielding a higher

[4] Please note that this is a monotonically increasing payoff function for positive-only payoffs. In the case of negative payoffs, we can set the utility to 0 as soon as the payoff value for one of the objectives becomes negative.

Table 5. A correlated equilibrium in the (Im)balancing act game under SER

	L	M	R
L	0	0.75	0
M	0	0	0
R	0	0.25	0

utility for agent 1 if agent 2 does not adjust its strategy. Hence, the SER and ESR cases are fundamentally different.

THEOREM 4 *In finite MONFGs, where each agent seeks to maximize the utility of its expected payoff vectors given a signal (single-signal CE under SER), correlated equilibria can exist when Nash equilibria do not.*

Proof Consider the action suggestions in Table 5 for the (Im)balancing act game.

It may easily be shown that the action suggestions in Table 5 satisfy the conditions given in Equation (12) for a single-signal CE in a MONFG under SER:

- When L is suggested to the row player, the expected payoff vectors and SER for it to play L, M, or R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.75 \cdot [3, 1])/0.75 = [3, 1]$, SER $= 3^2 + 1^2 = 10$
  - M: $\mathbb{E}(\mathbf{p}) = (0.75 \cdot [2, 2])/0.75 = [2, 2]$, SER $= 2^2 + 2^2 = 8$
  - R: $\mathbb{E}(\mathbf{p}) = (0.75 \cdot [1, 3])/0.75 = [1, 3]$, SER $= 1^2 + 3^2 = 10$
- When R is suggested to the row player, the expected payoff vectors and SER for it to play L, M, or R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [3, 1])/0.25 = [3, 1]$, SER $= 3^2 + 1^2 = 10$
  - M: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [2, 2])/0.25 = [2, 2]$, SER $= 2^2 + 2^2 = 8$
  - R: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [1, 3])/0.25 = [1, 3]$, SER $= 1^2 + 3^2 = 10$
- When M is suggested to the column player, the expected payoff vectors and SER for it to play L, M, or R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.75 \cdot [4, 0] + 0.25 \cdot [2, 2])/(0.75 + 0.25) = [3.5, 0.5]$, SER $= 3.5 \cdot 0.5 = 1.75$
  - M: $\mathbb{E}(\mathbf{p}) = (0.75 \cdot [3, 1] + 0.25 \cdot [1, 3])/(0.75 + 0.25) = [2.5, 1.5]$, SER $= 2.5 \cdot 1.5 = 3.75$
  - R: $\mathbb{E}(\mathbf{p}) = (0.75 \cdot [2, 2] + 0.25 \cdot [0, 4])/(0.75 + 0.25) = [1.5, 2.5]$, SER $= 1.5 \cdot 2.5 = 3.75$

In all the cases above, neither of the agents may increase the utility of their expected payoff vectors given the recommendations, by deviating from the suggested actions in Table 5, assuming that the other agent follows the suggestions. Therefore, CE may exist in MONFGs under SER when conditioning the expectation on a given signal, even in cases where Nash equilibria do not exist. □

THEOREM 5 *In finite MONFGs, where each agent seeks to maximize the utility of its expected payoff vectors over all the given signals (multi-signal CE under SER), correlated equilibria need not exist.*

Proof In the case of a multi-signal CE, the agents are interested in their expected payoff vectors across all possible signals. In other words, to compute the expected payoff vectors, the signal must be marginalized out first. Therefore, the CE previously discussed for the single-signal case (Table 5) is not a CE for the multi-signal case, that is, player 1 will have an incentive to deterministically take action L or R, irrespective of the given signal. If the correlated strategy tries to incorporate this tendency, player 2 will have an incentive to deviate toward the options that offer her the most balanced outcome. Hence, similar to the proof for the non-existence of Nash equilibria under SER, at least one of the agents will always have an incentive to deviate from the given recommendation, and therefore there is no multi-signal CE under SER.

Table 6. The (Im)balancing act game without action M (left), together with the corresponding correlated strategy (right)

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

	L	R
L	0.25	0.25
R	0.25	0.25

We thus conclude that a MONFG under ESR with *known* utility functions is equivalent to a single-objective NFG, and therefore all theories, including the existence of Nash equilibria and correlated equilibria, are implied. Under SER however, Nash equilibria and multi-signal correlated equilibria need not exist, and MONFGs with nonlinear utility functions are fundamentally more difficult than single-objective NFGs, even when the utility functions are known in advance.

4.3 Additional games for SER analysis

To further investigate the existence of Nash equilibria, single- and multi-signal correlated equilibria under SERs, we introduce two additional games that demonstrate different characteristics under these criteria. We consider for analysis the same nonlinear utility functions as above: $u_1([p^1, p^2]) = p^1 \cdot p^1 + p^2 \cdot p^2$ for player 1 and $u_2([p^1, p^2]) = p^1 \cdot p^2$ for player 2.

4.3.1 The (Im)balancing act game without action M

First, we derive a 2-player, 2-action, 2-objective game from the (Im)balancing act game (Table 3), by removing the middle action. The (Im)balancing act game without action M is presented in Table 6 (left).

Notice that in the case of NE and multi-signal CE, the dynamics of the game remain unchanged from the original 3-action version: player 1 will always have an incentive to deviate toward imbalanced payoffs, while player 2 desires the exact opposite. For the single-signal CE, we have the opportunity to offer the agents the chance to coordinate their actions and obtain a fair outcome over the two possible situations (i.e., a balanced outcome (2, 2) and an imbalanced outcome (4, 0) or (0, 4)), as shown in the right side of Table 6.

It may be shown that the correlated strategy in Table 6 (right) satisfies the conditions given in Equation (12) for a single-signal CE in a MONFG under SER:

- When L is suggested to the row player, the expected payoff vectors and SER for it to play L or R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [4, 0] + 0.25 \cdot [2, 2])/(0.25 + 0.25) = [3, 1]$, SER $= 3^2 + 1^2 = 10$
  - R: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [4, 0] + 0.25 \cdot [2, 2])/(0.25 + 0.25) = [3, 1]$, SER $= 3^2 + 1^2 = 10$
- When R is suggested to the row player, the expected payoff vectors and SER for it to play L or R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [4, 0] + 0.25 \cdot [2, 2])/(0.25 + 0.25) = [3, 1]$, SER $= 3^2 + 1^2 = 10$
  - R: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [4, 0] + 0.25 \cdot [2, 2])/(0.25 + 0.25) = [3, 1]$, SER $= 3^2 + 1^2 = 10$
- When L is suggested to the column player, the expected payoff vectors and SER for it to play Lor R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [4, 0] + 0.25 \cdot [2, 2])/(0.25 + 0.25) = [3, 1]$, SER $= 3 \cdot 1 = 3$
  - R: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [4, 0] + 0.25 \cdot [2, 2])/(0.25 + 0.25) = [3, 1]$, SER $= 3 \cdot 1 = 3$
- When R is suggested to the column player, the expected payoff vectors and SER for it to play Lor R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [4, 0] + 0.25 \cdot [2, 2])/(0.25 + 0.25) = [3, 1]$, SER $= 3 \cdot 1 = 3$
  - R: $\mathbb{E}(\mathbf{p}) = (0.25 \cdot [4, 0] + 0.25 \cdot [2, 2])/(0.25 + 0.25) = [3, 1]$, SER $= 3 \cdot 1 = 3$

In all the cases above, neither of the agents may increase the utility of their expected payoff vectors given the recommendations, by deviating from the suggested actions, assuming that the other agent follows the suggestions. Therefore, the signal suggested in the right side of Table 6 represents a single-signal CE for the (Im)balancing act game without action M.

Table 7. A 3-action MONFG which has three pure strategy Nash equilibria (left)—(L, L), (M, M), and (R, R) – when the row player uses utility function $u_1([p^1, p^2]) = p^1 \cdot p^1 + p^2 \cdot p^2$ and the column player uses utility function $u_2([p^1, p^2]) = p^1 \cdot p^2$, with the corresponding proposed correlated strategy (right)

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

	L	*M*	*R*
L	0.5	0	0
M	0	0.5	0
R	0	0	0

4.3.2 A 3-action MONFG with NE and CE under SER

The final game we introduce for this work presents an example of a MONFG for which all the studied equilibria (i.e., NE, single- and multi-signal CE) exist under SER (Table 7). There are three pure strategy Nash equilibria—(L, L), (M, M), and (R, R), under the nonlinear utility functions specified above. Notice that player 1 will receive the highest SER under (L, L), while player 2 will prefer the (M, M) outcome. (R, R) is also a NE, but it is Pareto dominated by (L, L) and (M, M) and does not offer the best possible SER for either agent.

Let us turn our attention to the single-signal CE. It may be shown that the correlated strategy proposed in Table 7 (right) satisfies the conditions given in Equation (12) for a single-signal CE in a MONFG under SER:

- When L is suggested to the row player, the expected payoff vectors and SER for it to play L, M, or R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [4, 1])/0.5 = [4, 1]$, SER $= 4^2 + 1^2 = 17$
  - M: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [3, 1])/0.5 = [3, 1]$, SER $= 3^2 + 1^2 = 10$
  - R: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [1, 2])/0.5 = [1, 2]$, SER $= 1^2 + 2^2 = 5$
- When M is suggested to the row player, the expected payoff vectors and SER for it to play L, M, or R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [1, 2])/0.5 = [1, 2]$, SER $= 1^2 + 2^2 = 5$
  - M: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [3, 2])/0.5 = [3, 2]$, SER $= 3^2 + 2^2 = 13$
  - R: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [2, 1])/0.5 = [2, 1]$, SER $= 2^2 + 1^2 = 5$
- When L is suggested to the column player, the expected payoff vectors and SER for it to play L, M, or R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [4, 1])/0.5 = [4, 1]$, SER $= 4 \cdot 1 = 4$
  - M: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [1, 2])/(0.5 = [1, 2]$, SER $= 1 \cdot 2 = 2$
  - R: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [2, 1])/0.5 = [2, 1]$, SER $= 2 \cdot 1 = 2$
- When M is suggested to the column player, the expected payoff vectors and SER for it to play L, M, or R are
  - L: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [3, 1])/0.5 = [3, 1]$, SER $= 3 \cdot 1 = 3$
  - M: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [3, 2])/0.5 = [3, 2]$, SER $= 3 \cdot 2 = 6$
  - R: $\mathbb{E}(\mathbf{p}) = (0.5 \cdot [1, 2])/0.5 = [1, 2]$, SER $= 1 \cdot 2 = 2$

In all the cases above, neither of the agents may increase the utility of their expected payoff vectors given the recommendations, by deviating from the suggested actions, assuming that the other agent follows the suggestions. Therefore, the signal suggested in the right side of Table 7 represents a single-signal correlated equilibria.

In single-objective NFGs, it is known that any convex combination of NE payoff profiles can be reached or achieved by a CE (Aumann, 1974). The relationship between Nash equilibria and correlated equilibria in MONFGs remains, however, an open question. In Section 5, we empirically test whether the proposed correlated strategy, representing a convex combination between twp pure Nash equilibria under SER is a multi-signal CE as well. We also note that in the single-objective case, CE can achieve payoffs that lie outside the convex hull of NE payoffs, again a property not validated in the case of MONFGs.

5 Experiments

To demonstrate the effect of the SER optimization criterion on equilibria in MONFGs, with no action recommendations and in the case of a *single- and multi-signal CE*, we conducted a series of experiments using the games introduced in the previous section in Tables 3, 6 and 7. All experiments were repeated 100 times and had a duration of 10 000 episodes, where the MONFG game was played once per episode[5].

Agents implemented a simple algorithm[6] to learn estimates of the expected vectors for each action according to the following update rule (i.e., a 'one-shot' vectorial version of Q-learning Watkins, 1989):

$$\mathbf{Q}(s_i, a_i) \leftarrow \mathbf{Q}(s_i, a_i) + \alpha[\mathbf{p}_i(s_i, a_i) - \mathbf{Q}(s_i, a_i)] \tag{16}$$

where $\mathbf{Q}(s_i, a_i)$ is an estimate of the expected value vector for selecting action a_i when a private signal s_i is received, $\mathbf{p}_i(s_i, a_i)$ is the payoff vector received by agent i for selecting action a_i when observing s_i, and α is the learning rate.

The private signals given to each agent allow us to test empirically whether agents will have an incentive to deviate from a single- or multi-signal CE in a MONFG under SER. For the experiments marked as 'No action recommendations', in each episode agents received unchanging private signals with probability 1 (i.e., equivalent to the case where no private signals are present). Otherwise, the private signals received by each agent corresponded to the correlated action recommendations indicated for each considered MONFG. When signals were given, for the first 500 episodes, both agents followed the action recommendations in their private signals deterministically, so that the CE behavior could be learned. For the last 9500 episodes, agents continued to receive action recommendations, but selected their actions autonomously.

Agents implemented the ϵ-greedy exploration strategy. As agents seek to optimize their action choices with respect to SERs, they will determine the optimal mixed strategy (given the recommendation, where applicable), with probability $1-\epsilon$, or chose a random action with probability ϵ. Agents determine their optimal mixed strategy by solving a nonlinear optimization problem with the goal of maximizing their SERs, under their utility function and current Q-values[7]. For all the experiments, the estimates of expected value vectors for each action were scalarized using the same utility functions as in Section 4.2. In the case of the single-signal CE, this expectation is taken under the given action recommendation, while for the multi-signal CE, the expectation is derived with respect to the entire CE signal the agent received, following Definitions 4 and 5, respectively. This also implies that for the multi-signal correlated equilibria, each agent has information regarding the CE distribution over her own actions, but not over the entire joint action space. For example, in the case of the (Im)balancing act game, player 1 knows that the CE distribution over her actions is [0.75, 0, 0.25], but is not aware that player 2 will be recommended action 'M' with probability 1, leaving this information to be acquired through the learning process.

All agents used a constant value of $\alpha = 0.05$ for the learning rate. For the experiments without action recommendations, ϵ was initially set to 0.1 in the first episode and decayed by a factor 0.999 in each subsequent episode. For the experiments where agents receive action recommendations, ϵ was set to 0.0 in for the first 500 episodes where the agents deterministically followed the recommendations from their private signals, after which ϵ was set to 0.1 for episode 501 and decayed by a factor 0.999 in each subsequent episode. No attempt was made to conduct comprehensive parameter sweeps to optimize the values of α and ϵ which were used in either experiment.

[5] A complete implementation can be found here: https://github.com/rradules/equilibria_monfg.

[6] We note that specialized algorithms exist to learn mixed strategy Nash equilibria (e.g., Fudenberg and Kreps, 1993) or correlated equilibria (e.g., Arifovic *et al*., 2016) in single-objective MAS. We leave the design and empirical evaluation of versions of these algorithms for learning or approximating equilibria in MOMAS under SER for future work.

[7] This nonlinear optimization problem is solved using the 'optimize' module of the Scipy Python package (Virtanen *et al.*, 2019).

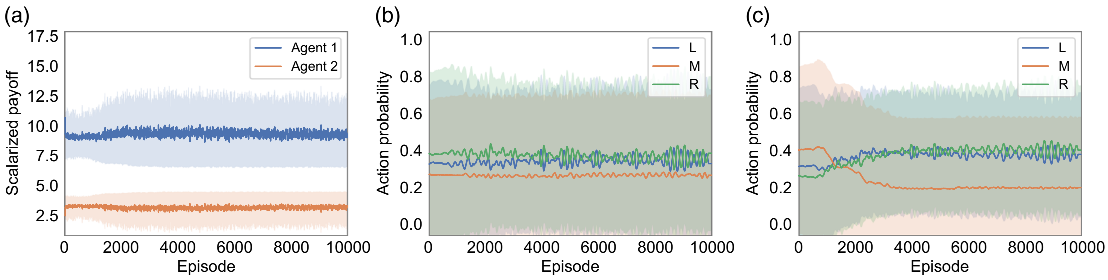

Figure 1. Game 1 under SER with no action recommendations. (a) Scalarized payoffs obtained by each agent. (b) Action selection probabilities of Agent 1. (c) Action selection probabilities of Agent 2

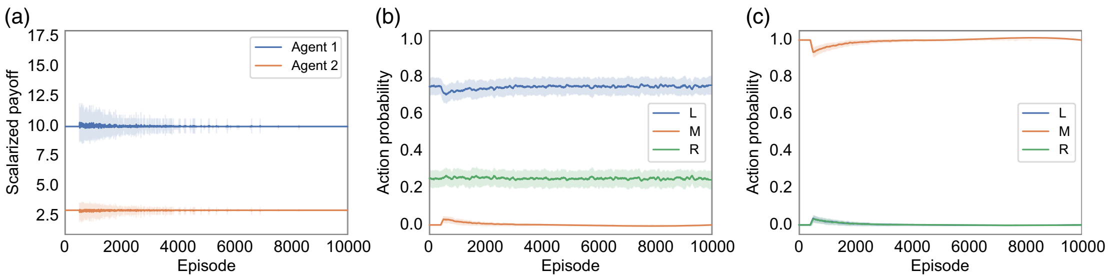

Figure 2. Game 1: single-signal CE under SER with action recommendations provided according to Table 5. (a) Scalarized payoffs obtained by each agent. (b) Action selection probabilities of Agent 1. (c) Action selection probabilities of Agent 2

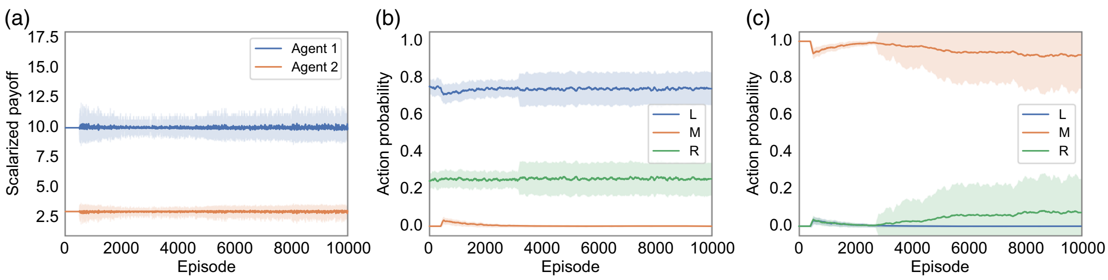

Figure 3. Game 1: multi-signal CE under SER with action recommendations provided according to Table 5. Scalarized payoffs obtained by each agent. (b) Action selection probabilities of Agent 1. (c) Action selection probabilities of Agent 2

5.1 Game 1: the (Im)balancing act game

For Game 1, the correlated signal was given in accordance with Table 5, that is, in a given episode, (L, M) was recommended with probability 0.75, or else (R, M) was recommended with probability 0.25.

The experimental results in terms of scalarized payoff are shown in Figures 1, 2, and 3. All figures show the scalarized payoffs received by the agents in each episode, averaged over 100 trials. Figure 4 presents the distribution of outcomes over the joint action space for the last 1000 interactions, averaged again over 100 trials. For each experiment, we also plot the action selection probabilities for each of the two players (Figures 1(b), (c), 2(b), (c), 3(b), and (c)). The probabilities are computed using a sliding window of size 100 over the history of taken actions and are also averaged over 100 trials. The shaded region around each plot shows one standard deviation from the mean. No smoothing was applied to any of the plots.

It is clear to see from the high standard deviations in Figure 1(a) that agents do not reliably converge on any one joint strategy when no correlated action recommendations are provided. This conclusion is further strengthened when observing the action selection probabilities of player 1 (Figure 1(b)) and player 2 (Figure 1(c)). Given our analysis in Theorem 3, this is to be expected, as agents will always have some

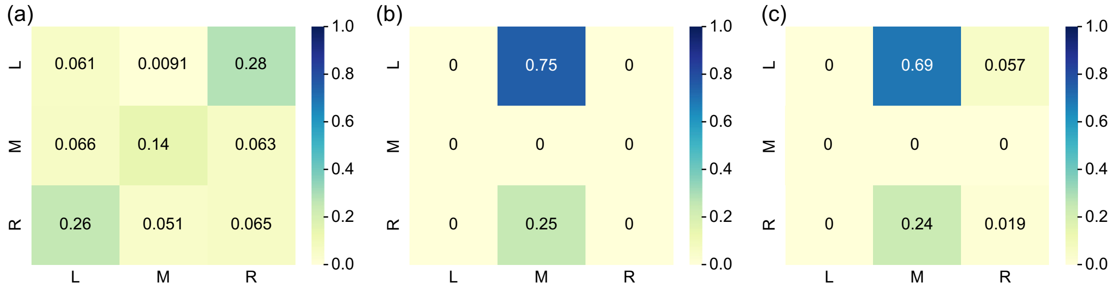

Figure 4. Game 1: joint action probabilities over the last 1000 episodes under SER. (a) No action recommendations. (b) Single-signal CE. (c) Multi-signal CE

incentive to deviate from a potential NE point in this game. As ϵ is decayed, the agents' behavior does not converge to any stable point, and the joint strategies learned in each run seem to always cycle among a few possibilities (e.g., predominant joint actions are (R, L), (L, R), and (M, M) as it can be seen from Figure 4).

In Figure 2, the effect of the single-signal CE may clearly be seen. As we would expect, for the first 500 episodes, a consistent scalarized payoff is received by both agents while they learn the CE. From episode 501, both agents are free to select actions autonomously and to explore and learn the effects of deviating from the action suggestions. As ϵ is gradually decayed toward zero, the agents consistently converge back to the CE, evidenced by the low standard deviations around the means of the scalarized payoffs near episode 10 000. Furthermore, Figures 2(b), (c), and 4(b) show that the action selection probabilities for each player nicely converge to the probabilities of the CE in Table 5 (i.e., agent 1 will select L with 25% probability and R with 75% probability, while agent 2 ends up selecting M with 100% of the time). This provides empirical support for our claim in Theorem 4 that single-signal correlated equilibria can exist in MONFGs under SER, demonstrating that neither agent has an incentive to deviate unilaterally given an action recommendation, when learning in this MONFG under SER.

For the case of multi-signal CE, Figure 3 clearly indicates how, after the initial 500 episodes, the agents slowly diverge from the given recommendations. From Figure 3(c), we can notice how agent 2 decays the use of the recommended action M, replacing it consistently with R, as it is trying to push the outcome toward the more imbalanced payoff outcome (L, R), given that his opponent is initially still taking the recommended actions L with 75% probability. We can then notice from Figure 3(b) an attempt from agent 1 to coordinate their actions to obtain (R, R), but with less success according to the join-action distribution outcome presented in Figure 4(c). This provides empirical support for our claim in Theorem 5 that multi-signal correlated equilibria need not exist in MONFGs under SER, demonstrating that the agents have incentives to deviate from the given action recommendations, when learning in this MONFG under SER.

5.2 Game 2: the (Im)balancing Act Game without action M

For Game 2, the correlated signal was given in accordance with Table 6 (right), that is, in a given episode, each possible joint action among the four (i.e., (L, L), (L, R), (R, L), or (R, R)) is recommended with equal probability.

The experimental results in terms of scalarized payoff are shown in Figures 5, 6, and 7, respectively. Figure 8 presents the distribution over the joint action space for the last 1000 interactions. Again, all experiments are run for 10 000 interactions, averaged over 100 trials.

Figure 5(b) and (c) highlight the dynamics between the agents of shifting between balanced and imbalanced outcomes, without being able to converge to any stable equilibrium strategies when no action recommendations are given, implying that there are no NE present in this game. According to Figure 8(a), player 2 seems to be more successful in obtaining her desired outcomes (L,R) or (R,L). Regarding the multi-signal CE, we see from Figure 7(b) that player 1 has a stronger incentive to deviate from the recommendations, probably obtained due to the higher loss in utility incurred when switching between the

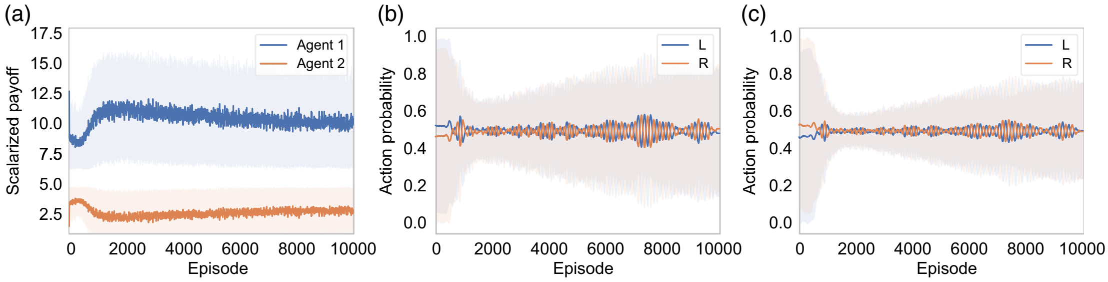

Figure 5. Game 2 under SER with no action recommendations. Scalarized payoffs obtained by each agent. (b) Action selection probabilities of Agent 1. (c) Action selection probabilities of Agent 2

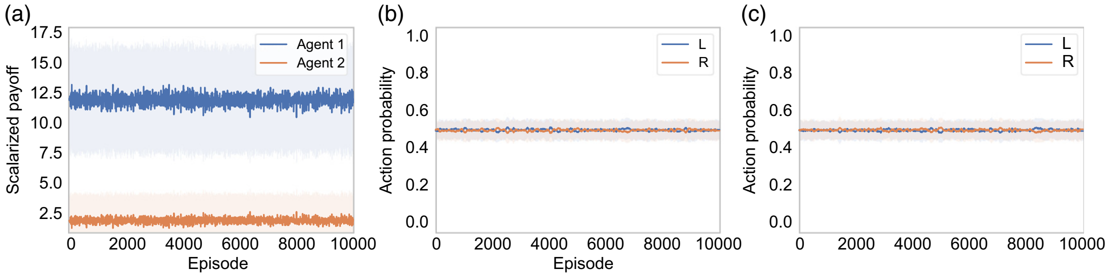

Figure 6. Game 2: single-signal CE under SER with action recommendations provided according to Table 5. (a) Scalarized payoffs obtained by each agent. (b) Action selection probabilities of Agent 1. (c) Action selection probabilities of Agent 2

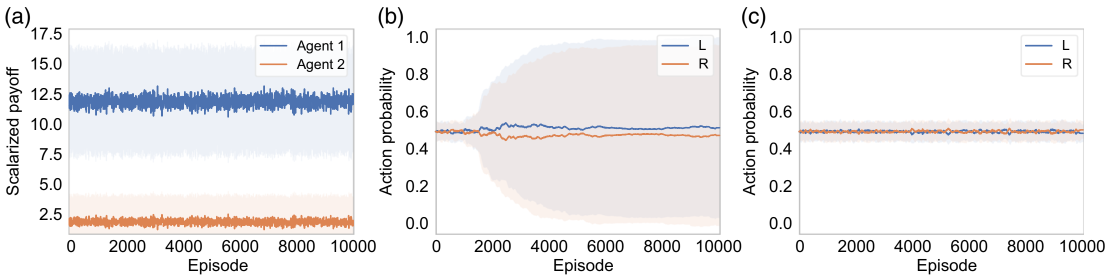

Figure 7. Game 2: multi-signal CE under SER with action recommendations provided according to Table 5. (a) Scalarized payoffs obtained by each agent. (b) Action selection probabilities of Agent 1. (c) Action selection probabilities of Agent 2

possible outcomes. In any case, similar to the previous experiment, agents are not able to converge to any stable strategy, implying that the set of action recommendations in Table 6 (right) do not constitute a multi-signal CE for this game. Finally, similar to the 3-action (Im)balancing act game, agents have no incentive to deviate from the action recommendations in Table 6 (right) in the case of a single-signal CE, as can be seen from Figure 6(b) and (c). Additionally, the distribution of outcomes over the joint action space (Figure 8(b)) also closely aligns with the action recommendations in Table 6 (right), thus allowing the agents to fairly coordinate between ending up half of the time in the imbalanced payoff outcomes, preferred by player 1, and the balanced payoff outcomes, preferred by player 2.

5.3 *Game 3: a 3-action MONFG with pure NE*

For Game 3, the correlated signal was given in accordance with Table 7 (right), that is, in a given episode, (L, L) was recommended with probability 0.5, or else (M, M) was recommended with the same probability.

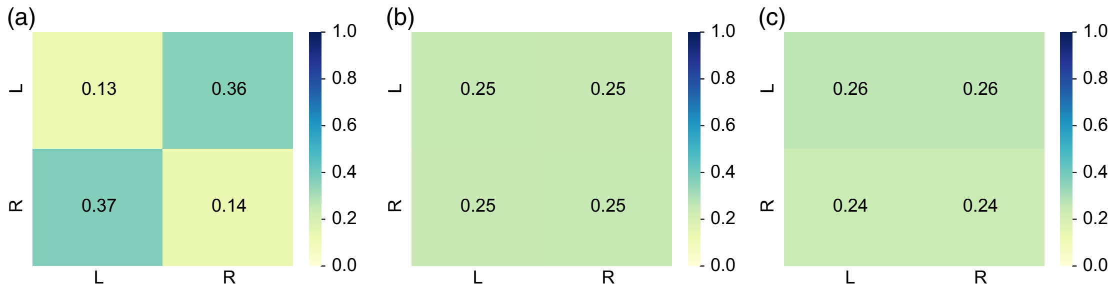

Figure 8. Game 2: joint action probabilities over the last 1000 episodes under SER. (a) No action recommendations. (b) Single-signal CE. (c) Multi-signal CE

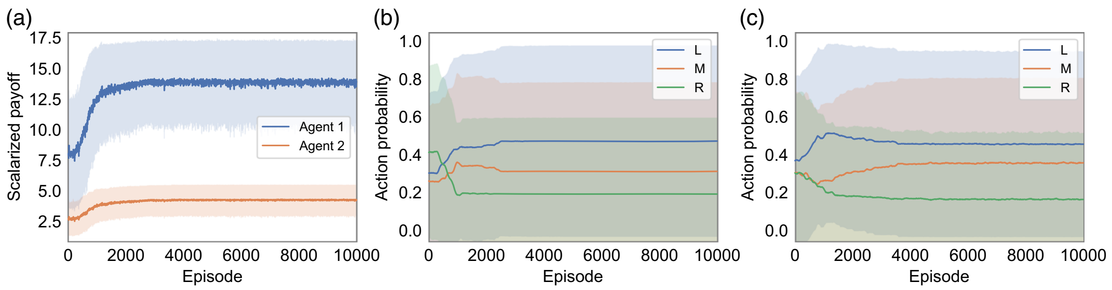

Figure 9. Game 3 under SER with no action recommendations. (a) Scalarized payoffs obtained by each agent. (b) Action selection probabilities of Agent 1. (c) Action selection probabilities of Agent 2

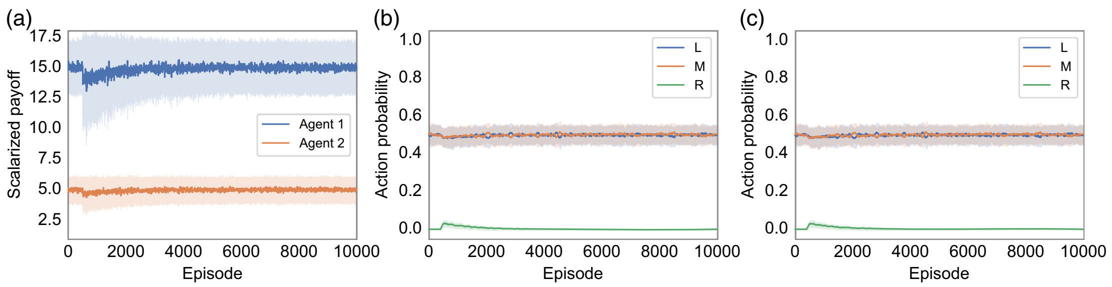

Figure 10. Game 3: single-signal CE under SER with action recommendations provided according to Table 5. (a) Scalarized payoffs obtained by each agent. (b) Action selection probabilities of Agent 1. (c) Action selection probabilities of Agent 2

Compared to the previous two games, we now have the opportunity to study the learning outcomes of the agents when all the considered equilibria exist. Figures 9 and 12(a) present the results for the setting in which the agents do not receive any action recommendations. Although the action selection probabilities (Figure 10(b) and (c)) might not exhibit any regular behavior over the considered trials, when looking at the distribution over the joint action space in Figure 12(a) more structure emerges. We notice that for about 95% of the time the agents converge to one of the pure Nash equilibria, described in Section 4.3.2—(L, L), (M, M), or (R, R)—with the Pareto-dominated outcome, (R, R), having the least probability mass. This indicates that our learning algorithm that combines a 'one-shot' vectorial Q-learning update rule with a ε-greedy action selection method, while allowing agents to determine their best mixed strategy by solving nonlinear optimization problems with respect to their Q-values and utility function, is quite successful in converging to NE in the case of independent learners.

When agents are able to receive action recommendations, we can notice that the selected correlated strategy is both a single-signal CE (Figures 10 and 12(b)) and a multi-signal CE (Figures 11 and 12(c)). By looking at the scalarized payoffs in Figures 10(a) and 11(a), we can also notice that even in a multi-objective setting, correlated equilibria can allow one to obtain better compromises between conflicting

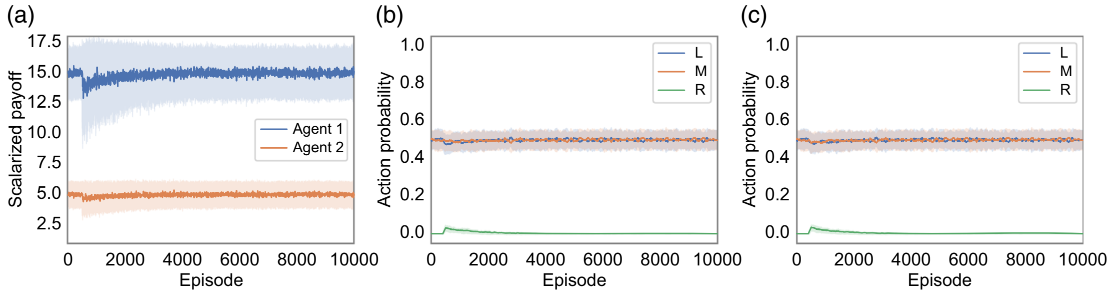

Figure 11. Game 3: multi-signal CE under SER with action recommendations provided according to Table 5. (a) Scalarized payoffs obtained by each agent. (b) Action selection probabilities of Agent 1. (c) Action selection probabilities of Agent 2

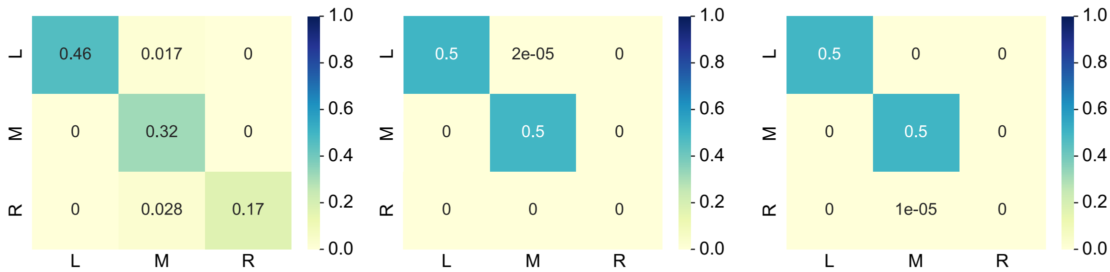

Figure 12. Game 3: joint action probabilities over the last 1000 episodes under SER. (a) No action recommendations. (b) Single-signal CE. (c) Multi-signal CE

utility functions (i.e., a SER of 14.99 for agent 1 and 5 for agent 2 in the case of single- and multi-signal CE) compared to the NE case (i.e., SER of 13.98 for agent 1 and 4.38 for agent 2), given that the agents are able to receive a correlation signal. This empirical result demonstrates that the well-known previous findings that CE can provide better payoffs than NE in single-objective games (see e.g., Aumann, 1974) can also apply in the more general class of multi-objective games, that is, that in a MOMAS where a coordination signal can be established CE can potentially lead to higher individual utilities than NE.

6 Conclusion and future work

In this work, we explored the differences between two optimization criteria for MOMAS: ESRs and SERs. Using the framework of MONFGs, we constructed sets of conditions for the existence of Nash equilibria and correlated equilibria, two of the most commonly used solution concepts in the single-objective MAS literature. Our analysis demonstrated that fundamental differences exist between the ESR and SER criteria in multi-agent settings.

While we have provided some theoretical results concerning the existence of equilibria in utility-based MONFGs, a number of deep and interesting open questions remain unanswered. Even though we provide examples of games where Nash equilibria and multi-signal correlated equilibria both exist (Table 7) or do not exist (proof of Theorems 3 and 5) under SER when considering nonlinear utility functions, we have no concrete conclusion on how or if the relation between NE and CE modifies under SER, that is, when we can expect equilibria to exist, and when they do not; therefore, further detailed theoretical analysis is required.

In the proof of Theorem 4, we provide an example where a single-signal CE does exist under SER, although it not known whether single-signal correlated equilibria always exist in this setting. The existence of correlated equilibria in single-objective NFGs has been proven by Hart and Schmeidler (1989) based on linear duality, an argument which does not rely on the existence of Nash equilibria (or by extension, the use of a fixed point theorem as per Nash, 1951) as the original proof by Aumann (1974) did. Extending the work of Hart and Schmeidler (1989) for utility-based MONFGs under SER is a promising direction for future work. As we saw in the example Chicken game in Table 1, correlated equilibria allow for better compromises to be achieved between conflicting payoff functions in single-objective NFGs, when compared with Nash equilibria. In utility-based MONFGs, we demonstrated that this property translates well, allowing compromises to be achieved between conflicting utility functions (and allowing a stable compromise solution to be reached in MONFGs where no stable compromise may be reached using Nash equilibria, when conditioning on the received signal).

The analysis in this paper has a number of important limitations which should be addressed in future work. Our worked examples considered MONFGs with two agents only, so the interaction between equilibria and optimization criteria should be further explored in larger MOMAS. It would also be worthwhile to conduct larger and more rigorous empirical studies to further expand upon our findings and to develop new learning algorithms for MOMAS. One promising direction to allow agents to avoid solving nonlinear optimization problems when computing their optimal mixed strategy is to adopt an actor-critic approach, as per a new learning algorithm recently proposed for MONFGs by Zhang *et al*., (2020). By adopting the MONFG model, we considered stateless decision-making problems only; our analysis should be extended to stateful MOMAS models such as multi-objective stochastic games (MOSGs) (Mannion *et al*., 2017b) or even multi-objective versions of partially observable stochastic games (Wiggers *et al*., 2016). We note that a similar equilibrium concept to the CE exists for single-objective stochastic games; the cyclic equilibrium (or cyclic CE) (Zinkevich *et al*., 2006). Little is currently known about equilibria in multi-objective multi-agent sequential decision-making settings. If the existence of Nash equilibria cannot be proven or demonstrated for MOSGs with nonlinear utility functions under SER in the future, the cyclic equilibrium is one alternate solution concept which is worthy of exploration.

Another interesting line of future research concerns the interaction between MOMAS, optimization criteria (ESR vs. SER), and reward shaping. Although reward shaping in MOMAS has received some attention to date (see e.g., Yliniemi and Tumer, 2016; Mannion *et al*., 2017a; Mannion *et al*., 2018), it has been primarily from the ESR perspective, and using linear and hypervolume scalarization functions only. Principled reward shaping techniques such as potential-based reward shaping and difference rewards come with convenient theoretical guarantees (e.g., preserving the relative value of policies and/or actions, and therefore Nash and Pareto relations between policies and/or actions in MAS/MOMAS (Devlin and Kudenko, 2011; Colby and Tumer, 2015; Mannion *et al*., 2017a, b); how well these techniques will work under SER with nonlinear utility functions is currently unknown.

How to best model users' utility functions for MOMAS remains a significant open question. Recent work on preference elicitation strategies for multi-objective decision support settings (Zintgraf *et al*., 2018) has delivered promising results in single-agent settings with nonlinear utility; this approach could feasibly be extended to generate utility functions for decision-making in MOMAS. It may also be beneficial for agents in MOMAS to learn opponent models; opponent modeling could potentially help to improve the utility of agents that implement it, as well as improving the probability of convergence to desirable equilibria. Initial work on opponent modeling for MONFGs under SER with nonlinear utility functions (Zhang *et al*., 2020) modeled opponent behavior via policy reconstruction using conditional action frequencies. Modeling opponent utility functions directly, for example, using Gaussian processes, is another promising avenue that could be explored in future research.

Finally, as we mentioned in Section 2.3, users may prefer either the SER or ESR criterion depending on their needs (e.g., whether they care more about average performance over a number of policy executions or just the performance of a policy single execution (Roijers *et al*., 2018)). In larger MOMAS, it is possible that not all users would choose the same optimization criterion, or that their preference for a specific optimization criterion may change over time, potentially adding further complexity to the process of computing equilibria.

Acknowledgement

We thank Bart Bogaerts for his useful feedback and discussions on this work. This research received funding from the Flemish Government under the 'Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaandere' program.

References

Arifovic, J., Boitnott, J. F. & Duffy, J. 2016. Learning correlated equilibria: an evolutionary approach. *Journal of Economic Behavior & Organization* **157**, 171–190.

Aumann, R. J. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* **1**(1), 67–96.

Aumann, R. J. 1987. Correlated equilibrium as an expression of bayesian rationality. *Econometrica: Journal of the Econometric Society* **1**, 1–18.

Bergstresser, K. and Yu, P. 1977. Domination structures and multicriteria problems in n-person games. *Theory and Decision* **8**(1), 5–48.

Blackwell, D. *et al.* 1956. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics* **6**(1), 1–8.

Borm, P., Tijs, S. & van den Aarssen, J. 1990. Pareto equilibria in multi-objective games. *Methods of Operations Research* **60**, 303–312.

Colby, M. & Tumer, K. 2015. An evolutionary game theoretic analysis of difference evaluation functions. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, 1391–1398. ACM.

Devlin, S. & Kudenko, D. 2011. Theoretical considerations of potential-based reward shaping for multi-agent systems. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 225–232.

Foster, D. P. & Vohra, R. 1999. Regret in the on-line decision problem. *Games and Economic Behavior* **29** (1–2), 7–35.

Fudenberg, D. & Kreps, D. M. 1993. Learning mixed equilibria. *Games and Economic Behavior* **5** (3), 320–367. ISSN 0899-8256.

Hart, S. & Schmeidler, D. 1989. Existence of correlated equilibria. *Mathematics of Operations Research* **14**(1), 18–25.

Igarashi, A. & Roijers, D. M. 2017. Multi-criteria coalition formation games. In *International Conference on Algorithmic DecisionTheory*, 197–213. Springer.

Jensen, J. L. W. V. *et al.* 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Actamathematica* **30**, 175–193.

Lozan, V. & Ungureanu, V. 2013. Computing the pareto-nash equilibrium set in finite multi-objective mixed-strategy games. *Computer Science Journal of Moldova*, **21** (2).

Lozovanu, D., Solomon, D. & Zelikovsky, A. 2005. Multiobjective games and determining pareto-nashequilibria. *Buletinul Academiei de Ştiinţe a Republicii Moldova. Matematica*, (3), 115–122.

Mannion, P., Devlin, S., Duggan, J. & Howley, E. 2018. Reward shaping for knowledge-based multi-objective multi-agent reinforcement learning. *The Knowledge Engineering Review* **33**, e23.

Mannion, P., Devlin, S., Mason, K., Duggan, J. & Howley, E. 2017a. Policy invariance under reward transformations for multi-objective reinforcement learning. *Neurocomputing* **263**, 60–73.

Mannion, P., Duggan, J. & Howley, E. 2016a. An experimental review of reinforcement learning algorithms for adaptive traffic signal control. In *Autonomic Road Transport Support Systems*, McCluskey, L. T., Kotsialos, A., Müller, P. J., Klügl, F., Rana, O. & Schumann, R. (eds), 47–66. Springer International Publishing.

Mannion, P., Duggan, J. & Howley, E. 2017b. A theoretical and empirical analysis of reward transformations in multi-objective stochastic games. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2017b.

Mannion, P., Mason, K., Devlin, S., Duggan, J. & Howley, E. 2016b. Multi-objective dynamic dispatch optimisation using multi-agent reinforcement learning. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2016b.

Mossalam, H., Assael, Y. M., Roijers, D. M. & Whiteson, S. 2016. Multi-objective deep reinforcement learning. In *NIPS Workshop on Deep Reinforcement Learning*.

Nash, J. 1950. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences* **36**(1), 48–49. ISSN 0027-8424.

Nash, J. 1951. Non-cooperative games. *Annals of Mathematics* **54**(2), 286–295.

Papadimitriou, C. H. & Roughgarden, T. 2008. Computing correlated equilibria in multi-player games. *Journal of the ACM (JACM)* **55**(3), 14.

Rădulescu, R., Legrand, M., Efthymiadis, K., Roijers, D. M. & Nowé, A. 2018. Deep multi-agent reinforcement learning in a homogeneous open population. In *Proceedings of the 30th Benelux Conference on Artificial Intelligence (BNAIC 2018)*, 177–191.

Rădulescu, R., Mannion, P., Roijers, D. & Nowé, A. 2019. Equilibria in multi-objective games: a utility-based perspective. In *Adaptive and Learning Agents Workshop (at AAMAS 2019)*, May 2019.

Rădulescu, R., Mannion, P., Roijers, D. M. and Nowé, A. 2020. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems* **34** (10).

Reymond, M., Patyn, C., Rădulescu, R., Deconinck, G. & Nowé, A. 2018. Reinforcement learning for demand response of domestic household appliances. In *Proceedings of the Adaptive and Learning Agents Workshop at FAIM 2018*.

Roijers, D. M. 2016. *Multi-Objective Decision-Theoretic Planning*. PhD thesis, University of Amsterdam.

Roijers, D. M., Steckelmacher, D. & Nowé, A. 2018. Multi-objective reinforcement learning for the expected utility of the return. In *Proceedings of the Adaptive and Learning Agents Workshop at FAIM 2018*.

Roijers, D. M., Vamplew, P., Whiteson, S. & Dazeley, R. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research* **48**, 67–113.

Roijers, D. M. & Whiteson, S. 2017. Multi-objective decision making. *Synthesis Lectures on Artificial Intelligence and Machine Learning* **11**(1), 1–129.

Shapley, L. S. & Rigby, F. D. 1959. Equilibrium points in games with vector payoffs. *Naval Research Logistics Quarterly* **6** (1), 57–61.

Talpert, V., Sobh, I., Kiran, B. R., Mannion, P., Yogamani, S., El-Sallab, A. & Perez, P. 2019. Exploring applications of deep reinforcement learning for real-world autonomous driving systems. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, February 2019.

Vamplew, P., Dazeley, R., Berry, A., Issabekov, R. & Dekker, E. 2011. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine Learning* **84** (1–2), 51–80.

Van Moffaert, K. & Nowé, A. 2014. Multi-objective reinforcement learning using sets of pareto dominating policies. *The Journal of Machine Learning Research* **15**(1), 3483–3512.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & Contributors, S. 2019. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. arXiv e-prints, art. arXiv:1907.10121, July 2019.

Voorneveld, M., Vermeulen, D. & Borm, P. 1999. Axiomatizations of paretoequilibria in multicriteria games. *Games and Economic Behavior* **280** (1), 146–154.

Walraven, E. & Spaan, M. T. J. 2016. Planning under uncertainty for aggregated electric vehicle charging with renewable energy supply. In *Proceedings of the European Conference on Artificial Intelligence*, 904–912.

Watkins, C. J. C. H. Learning from Delayed Rewards. PhD thesis, King's College, Cambridge, UK, 1989.

Wierzbicki, A. P. 1995. Multiple criteria games – theory and applications. *Journal of Systems Engineering and Electronics* **60** (2), 65–81.

Wiggers, A. J., Oliehoek, F. A. & Roijers, D. M. 2016. Structure in the value function of two-player zero-sum games of incomplete information. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, 1628–1629. IOS Press.

Yliniemi, L., Agogino, A. K. & Tumer, K. 2015. Simulation of the introduction of new technologies in air traffic management. *Connection Science* **270** (3), 269–287.

Yliniemi, L. & Tumer, K. 2016. Multi-objective multiagent credit assignment in reinforcement learning and nsga-ii. *Soft Computing* **200** (10), 3869–3887.

Zhang, Y., Rădulescu, R., Mannion, P., Roijers, D. M. & Nowé, A. 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 2020)*, May 2020.

Zinkevich, M., Greenwald, A. & Littman, M. L. 2006. Cyclic equilibria in markov games. *In Advances in Neural Information Processing Systems*, 1641–1648.

Zintgraf, L. M., Kanters, T. V., Roijers, D. M., Oliehoek, F. A. & Beau, P. 2015. Quality assessment of MORL algorithms: a utility-based approach. In *Benelearn 2015*: *Proceedings of the Twenty-Fourth Belgian-Dutch Conference on Machine Learning*.

Zintgraf, L. M., Roijers, D. M., Linders, S., Jonker, C. M. & Nowé, A. 2018. Ordered preference elicitation strategies for supporting multi-objective decision making. In *Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems*, 1477–1485. International Foundation for Autonomous Agents and Multiagent Systems.