Generalized Response Objectives for Strategy Exploration in Empirical Game-Theoretic Analysis

Yongzhao Wang
The Alan Turing Institute
London, United Kingdom
yongzhao.wang@turing.ac.uk

Michael P. Wellman
University of Michigan
Ann Arbor, USA
wellman@umich.edu

ABSTRACT

In the policy-space response oracle (PSRO) framework, strategy sets defining an empirical game are iteratively extended by computing each player's best response to a target profile. The method for selecting a target profile is called a *meta-strategy solver* (MSS), and a variety of MSSs have been proposed and analyzed for their effectiveness in exploring the strategy space. Here we investigate an alternative means to control strategy exploration: setting the *response objective* (RO) employed in deriving a strategy for a given target profile. In evaluating effectiveness of strategy exploration, we consider not only rate of convergence to a solution, but also the quality of solution(s) captured by the evolving empirical game. We perform our studyfi rst in the domain of sequential bargaining games, comparing the standard RO based on own payoffwith others that incorporate other players' payoffs. Wefi nd that other-regarding ROs can lead tofi nding equilibrium outcomes with significantly higher social welfare than the standard objective. For other proposed ROs, experiments demonstrate that they can differentially affect the makeup and value of solutions for different players. We further test PSRO with generalized ROs in large attack-graph games. We observe a similar impact and effectiveness of our ROs on strategy exploration. Finally, we establish a theoretical relationship between PSRO with generalized ROs and generalized weakened fictitious play in particular settings, and a connection between the social welfare related RO and Berge equilibrium.

KEYWORDS

Empirical Game-Theoretic Analysis; Strategy Exploration; Policy Space Response Oracles; Response Objectives

ACM Reference Format:
Yongzhao Wang and Michael P. Wellman. 2024. Generalized Response Objectives for Strategy Exploration in Empirical Game-Theoretic Analysis. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 13 pages.

1 INTRODUCTION

The methodology of ***Empirical Game-Theoretic Analysis*** (EGTA) [29, 34] offers a comprehensive collection of techniques for game reasoning with models based on simulation data. For multi-agent systems not amenable to analytic solution, EGTA provides a simulation-based alternative, where a selected set of strategies are evaluated, addressing the most important strategic considerations [2]. The challenge of efficiently assembling a suitable collection of strategies for EGTA is called the ***strategy exploration*** problem [12]. The clearest formulation of strategy exploration in EGTA is within an iterative process, in which the creation of new strategies alternates with the assessment and analysis of a game model. In particular, the ***Policy Space Response Oracle*** (PSRO) algorithm [13] provides afl exible framework for strategy exploration, with new strategies generated each iteration through a best response to a target other-players profile using reinforcement learning (RL). The component responsible for determining the target profile is called a meta-strategy solver (MSS), which takes an empirical game model as input and "solves" it to produce the target. PSRO can be viewed as generalizing some game-learning algorithms. For example, PSRO with Nash equilibrium (NE) as MSS is essentially the ***double oracle*** (DO) algorithm [19] with RL for computing (approximate) best responses. DO incrementally adds strategies that are best-responses to NE of the current game model, terminating once an NE is found. The NE found by DO and its features are almost uniquely determined given afi xed initialization and a tie-breaking rule for best responses.

Identifying a single NE is sometimes sufficient for the goals of game analysis, particularly in situations like two-player zero-sum games, where NE are interchangeable. In other cases, we might be interested in characterizing multiple equilibria, or identifying solutions with particular features. For example, in the Traveler's dilemma (Table 1) [3, 7], we might prefer the profile $(5, 5)$ to the NE $(1, 1)$ since the profile $(5, 5)$ has a reasonable low regret but a significantly higher social welfare than the NE. Unfortunately, DO and many classic learning dynamics (e.g., no-regret learning) will always converge to the NE $(1, 1)$ in this case.

	5	4	3	2	1
5	(6,6)	(3,7)	(2,6)	(1, 5)	(0,4)
4	(7,3)	(5,5)	(2,6)	(1, 5)	(0,4)
3	(6,2)	(6,2)	(4,4)	(1, 5)	(0,4)
2	(5,1)	(5,1)	(5,1)	(3,3)	(0,4)
1	(4,0)	(4,0)	(4,0)	(4,0)	(2,2)

Table 1 The Traveler's dilemma.

Based on this observation, we raise the question of how to steer strategy exploration toward NE with preferred characteristics, or more generally, a preferred game model. One natural hypothesis is that the choice of response objectives (ROs), which are objectives

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

(approximately) solved through RL at each iteration of PSRO, can substantially impact strategy exploration and equilibrium outcomes. For example, in the Traveler's dilemma, we observe that optimizing another player's payoff, as opposed to maximizing its own payoff, will yield strategies that make up high-welfare profiles. Stemming from this hypothesis, we introduce PSRO with generalized ROs. Generalized ROs are not limited to optimizing utility against other players' strategies, as in the standard PSRO framework, but allows ROs to take any forms compatible with RL and incorporate specified preferences. We propose four RO instances for PSRO with various strategy exploration preferences and evaluate them in sequential bargaining games and attack-graph games, comparing solutions found according to various criteria and revealing the impact of different combinations of MSSs and ROs on strategy exploration.

Sequential bargaining games involve two players engaging in a process to negotiate a deal over items. These games commonly exhibit multiple equilibria of varying preference, thus making them an especially interesting environment for investigating our questions. A natural preference in sequential bargaining games for strategy exploration is tofi nd equilibria with higher social welfare. We explore two methods to encode this preference into ROs by adding to the standard RO a term based on the Nash bargaining product and a term that trades off deviation payoff for other players' utilities, respectively. Our research shows that both methods can yield equilibria with significantly higher social welfare than other PSRO variants, regardless of the MSSs employed. In addition, we demonstrate that our other ROs can achieve their intended purposes and effectively reduce or enlarge the utility difference among players in equilibria. We further investigate the impact of generalized ROs in attack-graph games, where we obtain similar observations on the impact and efficacy of our ROs for strategy exploration, as in bargaining games.

Besides learning toward NE with particular features, PSRO with generalized ROs can also steer strategy exploration to solution concepts other than NE. As an example, we specify an MSS-RO combination for PSRO, enabling the computation of Berge equilibrium (BE) [5], a solution concept widely used in the study of social science. Compared to prior methods that compute BE by enumerating all profiles [8], our method is scalable and enables BE computation in large games. Finally, we connect PSRO with generalized ROs with generalized weakenedfi ctitious play (GWFP) [14] and show that PSRO with certain ROs share the convergence properties of GWFP.

Contributions of this study include:

(1) PSRO with generalized ROs: a generalization of PSRO framework for controlling strategy exploration through customized ROs;
(2) Introduction of four variant RO forms and a comprehensive analysis of PSRO with these ROs in sequential bargaining games and attack-graph games; Our key observation is that the choice of ROs can substantially improve the quality of strategy exploration and equilibrium outcomes;
(3) A demonstration that PSRO with generalized ROs can be employed to compute Berge equilibrium;
(4) A theoretical connection between PSRO with generalized ROs and GWFP.

2 RELATED WORK

Some classic learning dynamics and a few PSRO variants involve modifications in the best response operation for varies purposes (e.g., regularization). Compared to prior literature, our work systematically investigates the effectiveness of ROs combined with disparate MSSs for guiding strategy exploration toward preferred game games.

2.1 Variant Objectives in Classic Learning Dynamics

The best response operation is a well-established technique used in classic game-learning dynamics, such asfi ctitious play (FP) [6], weakened FP [30], GWFP [14], and iterated best response. Some of these dynamics involve modifications in the best response targets (i.e., ROs). For example, the smooth FP method [10] perturbs best responses by a smooth and positive definite function (e.g., the Gibbs Entropy). This perturbation is not intended to steer strategy exploration toward a particular equilibrium but aims to achieve convergence through a concave regularizer.

2.2 Variant Objectives in PSRO

Following thefi rst PSRO paper [13], PSRO have been significantly advanced in recent years, with various newly developed MSSs. For example, Wang et al. [32] employed a mixture of NE and uniform, which essentially randomizes over whether to apply DO or fictitious play (FP) on a given PSRO iteration. Wright et al. [35] proposed history-aware PSRO thatfi ne-tunes the best response against out-of-equilibrium but recently seen opponents' strategies. Marris et al. [17] proposed maximum welfare coarse correlated equilibrium (MWCCE) and maximum Gini coarse correlated equilibrium (MGCCE) as MSSs for computing correlated equilibria. Wang and Wellman [33] proposed an MSS called regularized replicator dynamics (RRD), which mitigates the overfitting problem [13] by regularizing the equilibrium search process based on a regret criterion. McAleer et al. [18] proposed to use minimum regret constrained profile [12, 31] as an MSS for PSRO for two-player zero-sum games.

Most of these works follow the standard PSRO framework, where the learning player optimizes its own payoff against other players' strategies (i.e., the original RO), though a few have considered some variants of the standard RO. One such example is the method called diverse PSRO [27], which includes a diversity measure defined through a determinantal point process in the response objective. Liu et al. [16] proposed the unified diversity measure (UDM), as a way to capture a variety of diversity metrics including effective diversity [2], expected cardinality [27], and population diversity [26]. Muller et al. [22] employed an MSS based on α-Rank [25] and proposed a preference-based objective to ensure the convergence of PSRO to α-Rank. Li et al. [15] deployed Monte Carlo tree search (MCTS) as the best response oracle using different values (e.g., social welfare) to update values of nodes along the sample path in the back-propagation step of MCTS. The employment of different back-propagation values can also be viewed as modifications in ROs.

3 PRELIMINARIES

A normal-form game $\mathcal{G} = (N, (S_i), (u_i))$ comprises afi nite set of players N indexed by i, a non-empty set of strategies S_i for player $i \in N$, and a utility function $u_i : \prod_{j \in N} S_j \to \mathbb{R}$ for player $i \in N$.

A mixed strategy σ_i is a probability distribution over strategies in S_i, with $\sigma_i(s_i)$ denoting the probability player i plays strategy s_i. We adopt conventional notation for the other-agent profile: $\sigma_{-i} = \prod_{j \neq i} \sigma_j$. Let $\Delta(\cdot)$ represent the probability simplex over a set. The mixed strategy space for player i is given by $\Delta(S_i)$. Similarly, $\Delta(S) = \prod_{i \in N} \Delta(S_i)$ is the mixed profile space.

Player i's ***best response*** to profile σ is the set of strategies yielding maximum payofffor i,fi xing other-player strategies:

$$br_i(\sigma_{-i}) \equiv \underset{\sigma_i' \in \Delta(S_i)}{\text{argmax}}\, u_i(\sigma_i', \sigma_{-i}).$$

Let $br(\sigma) \equiv \prod_{i \in N} br_i(\sigma_{-i})$ be the overall best-response correspondence for a profile σ. A ***Nash equilibrium*** (NE) is a profile σ^* such that $\sigma^* \in br(\sigma^*)$.

Player i's ***regret*** for profile σ in game $\mathcal{G}$ is given by

$$\rho_i^{\mathcal{G}}(\sigma) \equiv \max_{s_i' \in S_i} u_i(s_i', \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i}).$$

Regret captures the maximum player i can gain in expectation by unilaterally deviating from its mixed strategy in σ to an alternative strategy in S_i. An NE has zero regret for every player. A profile is said to be an ϵ-***Nash equilibrium*** (ϵ-NE) if no player can gain more than ϵ by unilateral deviation. We define the regret of a strategy profile σ as the sum over player regrets:

$$\rho^{\mathcal{G}}(\sigma) \equiv \sum_{i \in N} \rho_i^{\mathcal{G}}(\sigma).$$

A ***restricted game*** $\mathcal{G}_{S\downarrow X}$ is a projection of full game $\mathcal{G}$, in which players choose from restricted strategy sets $X_i \subseteq S_i$. An ***empirical game*** $\hat{\mathcal{G}}$ is a model of true game $\mathcal{G}$ where payoffs are estimated through simulation. Thus, $\hat{\mathcal{G}}_{S\downarrow X} = (N, (X_i), (\hat{u}_i))$ denotes an empirical game model where $\hat{u}$ is an estimated projection of u onto the strategy space X.

PSRO is presented below as Algorithm 1 (with line 7). In PSRO, each player is initialized with a set of strategies X_i and the utilities for profiles in the profile space X are simulated, resulting in an initial empirical game $\hat{\mathcal{G}}_{S\downarrow X}$. At each iteration of PSRO, a ***meta-strategy solver*** extracts a profile from the empirical game $\hat{\mathcal{G}}_{S\downarrow X}$ as the best response target profile σ. Then each player (i.e., the learning player) computes a best response s_i' against other players' strategies σ_{-i} in the profile σ. This is achieved byfi xing other players' strategies σ_{-i}, enabling the environment to become stationary to the learning player for computing an optimal strategy. The best response s_i' will be added to its strategy set X_i in the empirical game. This procedure will repeat until certain stopping criterion has been satisfied (e.g., a fixed number of iterations).

A ***response objective*** for player i in PSRO is a function of strategy profiles, denoted as $RO_i(\sigma)$. For example, in standard PSRO described above, the RO can be written as $RO_i(\sigma) = u_i(s_i', \sigma_{-i})$ and maximizing it over s_i' gives player i a best response against σ_{-i}. Since MSSs and ROs may have a coupled impact on strategy exploration and are not interchangeable, for simplicity, we refer to the choice of a pair of an MSS and an RO as an MSS-RO combination.

Algorithm 1 PSRO, parametrized by solver MSS

Input: Initial strategy sets X
1: Estimate $\hat{\mathcal{G}}_{S\downarrow X}$ by simulating $\sigma \in X$
2: Initialize target $\sigma \leftarrow \text{MSS}(\hat{\mathcal{G}}_{S\downarrow X})$
3: **for** PSRO iteration $\tau = 1, 2, \ldots, \mathcal{T}$ **do**
4: **for** player $i \in N$ **do**
5: **for** many RL training episodes **do**
6: Sample a profile $s_{-i} \sim \sigma_{-i}$
7: **Standard PSRO:** Train best response oracle s_i' against s_{-i}
8: **PSRO with Generalized ROs:** Train a RL agent s_i' against s_{-i} to optimize $RO_i(s_i', s_{-i})$
9: **end for**
10: $X_i \leftarrow X_i \cup \{s_i'\}$
11: **end for**
12: Update $\hat{\mathcal{G}}_{S\downarrow X}$ by simulating missing profiles over X
13: Compute response target $\sigma \leftarrow \text{MSS}(\hat{\mathcal{G}}_{S\downarrow X})$
14: **end for**
15: **Return** $\hat{\mathcal{G}}_{S\downarrow X}$

4 PSRO WITH GENERALIZED RESPONSE OBJECTIVES

We introduce PSRO with generalized ROs in Algorithm 1 (with line 8), which generalizes the standard PSRO by allowing ROs to be customized for each player. The customized ROs will be (approximately) solved through RL at each iteration of PSRO and details for this optimization is described in Appendix A.2. Our key observation for generalized ROs is that ROs will substantially steer strategy exploration toward preferred equilibria, or more broadly, empirical game models. We demonstrate this byfi rst proposing four RO instances for PSRO with various strategy exploration preferences and then accessing their impacts in sequential bargaining games and attack-graph games.

In Table 2, we describefi ve ROs considered in this work.[1] In each, the learning player i maximizes the RO over $s_i' \in S_i$, responding to thefi xed other-player strategy σ_{-i}. First is the ***Original RO***—standard in PSRO—which maximizes i's own utility against σ_{-i} (i.e., the *deviation payoff*). Ourfi rst variant RO is named the ***Nash Product Response Objective*** (NPRO), which trades offthe deviation payoff for the Nash product (i.e., the product of players' utilities). It is well-known that maximizing the Nash product will yield the Nash bargaining solution [23]. By replacing the Nash product with other players' utilities, we obtain our second variant RO, called ***Social Welfare Response Objective*** (SWRO). When $\alpha = 0.5$, SWRO reproduces social welfare (i.e., the sum of players' utilities). Our next RO, the ***Social Equity Response Objective*** (SERO), aims to balance utilities among players. SERO penalizes the deviation payoff by the difference in utilities among players. Thefi nal RO, ***Minimizing Opponent Response Objective*** (MORO), seeks to explicitly minimize other-player utility, while also maximizing deviation payoff.

[1] The ROs in Table 2 are defined for two-player games, so $u_{-i}(s_i', \sigma_{-i})$ is a scalar. Some ROs like SWRO can be generalized straightforward for $|N|$ players.

RO Name	Formula
Original Response Objective	$u_i(s_i', \sigma_{-i})$
Nash Product Response Objective	$\alpha u_i(s_i', \sigma_{-i}) + (1-\alpha)u_i(s_i', \sigma_{-i})u_{-i}(s_i', \sigma_{-i})$
Social Welfare Response Objective	$\alpha u_i(s_i', \sigma_{-i}) + (1-\alpha)u_{-i}(s_i', \sigma_{-i})$
Social Equity Response Objective	$\alpha u_i(s_i', \sigma_{-i}) - (1-\alpha)\lvert u_i(s_i', \sigma_{-i}) - u_{-i}(s_i', \sigma_{-i})\rvert$
Minimizing Opponent Response Objective	$\alpha u_i(s_i', \sigma_{-i}) - (1-\alpha)u_{-i}(s_i', \sigma_{-i})$

Table 2 Five response objective forms. $\alpha \in [0, 1]$ is a weighting parameter.

5 SEQUENTIAL BARGAINING GAMES

Sequential bargaining games represent a broad class of situations where two parties attempt to reach a deal through a series of proposals and counter-proposals [11, 28]. Variations of this model have been applied extensively, to scenarios including negotiations between nations in trade agreements, and private individuals bargaining over salaries. Sequential bargaining is a salient domain for EGTA due to its strategic complexity, and ubiquity in practice. These games also commonly exhibit multiple equilibria of varying preference, thus making them an especially interesting environment for studying how strategy exploration can affect which equilibria are captured by alternative paths of empirical game models.

5.1 Game Setup

We consider a non-zero-sum incomplete-information bargaining game, in which two players alternatively make offers to reach a deal over K types of items within time horizon T. The item of type k has M_k units available. For each bargaining instance, M_k is drawn from a uniform distribution, and revealed to both players. Each player has a private per-unit valuation for each item type, drawn independently from a specified distribution. Also for each instance the players are assigned independently drawn *disagreement values*, from player-specific distributions.

During each time step $t \leq T$, one player makes an offer and the other player decides whether to accept or reject it. Offers are made in vector form, representing the quantities of each item requested by the player (e.g., (3, 1, 1) requests 3 units of thefi rst item and 1 unit each of the second and third items). If a deal is reached, the players receive a sum of their private values for the items in the offer, discounted by a factor of γ^t. If no deal is reached, they receive their disagreement values.

5.2 Experimental Results

In sequential bargaining games, a natural consideration for the bargaining outcome is social welfare. Under the PSRO framework, it means that an empirical game model that incorporates equilibria with greater social welfare will be more desirable than others. In ourfi rst experiment, we show that PSRO with NPRO and SWRO can yield such game models compared to other PSRO variants. Our experimental results support the claim that ROs can substantially impact strategy exploration and equilibrium outcomes.

Specifically, we run PSRO with a combination offi ve MSSs (RRD, Nash equilibrium, uniform, MWCCE, and MGCCE) and three ROs (the original RO, the NPRO, and the SWRO), producingfi fteen MSS-RO combinations in total (seven combinations are explained in Fig. 1 while the remaining eight are depicted in Figs. 2 and 3). PSRO with each MSS-RO combination will generate a sequence of empirical games. We evaluate the quality of solutions in the terminal empirical game for each combination. Our comparisons follow a consistency criterion [31], which states that whereas empirical games can be generated by different MSS-RO combinations, they should be evaluated based on measures of interest (e.g., regret, social welfare) applied to the same solution concept. For our purposes, we choose to compare the social welfare of the same solution concepts across the generated empirical games. We selectfi ve solution concepts for evaluation (shown in Table 3), reflecting the quality of solutions in the empirical games from different angles. For example, NE, MWCCE, and MGCCE represent the common solution concepts whilst the uniform reflects the average performance of strategies in the empirical game.

In Fig. 1, wefi rst show the impact of NPRO and SWRO on strategy exploration by replacing the original RO in Nash with them, respectively. Specifically, each color represents an MSS-RO combination (if the RO is the original RO, it is omitted for simplicity), and seven empirical games were generated in total, one for each combination. Then the social welfare (averaged over 15 random seeds) of the same conceptacross the seven empirical games were bundled. For example, in the Max SW group (i.e., the left-most bundle), we plot the maximal social welfare in pure strategy profiles for each of the seven empirical games. Moreover, in the NE SW group, the social welfare of NE of each empirical game is listed for comparison.

From Fig. 1, we observed that the Nash-NPRO combination generates the greatest social welfare across allfi ve solution concepts than other variants and Nash-SWRO is ranked second. By comparing Nash with Nash-NPRO and Nash-SWRO, we noticed a significant increase in social welfare for equilibria in the resulting empirical games after replacing the original RO with NPRO or SWRO. This observation verifies our concern for DO (i.e., PSRO with Nash) that it can stop at an NE with arbitrary features, and shows that either NPRO or SWRO can steer strategy exploration toward solutions with higher social welfare. As discussed later, our observation remains valid, regardless of the MSSs employed. It is worth mentioning that Nash-SWRO achieves the highest social welfare with a weighting parameter $\alpha = 0.8$, as opposed to the setting $\alpha = 0.5$ that exactly captures social welfare. In other words, there is a benefit to considering the other-agent value in constructing a response strategy, but not to the same degree as one's own value.

Solution Concept	Description
Max SW	maximum social welfare across pure strategy profiles
NE SW	social welfare of Nash equilibrium
Uniform SW	social welfare of a uniform distribution over strategies
MWCCE SW	social welfare of maximum social welfare coarse correlated equilibrium
MGCCE SW	social welfare of maximum Gini coarse correlated equilibrium

Table 3 Five solution concepts used for evaluation.

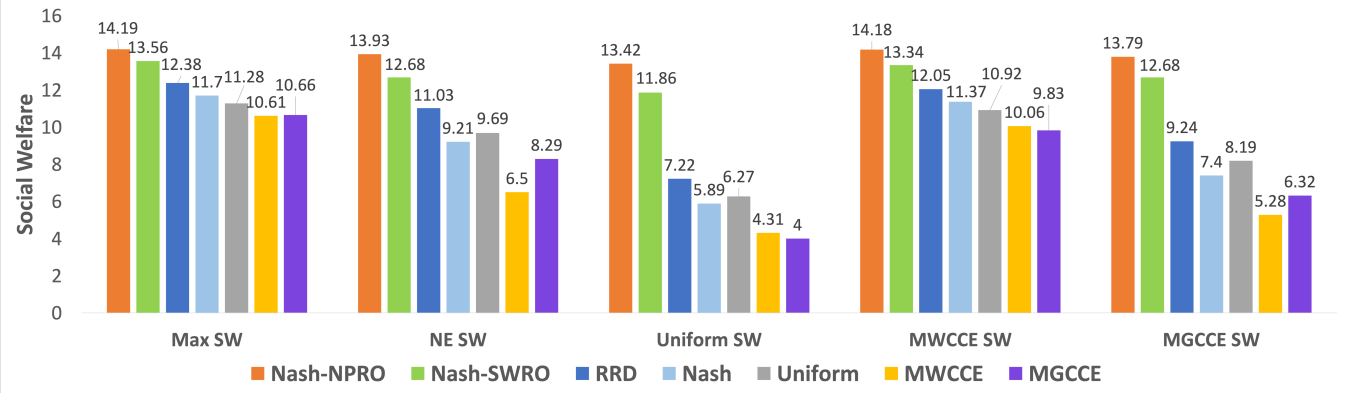


Figure 1 Social Welfare of PSRO with various MSS-RO combinations. Each color represents an MSS and each bundle of colors shows the SW of a given solution concept in the corresponding empirical games. Max SW is the maximum SW among pure strategy profiles.

Since there might exist multiple equilibria in an empirical game, which equilibrium to select for evaluation is pivotal. We demonstrate that this issue is less stressful given the results in our particular situation. In particular, we assume the solution concept of interest is NE and use Nash-SWRO as an example. From Figure 1, we can see that the social welfare of NE found by Nash-SWRO (i.e., 12.68) is higher than that of any other combinations in Max SW. Since the social welfare of any mixed strategy profile is upper bounded by the maximal social welfare over pure strategy profiles, the social welfare of NE found by Nash-SWRO is determined to be higher than the social welfare of any profiles (including NE) found by other MSS-RO combinations. Another way to reason about this argument is that since the set of NE is a subset of CCE, the social welfare of NE is upper bounded by the social welfare of MWCCE in the corresponding empirical game, which is further bounded by Max SW. As the social welfare of NE found by Nash-SWRO is higher than that of MWCCE given by other MSS-RO combinations, Nash-SWRO must result in NE with higher social welfare than others. The same argument can be simply applied to Nash-NPRO.

In Fig. 2, we combine NPRO with each MSS and plot the social welfare of the samefi ve evaluation concepts. We observed that the social welfare of solutions given by Nash-NPRO remains highest across all combinations. One interesting observation is that Nash-NPRO outperforms RRD-NPRO even though Nash performs worse than RRD (light blue vs dark blue in Fig. 1). This observation indicates that MSSs and ROs have a coupled influence on strategy exploration. Our hypothesis for the reduced performance of RRD with NPRO is that the regularization imposed by RRD is superfluous given that NPRO already varies from Nash.

In Fig. 3, we plot the social welfare of solutions before and after integrating NPRO and SWRO with each individual MSS. We observed that the social welfare of all evaluation concepts increases after applying either NPRO or SWRO, regardless of the MSS employed. This shows that either NPRO or SWRO alone can direct strategy exploration and identify strategy spaces that cover solutions with higher social welfare, though a proper choice of MSSs will further raise the social welfare (i.e., Nash-NPRO yields the highest SW). In Appendix A.1, we provide more data that shows the social welfare given by Nash-NPRO and Nash-SWRO is strikingly high compared to global maximum of the full game.

Fig. 4 plots individual player utilities in thefi nal NEs produced by 11 PSRO runs with different MSS-RO combinations. The convex hull of these points represent the empirical Pareto frontier of equilibria of the bargaining game. Points with the same color are obtained by running PSRO with different random seeds. From the plot, we observed that the equilibria given by both Nash-NPRO and Nash-SWRO are on the frontier and appear dense while the equilibria found by other combinations spread out in the utility space. This means that both Nash-NPRO and Nash-SWRO can steer strategy exploration toward preferred game models, in a relatively stable manner. Moreover, we found that player 1 earns a higher utility than player 2 in all equilibria found by Nash-NPRO and Nash-SWRO, which reveals the advantage of movingfi rst in these equilibria with higher social welfare.

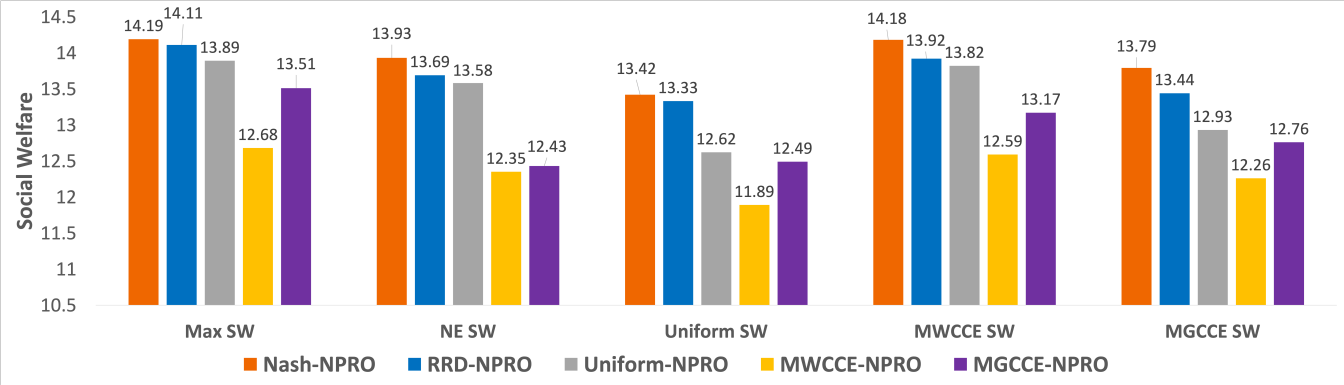
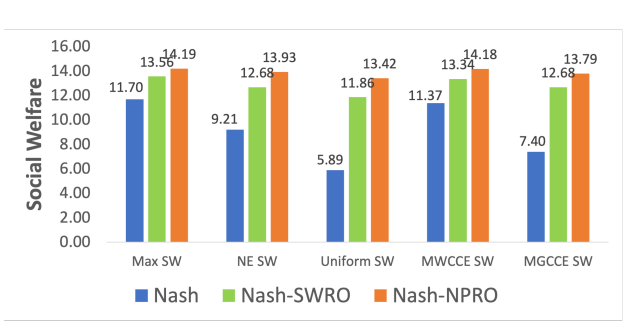
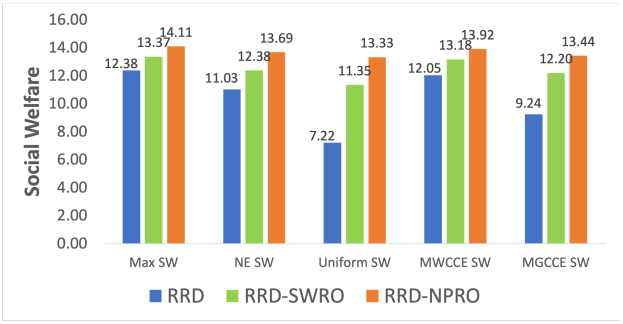


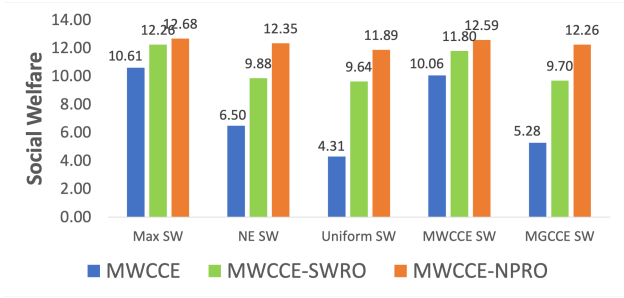
Figure 2 Social welfare of PSRO with MSSs and NPRO evaluated under the same solution concepts.



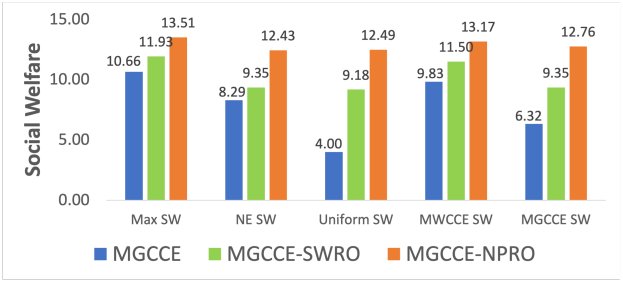
(a) Nash as an MSS.



(b) RRD as an MSS.



(c) MWCCE as an MSS.



(d) MGCCE as an MSS.

Figure 3 Social welfare of MSSs with and without SWRO and NPRO.

To further demonstrate the impact of ROs on strategy exploration, we combine Nash with SERO and list the averaged utilities in equilibria given by some selected MSS-RO combinations in Table 4. we observed that Nash-SERO can efficaciously reduce the utility difference between two players, compared to other combinations. Moreover, we noticed that Nash-SERO causes an increase in social welfare from Nash. This rise can be attributed to the transformation of SERO into a formula that accounts for the utility of both players when $u_i(s_i', \sigma_{-i}) - u_{-i}(s_i', \sigma_{-i}) \geq 0$ and $\alpha > 0.5$.

MSS-RO	$u_1(\sigma^*)$	$u_2(\sigma^*)$	$\lvert u_1(\sigma^*) - u_2(\sigma^*)\rvert$	SW
Nash-NPRO	7.82	6.11	1.71	13.93
Nash-SWRO	9.24	3.44	5.80	12.68
Nash-SERO	5.56	5.83	**0.27**	11.39
Nash	4.26	4.95	0.69	9.21

Table 4 A shrinkage in the utility gap caused by the SERO.

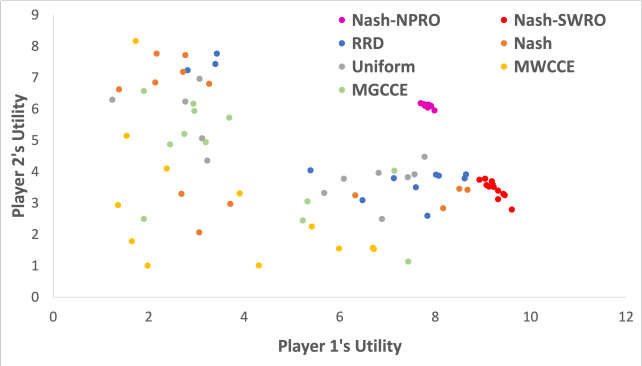


Figure 4 NE scatters in the utility space. Each color represents an MSS-RO combination. Points with the same color are obtained by running PSRO with different random seeds.

5.3 Verification of NE given by Nash-NPRO and Nash-SWRO

To confirm that the NE given by Nash-NPRO and Nash-SWRO is indeed a full-game NE, we compute deviation payoffs using deep Q-network (DQN) [21]. Wefi nd the regrets are on the order of 10^{-2}, which confirms the NE is sufficiently stable with respect to the full game. This stability is supported by the fact that with NE as MSS, iteratively solving ROs can be viewed as iterative ϵ-best responses to NE, which will result in ϵ-NE after convergence. In our scenario, the value of ϵ is closely related to the weighting parameter α and a small or an annealed α will lead to a small ϵ.

In fact, in the context of strategy exploration, the convergence of PSRO can be asserted as long as the empirical strategy space encompasses a full-game NE, as verified by examining the full-game regret. Importantly, this convergence does not necessitate the best-response target aligning with the NE. In general, achieving exact convergence in large games is often unattainable. Consequently, the primary focus of strategy exploration in large games revolves around the development of new algorithms that exhibit strong empirical performance.

6 ATTACK-GRAPH GAMES

Attack graphs [20] are tools in cyber-security analysis employed to model the paths by which an adversary may compromise a system. An ***attack-graph game*** is a two-player general-sum game defined on the attack graph where an attacker attempts to compromise a sequence of nodes to reach *goal* nodes and a defender endeavors to protect any node (e.g., deny an access). Reaching the *goal* nodes within afi nite horizon provides a large benefit for the attacker and a substantial loss for the defender. Both offensive and defensive actions are associated with a cost. The ability of the attacker to choose any subset of feasible nodes and of the defender to defend any subset of the nodes induces action spaces of combinatorial size. We consider an attack-graph game instance with 100 nodes and hence 2^{100} possible combinatorial actions. Since the game is too large to analyze exhaustively, wefi rst construct a particular set of DQN strategies with 125 strategically-diverse strategies in total, following the strategy sampling approach by Czarnecki et al. [9]. Then we apply game-theoretic analysis to this set of strategies. The attack-graph games often have several equilibria exhibiting differing offensive and defensive interactions.

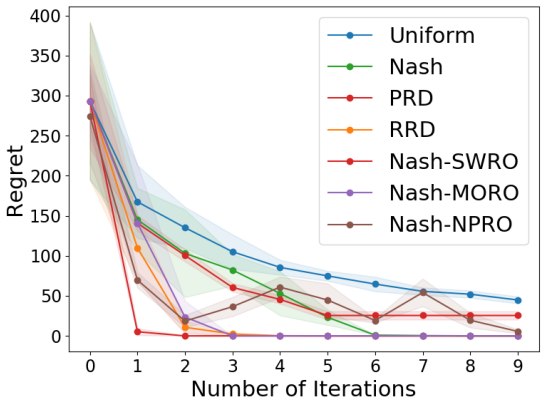


Figure 5 PSRO with different MSS-RO combinations in attack-graph games.

In Fig. 5, we plot the regret curves of different MSS-RO combinations in the attack graph game. We observed that strategy exploration with generalized ROs can affect the convergence speed to an NE. In particular, Nash-SWRO converges to an NE faster than others in this instance. Then we compute the averaged utilities in equilibrium strategies for both players, where the defender (D) and the attacker (A) earn utilities (D: -56.67, A: 27.50) for Nash, (D: -13.43, A: 53.43) for Nash-SWRO, (D: -29.19, A: 47.68) for Nash-NPRO, and (D: -84.91, A: 84.98) for Nash-MORO. An interesting observation is that Nash-SWRO can improve both players' utilities in the equilibrium, though the attack-graph games appear to be purely adversarial. Additionally, we observed that Nash-MORO can enlarge the utility difference between two players and the attacker can cause more damage to the defender.

7 THE CONNECTION OF GENERALIZED ROS TO GWFP

We establish a theoretical connection between PSRO with generalized ROs and GWFP [14]. Inspired by the fact that some PSRO variants can be viewed as classic game-learning dynamics (e.g., PSRO with uniform MSS recovers FP with RL), we establish a theoretical connection between PSRO with generalized ROs and a generalized version of FP, called generalized weakened FP (GWFP) [14]. GWFP unifies FP and weakened FP and generalizes them by allowing perturbations in the strategy updates and by relaxing restrictions on the step sizes. Formally, GWFP is any process $\{\sigma_n\}_{n\geq 0}$ with $\sigma_n \in \prod_{i\in N} \Delta(S)$, where n is index of iterations for FP, such that

$$\sigma_{n+1} \in (1-\alpha_{n+1})\sigma_n + \alpha_{n+1}(br_{\epsilon_n}(\sigma_n) + M_{n+1}), \tag{1}$$

with br_{ϵ_n} being an ϵ_n- best response correspondence, $\alpha_n \to 0$ and $\epsilon_n \to 0$ as $n \to 0$, and $\{M_n\}_{n\geq 1}$ being a sequence of perturbations

such that, for any $T > 0$,

$$\lim_{n\to\infty} \sup_{k} \left\{ \left\| \sum_{i=n}^{k-1} \alpha_{i+1} M_{i+1} \right\| : \sum_{i=n}^{k-1} \alpha_i \le T \right\} = 0. \quad (2)$$

It is known that GWFP recovers FP by setting $\epsilon_n = 0$, $M_n = 0$, and $\alpha_n = \frac{1}{n}$ for all n, and shares the same convergence property with FP (e.g., convergence to NE in two-player zero-sum games, potential games, and $2 \times |S_i|$ generic games).

Intuitively, PSRO with uniform MSS and a regularizer weighted by $1 - \alpha$ (e.g., the other players' utility in SWRO) will almost surely converge to GWFP as $\alpha \to 1$ (i.e., the regularizer vanishes and the RO will converge to the original RO) and thus sharing the same convergence properties as GWFP. Formally, we have the following theorem with a detailed proof in Appendix D.1.

Theorem 1. With a strictly differentiable concave regularizer associated with an infinitesimal temperature parameter in ROs, PSRO with uniform MSS and an exact RO solver belongs to the class of GWFP.

8 PSRO WITH GENERALIZED ROS FOR COMPUTING BERGE EQUILIBRIUM

Besides learning toward NE with particular features, PSRO with generalized ROs can also be employed as a tool for computing other solution concepts. In particular, we focus on computing Berge equilibrium [5], a common solution in the game-theoretic study of philosophy and social interactions, in which each player ensures that all other players will receive the highest payoff. We employ BE as an MSS and introduce a generalized RO based on the definition of BE. We show that PSRO with this MSS-RO combination will stop at a full-game BE in two-player games. Compared to prior methods that compute BE by enumerating all profiles [8], our method enables BE computation in large games.

8.1 Berge Equilibrium

We follow the definition of BE from an individual perspective given by Zhukovskii [36], though BE wasfi rst defined in terms of coalitions by Berge [5].

A strategy profile $\sigma^B \in \Delta(S)$ is a ***Berge equilibrium*** if for $i \in \{1, 2\}$ and all $s_{-i} \in S_{-i}$,

$$u_i(\sigma^B) \ge u_i(\sigma_i^B, s_{-i}). \quad (3)$$

This definition means that for any particular player i, its utility would not increase if it sticks to its own BE strategy while other players can change their strategies. This can be viewed as the altruism in the game playing since in a BE each player ensures the highest payoff for all other players who are also employing their BE strategy. Note that this is different from the spirit of NE, where players are assumed to be selfish and only maximize their own payoff.

8.2 Computing Berge Equilibria with PSRO

To adapt PSRO for computing BE, we should answer the following two questions: a) How can we compute a BE in the current empirical game if BE is employed as an MSS? b) Which response objective will steer strategy exploration toward a full-game BE effectively?

To answer thefi rst question, based on the definition of BE, it can be simply proved that a BE of the empirical game can be obtained by computing NE of the corresponding utility-swapping game, assuming a BE of the empirical game exists. Details of the proof is included in Appendix D.2.

Proposition 1. Given a two-playerfi nite game $\mathcal{G} = (\{1, 2\}, (S_i), (u_i))$, σ^B is a BE of $\mathcal{G}$ if and only if it is an NE of the utility function swapping game $\mathcal{G}' = (\{1, 2\}, (S_i), (z_i))$, where $z_i = u_{-i}$ for $i \in \{1, 2\}$.

Now we can confirm that a BE exists in an empirical game since an NE exists in the utility-swapping game [24].

Corollary 1. For two-playerfi nite games, a BE exists in both the full game and the empirical game.

To answer the second question, we propose the ***Berge Equilibrium Response Objective*** (BERO) based on the definition of BE, which simply takes the form of $u_{-i}(s_i, \sigma_{-i})$. BERO can be viewed as a special case of SWRO where $\alpha = 0$. With BE as an MSS and BERO, we show the Berge PSRO algorithm for computing BE in Appendix C.

Proposition 2. With exact best response oracles, PSRO with BE as an MSS and BERO will stop at a full-game BE in two-playerfinite games.

8.3 Revisiting the Traveler's Dilemma

In the Traveler's dilemma shown in Table 1, Berge PSRO will stop at the profile (5, 5) with payoff (6, 6), which is a BE of this game since if one player deviates, another player's payoff will not increase. We also noticed that the profile (5, 5) is a desired profile with high social welfare and low regret in this game.

Despite the higher social welfare of BE in the Travler's dilemma, it is not always the case that a BE will have higher social welfare than an NE. For example, consider a two-player game with the utility function $u_i(s_i, s_{-i}) = -5s_i + s_{-i}$ where $s_i \in \{+1, 1\}$ for $i \in \{1, 2\}$. The BE of this game is (1, 1) with social welfare -8 while the NE is (-1, -1) with social welfare 8. This is caused by individual irrationality of BE and can be addressed by refinements of BE [1, 37].

9 CONCLUSION

We study the effectiveness of setting customized ROs for guiding strategy exploration toward desired empirical games under the PSRO framework. Through systematically investigating various MSS-RO combinations in sequential bargaining games and attack-graph games, we show that ROs can steer strategy exploration toward empirical games with solutions aligned with specified objectives. Using BE as an example, we show that PSRO with generalized ROs can be employed for computing solution concepts other than NE. Theoretically, we prove that PSRO with certain ROs belongs to the class of GWFP.

ACKNOWLEDGMENTS

This work was supported in part by funding from the US Army Research Office (MURI grant W911NF-18-1-0208), and a grant from the Effective Altruism Foundation.

REFERENCES

[1] Kokou Y. Abalo and Michael M. Kostreva. 2004. Some existence theorems of Nash and Berge equilibria. *Applied Mathematics Letters* 17, 5 (2004), 569–573.

[2] David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech M Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. 2019. Open-ended learning in symmetric zero-sum games. In *36th International Conference on Machine Learning*. 434–443.

[3] Kaushik Basu. 1994. The traveler's dilemma: Paradoxes of rationality in game theory. *American Economic Review* 84, 2 (1994), 391–395.

[4] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. 2005. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization* 44, 1 (2005), 328–348.

[5] Claude Berge. 1957. *Théorie Générale des Jeux à n Personnes*. Vol. 138. Gauthier-Villars Paris.

[6] George W. Brown. 1951. Iterative solution of games byfi ctitious play. In *Activity Analysis of Production and Allocation*, T. C. Koopmans (Ed.). Wiley, 374–376.

[7] Vincent Conitzer and Caspar Oesterheld. 2022. Foundations of cooperative AI. *AAAI-23 Senior Member Blue Sky Ideas track* (2022).

[8] H. W. Corley and Phantipa Kwain. 2015. An algorithm for computing all Berge equilibria. *Game Theory* 2015, 862842 (2015).

[9] Wojciech Marian Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, and Max Jaderberg. 2020. Real world games look like spinning tops. In *34th Conference on Neural Information Processing Systems*.

[10] Drew Fudenberg and David K. Levine. 1995. Consistency and cautiousfictitious play. *Journal of Economic Dynamics and Control* 19, 5-7 (1995), 1065–1089.

[11] Drew Fudenberg and Jean Tirole. 1983. Sequential bargaining with incomplete information. *Review of Economic Studies* 50, 2 (1983), 221–247.

[12] Patrick R. Jordan, L. Julian Schvartzman, and Michael P. Wellman. 2010. Strategy Exploration in Empirical Games. In *9th International Conference on Autonomous Agents and Multi-Agent Systems* (Toronto). 1131–1138.

[13] Marc Lanctot, Vinicius Zambaldi, Audrūnas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *31st Annual Conference on Neural Information Processing Systems* (Long Beach, CA). 4190–4203.

[14] David S. Leslie and Edmund J. Collins. 2006. Generalised weakenedfictitious play. *Games and Economic Behavior* 56, 2 (2006), 285–298.

[15] Zun Li, Marc Lanctot, Kevin R. McKee, Luke Marris, Ian Gemp, Daniel Hennes, Paul Muller, Kate Larson, Yoram Bachrach, and Michael P. Wellman. 2023. Combining tree-search, generative models, and Nash bargaining concepts in game-theoretic reinforcement learning. *arXiv preprint arXiv:2302.00797* (2023).

[16] Zongkai Liu, Chao Yu, Yaodong Yang, Peng Sun, Zifan Wu, and Yuan Li. 2022. A unified diversity measure for multiagent reinforcement learning. In *36th Conference on Neural Information Processing Systems*. 10339–10352.

[17] Luke Marris, Paul Muller, Marc Lanctot, Karl Tuyls, and Thore Graepel. 2021. Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers. In *38th International Conference on Machine Learning*. 7480–7491.

[18] Stephen McAleer, Kevin Wang, John B Lanier, Marc Lanctot, Pierre Baldi, Tuomas Sandholm, and Roy Fox. 2022. Anytime PSRO for two-player zero-sum games. *CoRR, abs/2201.07700, 2022b. URL https://arxiv. org/abs/2201.07700* (2022).

[19] H. Brendan McMahan, Geoffrey J. Gordon, and Avrim Blum. 2003. Planning in the presence of cost functions controlled by an adversary. In *20th International Conference on Machine Learning*. 536–543.

[20] Erik Miehling, Mohammad Rasouli, and Demosthenis Teneketzis. 2015. Optimal defense policies for partially observable spreading processes on Bayesian attack graphs. In *2nd ACM Workshop on Moving Target Defense*. 67–76.

[21] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[22] Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, et al. 2020. A generalized training approach for multiagent learning. In *8th International Conference on Learning Representations* (virtual).

[23] John F. Nash Jr. 1950. The bargaining problem. *Econometrica: Journal of the econometric society* (1950), 155–162.

[24] John F. Nash Jr. 1950. Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences* 36, 1 (1950), 48–49.

[25] Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. 2019. α-rank: Multi-agent evaluation by evolution. *Scientific Reports* 9, 1 (2019), 1–29.

[26] Jack Parker-Holder, Aldo Pacchiano, Krzysztof M. Choromanski, and Stephen J. Roberts. 2020. Effective diversity in population based reinforcement learning. In *34th Conference on Neural Information Processing Systems*. 18050–18062.

[27] Nicolas Perez-Nieves, Yaodong Yang, Oliver Slumbers, David H. Mguni, Ying Wen, and Jun Wang. 2021. Modelling behavioural diversity for learning in open-ended games. In *International Conference on Machine Learning*. PMLR, 8514–8524.

[28] Ariel Rubinstein and Asher Wolinsky. 1985. Equilibrium in a market with sequential bargaining. *Econometrica: Journal of the Econometric Society* (1985), 1133–1150.

[29] Karl Tuyls, Julien Pérolat, Marc Lanctot, Edward Hughes, Richard Everett, Joel Z. Leibo, Csaba Szepesvári, and Thore Graepel. 2020. Bounds and dynamics for empirical game theoretic analysis. *Autonomous Agents and Multi-Agent Systems* 34 (2020), 7.

[30] Ben Van der Genugten. 2000. A weakened form offi ctitious play in two-person zero-sum games. *International Game Theory Review* 2, 04 (2000), 307–328.

[31] Yongzhao Wang, Qiurui Ma, and Michael P. Wellman. 2022. Evaluating Strategy Exploration in Empirical Game-Theoretic Analysis. In *23th International Conference on Autonomous Agents and Multi-Agent Systems*. 1346–1354.

[32] Yufei Wang, Zheyuan Ryan Shi, Lantao Yu, Yi Wu, Rohit Singh, Lucas Joppa, and Fei Fang. 2019. Deep reinforcement learning for green security games with real-time information. In *33rd AAAI Conference on Artificial Intelligence*. 1401–1408.

[33] Yongzhao Wang and Michael P Wellman. 2023. Regularization for Strategy Exploration in Empirical Game-Theoretic Analysis. *arXiv preprint arXiv:2302.04928* (2023).

[34] Michael P. Wellman. 2016. Putting the agent in agent-based modeling. *Autonomous Agents and Multi-Agent Systems* 30 (2016), 1175–1189.

[35] Mason Wright, Yongzhao Wang, and Michael P. Wellman. 2019. Iterated Deep Reinforcement Learning in Games: History-Aware Training for Improved Stability. In *20th ACM Conference on Economics and Computation* (Phoenix). 617–636.

[36] Vladislav I. Zhukovskii. 1985. Some problems of non-antagonistic differential games. *Matematiceskie Metody v Issledovanii Operacij* (1985), 103–195.

[37] Vladislav I Zhukovskii and Arkadii A Chikrii. 1994. Linear quadratic differential games. *Naoukova Doumka, Kiev* (1994).