Multi-agent learning in mixed-motive coordination problems

Julian Stastny Johannes Treutlein Maxime Riché Jesse Clifton

Abstract

Cooperation in settings where agents have different but overlapping preferences (*mixed-motive* settings) has recently received considerable attention in multi-agent learning. However, the mixed-motive environments typically studied are simplistic in that they have a single cooperative outcome on which all agents can agree. Multi-agent systems in general may exhibit many payoff profiles which might be called cooperative, but which agents have different preferences over. This causes problems for independently trained agents that do not arise in the case of that there is a unique cooperative payoff profile. In this note, we illustrate this problem with a class of games called *mixed-motive coordination problems (MCPs)*. We demonstrate the failure of several methods for achieving cooperation in sequential social dilemmas when used to independently train policies in a simple MCP. We discuss some possible directions for ameliorating MCPs.

1 Introduction

In *An Essay on Bargaining* [1956], Thomas Schelling describes a negotiation over the price of a house. Considering different strategies the negotiators might use, he points out that

> If the buyer is an agent authorized by a board of directors to buy at $16000 but not a cent more, and the directors cannot constitutionally meet again for several months and the buyer cannot exceed his authority, and if all this can be made known to the seller, then the buyer 'wins' — if, again, the seller has not tied himself up with a commitment to $19000. (pg. 284)

Imagine various actors with different preferences — a *mixed-motive* setting [Schelling, 1958] — developing machine learning systems to act on their behalf. These actors are like Schelling's bargainers, in that their systems implicitly encode a commitment to pursuing a particular tradeoff between their own preferences and those of other players. Just as the price negotiation in Schelling's example may fail if buyer and seller are each committed to insisting on incompatible prices, machine learning systems may fail to cooperate if they encode commitments to incompatible tradeoffs between their principals' goals.

Multi-agent learning (MAL) has in recent years paid considerable attention to problems of cooperation between agents with different, but not zero-sum, preferences. However, much of this work has been in the setting of *(sequential) social dilemmas (SSDs)* (e.g., Peysakhovich and Lerer 2017a, Lerer and Peysakhovich 2017, Peysakhovich and Lerer 2017b, Foerster et al. 2018, Wang et al. 2018, Eccles et al. 2019). The classic example of a social dilemma is the Prisoner's Dilemma (PD), and the SSDs studied in this literature are similar to the Prisoner's Dilemma in that there is a single salient notion of "cooperation". This means that it is relatively easy for actors to coordinate in their selection of policies to deploy in these settings. This situation is quite unlike that faced by Schelling's bargainers, for whom there are many possible agreements to commit to, and who thus may fail to reach an agreement if they make incompatible commitments[1].

Cao et al. [2018] are an exception in that they look at negotiation between deep reinforcement learners, such that there are multiple incompatible ways in which agents could choose among possible agreements. However, they do not evaluate the performance of independently trained policies when played against one

Working draft. Last updated March 8, 2021.

[1]The inadequacy of the Prisoner's Dilemma as a model for cooperation problems, and the need to study games that exhibit competing "cooperative" solutions, has been pointed out in other fields. See Snidal [1985] for a discussion in international relations, McAdams [2008] in law, and Binmore [1998] in evolutionary game theory.

another ("cross-play" [Hu et al., 2020]), which is where the existence of multiple incompatible cooperative solutions creates problems.

In this note, we argue that a focus on sequential social dilemmas and the lack of cross-play evaluation obscures a fundamental problem for decentralized MAL in mixed-motive settings. We discuss a class of games that we call *mixed-motive coordination problems (MCPs)*, which highlight this problem. We compare MCPs to pure coordination problems and SSDs. We then illustrate how MAL methods which achieve cooperation in other social dilemmas fail in some simple MCPs, and argue that future methods for promoting cooperation in mixed-motive settings should be benchmarked in these more difficult cases. We close with some tentative positive contributions, by sketching some partial solutions to MCPs.

2 Preliminaries

We work in the setting of partially observable stochastic games (POSGs). For simplicity we assume two players, $i = 1, 2$. We will index player i's counterpart by $-i$. Each player has an action space $\mathcal{A}_i$, there is an underlying state space $\mathcal{S}$ which evolves according to a Markovian transition function $P(S^{t+1} \mid S^t, A_1^t, A_2^t)$. At each time step, each player sees an observation O_i^t which depends on S^t. Thus each player has an accumulating history of observations $\mathcal{H}_i^t = \{O_1^v, A_1^v\}_{v=1}^t$. Call the set of all histories $\mathcal{H}_i = \cup_{t=1}^{\infty} \mathcal{H}_i^t$. Assume for simplicity that the initial observation history is fixed and common knowledge, $h_1^0 = h_2^0 \equiv h^0$. Finally, principals choose among policies $\pi_i : \mathcal{H} \to \Delta(\mathcal{A}_i)$, which we imagine as artificial agents deployed by the principals. We will refer to policy profiles generically as $\pi \in \Pi := \Pi_1 \times \Pi_2$.

Each player has a reward function r_i, such that $r_i(S^t, A_1^t, A_2^t)$ is their reward at time t. Define the value to player i of policy profile π starting at history h_i^t as $V_i(h_i^t, \pi) = \mathbb{E}_\pi \left[\sum_{v=t}^{\infty} \gamma^{v-t} r_i(S^v, A_1^v, A_2^v) \mid H_i^t = h_i^t\right]$, where $\gamma < 1$ is a discount factor. Define the value of a policy profile to player i as $V_i(\pi) = V_i(h^0, \pi)$. A payoff profile is then a tuple $(V_1(\pi), V_2(\pi))$. We say that π is a (Nash) equilibrium of a POSG if $\pi_i \in \arg\max_{\pi_i' \in \Pi_i} V_i(h^0, \pi_i', \pi_{-i})$ for $i = 1, 2$. We say that π is Pareto-optimal if for $i = 1, 2$ and $\pi' \in \Pi$, we have that $V_i(\pi') > V_i(\pi)$ implies $V_{-i}(\pi') < V_{-i}(\pi)$.

3 Mixed-motive coordination problems

We envision multiple actors ("principals") deploying machine learning systems, defined by policies, into the world. The principals may be human developers, or they may be machine learning systems themselves. The principals simultaneously deploy policies, without communication (in Section 6 we discuss ways in which communication may change the analysis). Let $\mathcal{W}$ be a set of *welfare functions*, which are supposed to encode normative properties (total welfare, fairness, etc.) of different payoff profiles. Two uncontroversial properties of a welfare function are *Pareto-optimality* (i.e., its optimizer should be Pareto-optimal) and *symmetry* (the welfare of a policy profile should be invariant to permutations of player indices). Beyond that, there are several appealing properties which cannot all be satisfied by the same welfare function (see Table 1). The experimental evidence regarding people's preferences over welfare functions is inconclusive [Felsenthal and Diskin, 1982]. We will say that a game is a *mixed-motive coordination problem (MCP)* if it has several equilibria[2] which are optimal with respect to some welfare function, but are incompatible with each other.

Definition 3.1 (Mixed-motive coordination problem)**.** Let G be a POSG. Let $\mathcal{W}$ be a set of welfare functions, and let $\mathcal{W}^E \subseteq \mathcal{W}$ be the set of welfare functions in $\mathcal{W}$ which are optimized by some equilibrium of G. Then G is a mixed-motive coordination problem (with respect to $\mathcal{W}$) if there exist distinct $w, w' \in \mathcal{W}^E$ such that $(\pi_1^w, \pi_2^{w'})$ has a Pareto-suboptimal policy profile for all $\pi^w \in \arg\max_{\pi'} w(\pi')$, $\pi^{w'} \in \arg\max_{\pi'} w'(\pi')$.

An example of a mixed-motive coordination problem is the Bach or Stravinsky game. A variant of this game with asymmetric payoffs is given in Figure 1 (left). The folk theorem [Friedman, 1971] tells us that any payoff profile in the convex hull of these payoffs is attainable in an equilibrium of the repeated game, for players with sufficiently high discount factors. And, as is suggested by the shape of the Pareto frontier for

[2]The spirit of an MCP is that independent training predictably leads to policy profiles which optimize different welfare functions, but are incompatible. So one could modify the requirement that welfare-optimal policies be equilibria and retain this spirit. One could replace the requirement of Nash equilibrium with a different or a broader set of solution concepts, e.g., the recently-proposed Markov-Conley chains used in the α-rank evaluation methodology [Omidshafiei et al., 2019].

Name of welfare function w	**Form of $w(\pi)$**
Nash [Nash, 1950]	$[V_1(\pi) - d_1] \cdot [V_2(\pi) - d_2]$
Kalai-Smorodinsky [Kalai and Smorodinsky, 1975]	$\left\lvert \frac{V_1(\pi)-d_1}{V_2(\pi)-d_2} - \frac{\sup_\pi V_1(\pi)-d_1}{\sup_\pi V_2(\pi)-d_2} \right\rvert +$ $\iota\{\pi \text{ Pareto-optimal }\}$
Egalitarian [Kalai, 1977]	$\min\{V_1(\pi) - d_1, V_2(\pi) - d_2, \}$
Utilitarian (e.g. Harsanyi 1955)	$V_1(\pi) + V_2(\pi)$

Table 1: Welfare functions, adapted to the multi-agent RL setting where two agents with value functions V_1, V_2 are bargaining over the policy profile π. to deploy. The function ι in the definition of the Kalai-Smorodinsky welfare is the ∞-0 indicator, used to enforce the constraint in its argument. $V_i^d = V_i(\pi^d)$ are the players' disagreement values.

Note that in bargaining problems in which the set of feasible payoffs is convex, the Nash welfare function uniquely satisfies the properties of (1) Pareto optimality, (2) symmetry, (3) invariance to affine transformations, and (4) independence of irrelevant alternatives. See Zhou [1997] for a characterization of the Nash welfare in non-convex bargaining problems. The Nash welfare can also be obtained as the subgame perfect equilibrium of an alternating-offers game as the "patience" of the players goes to infinity [Binmore et al., 1986].

On the other hand, Kalai-Smorodinsky uniquely satisfies (1)-(3) plus (5) resource monotonicity, which means that all players are weakly better off when there are more resources to go around. The egalitarian solution instead satisfies (1), (2), (4), and (5). The utilitarian welfare function is implicitly used in the work of Peysakhovich and Lerer [2017a], Lerer and Peysakhovich [2017], Wang et al. [2018] on cooperation in sequential social dilemmas.

BoS	B	S
B	4, 1	0, 0
S	0, 0	2, 2

PD	C	D
C	-1, -1	-3, 0
D	0, -3	-2, -2

Figure 1: Payoffs for asymmetric Bach or Stravinsky (BoS) and the Prisoner's Dilemma (PD).

BoS (Figure 2, left), different welfare functions pick out different Pareto-optimal payoff profiles. For instance, taking the disagreement point to be $d_1, d_2 = (0, 0)$, the Nash welfare is optimized at $(3.0, 1.5)$, whereas the egalitarian welfare is optimized at $(2.0, 2.0)$. On the other hand, all welfare functions in Table 1 agree on the payoffs corresponding to (C, C) in the iterated Prisoner's Dilemma (Figure 2, right).

In the following, we compare mixed-motive coordination problems to the more general equilibrium selection problem and to sequential social dilemmas.

3.1 Relation to other cooperation problems

Equilibrium and Pareto selection problems It is well-known that achieving cooperation between independently-trained agents this is a challenging problem for MAL in general (e.g., Boutilier 1999, Matignon et al. 2012, Hu et al. 2020). One reason is that there are often multiple equilibria, such that (to the extent that one wants to deploy a policy which is a best-response to that of one's counterpart) the choice of a policy constitutes an *equilibrium selection problem*.

Such problems even arise in the fully cooperative case, i.e., when $V_1 = V_2 \equiv V$, and are then also called Pareto selection problems [Matignon et al., 2012] or pure coordination problems. In a pure coordination problem, any $\pi \in \arg\max_\pi V(\pi)$ has an optimal value and is thus in equilibrium. There is an equilibrium selection problem if there exist $\pi, \pi' \in \arg\max_\pi V(\pi)$ such that (π_1, π'_2) has Pareto-suboptimal payoff profile.

Mixed-motive coordination problems that occur between policy profiles in equilibrium are a subclass of equilibrium selection problems. They only occur in general-sum, non-fully cooperative cases and pose a challenge which is qualitatively different from the fully cooperative case: players have to coordinate among

options over which they have conflicting preferences. The pure coordination problem can be solved by merely giving the players a chance to converge upon a coordinated policy profile (as in Boutilier [1999], for instance) and any such policy will do equally well. This is not so in mixed-motive coordination problems, since agents that are open to converging upon any Pareto-optimal outcome would be exploitable by opponents that insist on an outcome they prefer. (See Example 6.0.1 for a concrete illustration of this difference.)

Sequential social dilemmas Sequential social dilemmas (SSDs), as defined by Leibo et al. [2017], are POSGs whose policy space contains subsets corresponding to cooperation and defection, called Π_i^C and Π_i^D, respectively. At each state, the normal form game whose payoffs are value functions under profiles of policies in Π_i^C, Π_i^D satisfies the social dilemma inequalities [Macy and Flache, 2002], where atomic "cooperate" and "defect" actions are replaced with elements of Π_i^C, Π_i^D. An SSD is not necessarily an equilibrium selection problem. For instance, any Pareto-optimal and symmetric policy profile will select an equilibrium profile that leads to constant cooperation in the iterated Prisoner's Dilemma, and which is compatible with other such profiles. But other SSDs may exhibit multiple incompatible ways of implementing cooperation.

SSDs with symmetric Pareto-optimal equilibria are in any case not mixed-motive coordination problems. This includes the Gathering, Wolfpack [Leibo et al., 2017], and Coin Game [Peysakhovich and Lerer, 2017a] environments. For instance, in the Coin Game, the single cooperative payoff profile is attained when agents get only the coin of their own color, rather than attempting to take their counterpart's coin as well.

4 Learning algorithms

Play between policies trained by different runs or variations of an algorithm, or "cross-play" [Hu et al., 2020] is problematic for MCPs. To illustrate the need for evaluating policies in cross-play, we train policies using two multi-agent learning algorithms which have been shown to achieve cooperation in SSDs. The first is a variant of Lerer and Peysakhovich [2017]'s approximate Markov tit-for-tat (amTFT). The original amTFT algorithm trains a cooperative policy on the utilitarian welfare (Table 1), as well as a punishment policy, and switches from the cooperative to the punishment policy when it detects that the other player is defecting from the cooperative policy. We consider the more general class of algorithms in which a cooperative policy is constructed by optimizing some welfare function. Call the version of amTFT in which the cooperative policy optimizes welfare function w as $\texttt{amTFT}(w)$.

In our experiments, we use the utilitarian welfare function w^{Util} and an *inequity averse* welfare function w^{IA}. Following Hughes et al. [2018], this welfare function is constructed using modified reward functions which penalize inequitable outcomes:

$$V_i^{\text{IA}}(a_1^t, a_2^t; \alpha, \beta) = r_1(a_1^t, a_2^t) + r_1(a_2^t, a_2^t) - \alpha[e_1^t(a_1^t, a_2^t) - e_2^t(a_1^t, a_2^t)]_+ - \beta[e_2(a_1^t, a_2^t) - e_1(a_1^t, a_2^t)]_+$$

where $[x]_+ = \max\{x, 0\}$ and e_i^t are smoothed cumulative rewards computed as $e_i^t(a_1^t, a_2^t) = \gamma\lambda e_i^{t-1}(a_1^{t-1}, a_2^{t-1}) + r_i(a_1^t, a_2^t)$ with hyperparameter λ. and discount factor γ α and β control how much unequal outcomes are penalized.[3]

The second learning algorithm is learning with opponent-learning awareness (LOLA; Foerster et al. 2018) a policy gradient method in which learners attempt to shape the others' learning through the choice of their own policy updates. Write $V_i(\theta_1, \theta_2)$ as the value to player i under a profile of policies with parameters θ_1, θ_2. Then, the LOLA update for player 1 at time t with parameters $\delta, \eta > 0$ is

$$\theta_1^t = \theta_1^{t-1} + \delta\nabla_{\theta_1} V_i(\theta_1, \theta_2) + \delta\eta \left[\nabla_{\theta_2} V_1(\theta_1, \theta_2)\right]^{\mathsf{T}} \nabla_{\theta_1}\nabla_{\theta_2} V_2(\theta_1, \theta_2).$$

We use the discount factor $\gamma = 0.96$ throughout.

[3] We choose the utilitarian and inequity-averse welfare functions largely for convenience, and not because they have particularly compelling normative properties. Specifically, these welfare functions admit a Bellman equation, allowing us to straightforwardly apply reinforcement learning methods in order to maximize them. Methods for optimizing other welfare functions (Table 1), which may be less amenable to off-the-shelf reinforcement learning methods, are a direction for future work.

5 Results

Here we illustrate the special problem posed by MCPs by comparing the performance of the learning algorithms above in iterated BoS (IBoS) and in the iterated Prisoner's Dilemma (IPD). The folk theorems [Mailath et al., 2006] tell us that, for players with sufficiently high discount factors, any payoff profile of the stage game in which players get at least as much as they can unilaterally guarantee themselves can be attained in an equilibrium of the repeated game. IBoS is an MCP with respect to the set of welfare functions $\mathcal{W} = \{w^{\text{Util}}, w^{\text{IA}}\}$, since these welfare functions are optimized by distinct payoff profiles for sufficiently large values of α, β.

We measure the performance of these algorithms in *self-play* and *cross-play*. Self-play payoffs are the average payoffs attained by a profile of policies produced by the same training run. Cross-play payoffs are the average payoffs attained by a profile of two policies produced by two different training runs. In the case of `amTFT`, we also consider cross-play between policies trained with `amTFT`(w^{Util}) and those trained with `amTFT`(w^{IA}).

5.1 LOLA

Following Foerster et al. [2018]'s parameterization of policies in the iterated Prisoner's Dilemma, we used policies which condition on the previous pair of actions played, with one parameter for every possible previous action profile. To learn policy profiles, we used a discount factor of $\gamma = 0.96$ and applied LOLA to the closed-form value functions (which can be computed for each policy profile using the Bellman equations; this is called exact LOLA in Foerster et al. [2018]).

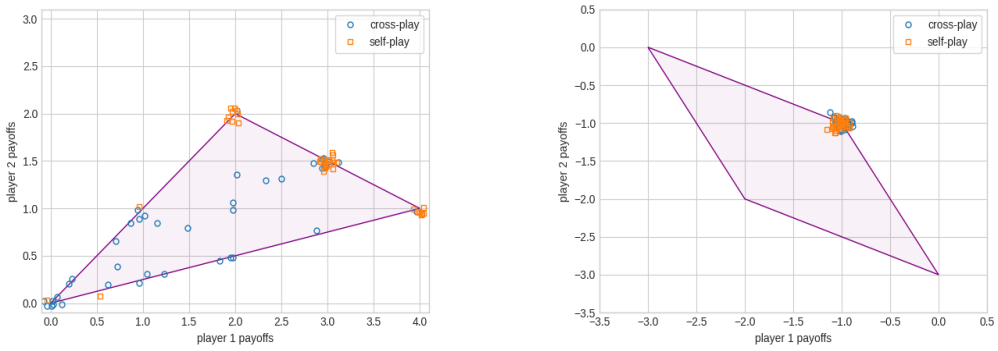

Figure 2: LOLA average payoff profiles for 40 training runs in IBoS (left) and IPD (right). Cross-play payoffs were obtained by randomly sampling 40 pairs (π_1, π_2), where π_1 and π_2 come from different runs. A small amount of random noise is applied to each point to enhance visibility.

On average, self-play yields payoffs $(2.69, 1.34)$, while cross-play yields $(1.30, 0.70)$. As Figure 2 shows, the self-play payoffs almost all lie close to the empirical Pareto frontier, which consists of payoff profiles roughly equal to $\{(4, 1), (3, 1.5), (2, 2)\}$. This clustering is caused by the policies depending only on actions taken at the previous timestep, which allows for always playing (S,S), alternating between them, or always playing (B,B). These coincide with the maxima for w^{Util}, the Nash welfare [Nash, 1950] under disagreement profile $(0, 0)$, or w^{IA}.

On the other hand, the cross-play points are much more often Pareto-suboptimal. This is not surprising, as if LOLA produces tit-for-tat-like strategies which punish deviations from a particular strategy profile, then playing policies from separate runs is liable to lead to incompatible policies and therefore punishments.

5.2 amTFT

We evaluated self- and cross-play for amTFT in IBoS and IPD, where each episode lasted 100 of steps. Payoff profiles from individual runs are displayed in Figure 3, and average payoffs under self- and cross-play for each

combination of welfare functions are displayed in Table 2.

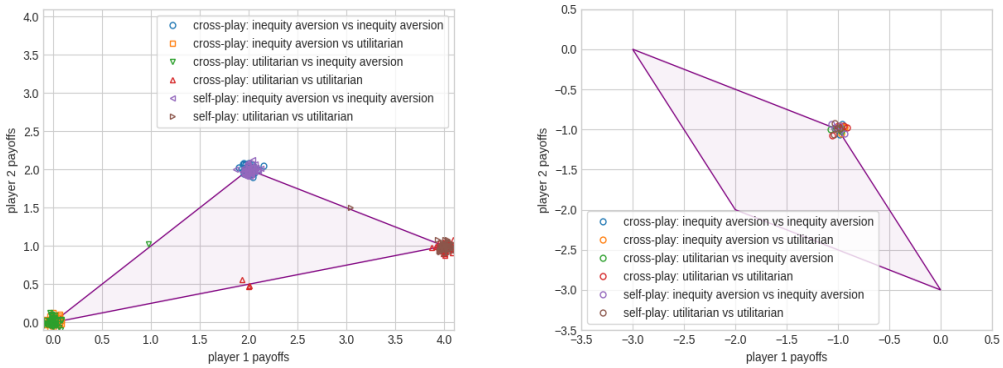

Figure 3: amTFT payoff profiles for 40 training runs in IBoS (left) and IPD (right). A small amount of random noise is applied to each point to enhance visibility.

Table 2: Self-play ("$\circ$") and cross-play ("$\times$") payoffs in IBoS for $\texttt{amTFT}(w^{\text{Util}})$ and $\texttt{amTFT}(w^{\text{IA}})$. Payoffs are averages over 40 training runs. $w^{\text{Util}} \times w^{\text{IA}}$ leads to the same payoff regardless of which player operates under which welfare function. All standard errors were below 0.01.

Policy profile	$w^{\text{Util}} \circ w^{\text{Util}}$	$w^{\text{IA}} \circ w^{\text{IA}}$	$w^{\text{Util}} \times w^{\text{Util}}$	$w^{\text{IA}} \times w^{\text{IA}}$	$w^{\text{Util}} \times w^{\text{IA}}$
Player 1	3.97	2.00	3.87	2.00	0.00
Player 2	1.01	2.00	0.97	2.00	0.00

6 Possible solutions to MCPs

Of course, the goal of developing benchmarks is not only to point out flaws in existing approaches, but also to help develop better ones. In this section we discuss several research directions for ameliorating MCPs, and give a toy illustration of one approach.

Coordination by the principals MCPs can be solved if the principals can agree on a way of selecting among policy profiles to deploy. For instance, principals might agree on a welfare function, and choose learning algorithms which converge to an optimizer of that welfare function. Coordination on a welfare function (and among equilibria which all maximize that welfare function) may be easy in toy environments such as the ones we have considered. But as systems become more complex, it will likely be intractable for principals to fully coordinate so as to train welfare-optimal agents. Thus there is a need to develop methods for low-bandwidth coordination between principals training agents in MCPs which captures most of the benefits of full coordination. Bandwidth-constrained coordination has already been studied in fully cooperative games (e.g., Stone and Veloso 1999, Zhang et al. 2019).

Compromise between welfare functions We have not discussed how giving agents the opportunity to communicate post-deployment might ameliorate MCPs. One possibility is to have them reach a compromise over different welfare-optimal policy profiles before deciding which to deploy. On possibility is for the train to several policies corresponding to several welfare functions according to which they are willing to play. When given the chance to communicate, their agents can then apply a bargaining protocol to this reduced set of policy profiles. Here is a simple example of such a protocol. Agents first announce the sets of policy profiles they are willing to play. If these sets overlap, a policy profile is taken at random from the intersection (compare with Chakravorti and Conley [2004]'s discussion of randomization between the Nash and Kalai-Smorodinsky

Pure Coordination	A	B
A	1, 1	0, 0
B	0, 0	1, 1

Figure 4: Payoffs for a pure coordination problem.

bargaining solutions in the asymmetric iterated Prisoner's Dilemma). Only when these sets do not overlap do agents revert to a default policy. We give an illustration in Example 6.0.1.

Example 6.0.1 (Bargaining over amTFT policies)**.** An example of a protocol for bargaining over welfare-optimal policy profiles is given in Algorithm 1. Policies are obtained by applying amTFT to a welfare function w, returning a policy $\texttt{amTFT}(w)_i$ for player i. Refer to the strategy which reports welfare functions $\mathcal{W}_i$ and default policy π_i^d as $(\mathcal{W}_i, \pi_i^d)$. Table 3 displays the payoff table of the game with strategy profiles $\{(\mathcal{W}_1, \pi_1^d), (\mathcal{W}_2, \pi_2^d)\}$ in this bargaining protocol applied to IBoS.

This payoff matrix clearly shows the tradeoff between exploitability and robustness to independently-trained policies. The strategies which play $\{w^{\text{Util}}, w^{\text{IA}}\}$ are of more robust, in the sense that they avoid the mutually dispreferred profile of $(1.0, 0.3)$. However, these strategies are *not* played in equilibrium, because they are exploitable: an agent can deviate from this profile by instead playing only their preferred welfare function. Nevertheless, one has the intuition that it is reasonable to play $\{w^{\text{Util}}, w^{\text{IA}}\}$ in view of its robustness, despite the fact that it is exploitable. On the other hand, the situation becomes more complicated when the set of welfare functions under consideration is much larger. Identifying reasonable solution concepts for this setting is an important direction for future work.

Besides the fact that avoiding bargaining failure is still not guaranteed, there are additional problems for this protocol. First, it may be ex-ante Pareto suboptimal, because randomization between Pareto optimal payoff profiles may be ex-ante Pareto suboptimal. Second, while it assumes that agents agree to uniform randomization over the intersection of their sets, agents will in fact have differing preferences over different distributions with which to randomize, generating another bargaining problem. Less ad-hoc solutions might be found by using a dynamic bargaining protocol (as in the classical Rubinstein bargaining game Rubinstein 1982), which may narrow down the set of equilibria.

As a final note, compare this protocol with an analogous protocol for a pure coordination problem (Figure 4). In this protocol, agents play a default action but also announce a set of actions that they are willing to play. Again, actions are taken uniformly at random from the overap of the sets if it is nonempty, and otherwise the default action is taken. Clearly playing $\{A, B\}$ is a dominant strategy for each player, unlike the analogous strategy in the protocol for IBoS, where greater flexibility comes at the price of greater exploitability.

Algorithm 1: Welfare function selection protocol

Input: Sets of welfare functions $\mathcal{W}_i$, default policies π_i^d
`// Independently train policies to maximize welfare functions`
for $i = 1, 2$ **do**
  $\Pi_i \leftarrow \emptyset$
  for $w \in \mathcal{W}_i$ **do**
    $\pi_i^w \leftarrow \texttt{amTFT}(w)_i$
    $\Pi_i \leftarrow \Pi_i \cup \{\pi_i^w\}$
`// Communicate and decide what policies to deploy`
Players announce $\mathcal{W}_1, \mathcal{W}_2$
if $\mathcal{W}_1 \cap \mathcal{W}_2 \neq \emptyset$ **then**
  $w \sim \texttt{Unif}(\mathcal{W}_1 \cap \mathcal{W}_2)$
  $\pi_1, \pi_2 \leftarrow \pi_1^w, \pi_2^w$
else
  $\pi_1, \pi_2 \leftarrow \pi_1^d, \pi_2^d$
return π_1, π_2

Table 3: Empirical payoff matrix for agents playing IBoS and following bargaining protocol in Algorithm 1 with welfare functions in $\{w^{\text{Util}}, w^{\text{IA}}\}$.

	$\{w^{\text{Util}}\}, \pi_2^{\text{IA}}$	$\{w^{\text{IA}}\}, \pi_2^{\text{IA}}$	$\{w^{\text{Util}}, w^{\text{IA}}\}, \pi_2^{\text{IA}}$
$\{w^{\text{Util}}\}, \pi_1^{\text{Util}}$	4.0, 1.0	0.0, 0.0	4.0, 1.0
$\{w^{\text{IA}}\}, \pi_1^{\text{Util}}$	0.0, 0.0	2.0, 2.0	2.0, 2.0
$\{w^{\text{Util}}, w^{\text{IA}}\}, \pi_1^{\text{Util}}$	4.0, 1.0	2.0, 2.0	3.0, 1.5

Table 4: Payoff matrix for agents playing Pure Coordination and following a variant of the bargaining protocol in 1 where sets of rather than welfare functions are announced.

	$\{A\}, B$	$\{B\}, B$	$\{A, B\}, B$
$\{A\}, A$	1.0, 1.0	0.0, 0.0	1.0, 1.0
$\{B\}, A$	0.0, 0.0	1.0, 1.0	1.0, 1.0
$\{A, B\}, A$	1.0, 1.0	1.0, 1.0	1.0, 1.0

References

Ken Binmore, Ariel Rubinstein, and Asher Wolinsky. The nash bargaining solution in economic modelling. *The RAND Journal of Economics*, pages 176–188, 1986.

Ken G Binmore. The evolution of fairness norms. *Rationality and Society*, 10(3):275–301, 1998.

Craig Boutilier. Sequential optimality and coordination in multiagent systems. In *IJCAI*, volume 99, pages 478–485, 1999.

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. Emergent communication through negotiation. *arXiv preprint arXiv:1804.03980*, 2018.

Bhaskar Chakravorti and John Conley. Bargaining efficiency and the repeated prisoner's dilemma. *Economics Bulletin*, 3(3):1–8, 2004.

Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z Leibo. Learning reciprocity in complex sequential social dilemmas. *arXiv preprint arXiv:1903.08082*, 2019.

Dan S. Felsenthal and Abraham Diskin. The Bargaining Problem Revisited: Minimum Utility Point, Restricted Monotonicity Axiom, and the Mean as an Estimate of Expected Utility. *Journal of Conflict Resolution*, 26(4):664–691, December 1982. ISSN 0022-0027. doi: 10.1177/0022002782026004005. URL `https://doi.org/10.1177/0022002782026004005`. Publisher: SAGE Publications Inc.

Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

James W. Friedman. A Non-cooperative Equilibrium for Supergames12. *The Review of Economic Studies*, 38(1):1–12, 1971. ISSN 0034-6527. doi: 10.2307/2296617. URL `https://doi.org/10.2307/2296617`. _eprint: https://academic.oup.com/restud/article-pdf/38/1/1/4362169/38-1-1.pdf.

John C Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4):309–321, 1955.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. " other-play" for zero-shot coordination. *arXiv preprint arXiv:2003.02979*, 2020.

Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. Inequity aversion improves

cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*, pages 3326–3336, 2018.

Ehud Kalai. Proportional solutions to bargaining situations: interpersonal utility comparisons. *Econometrica: Journal of the Econometric Society*, pages 1623–1630, 1977.

Ehud Kalai and Smorodinsky. Other solutions to nash's bargaining problem. *Econometrica*, 43(3):513–518, 1975.

Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

Adam Lerer and Alexander Peysakhovich. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068*, 2017.

Michael W Macy and Andreas Flache. Learning dynamics in social dilemmas. *Proceedings of the National Academy of Sciences*, 99(suppl 3):7229–7236, 2002.

George J Mailath, J George, Larry Samuelson, et al. *Repeated games and reputations: long-run relationships*. Oxford university press, 2006.

Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. 2012.

Richard H McAdams. Beyond the prisoners' dilemma: Coordination, game theory, and law. *S. Cal. L. Rev.*, 82:209, 2008.

John F Nash. The bargaining problem. *Econometrica: Journal of the Econometric Society*, pages 155–162, 1950.

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

Alexander Peysakhovich and Adam Lerer. Consequentialist conditional cooperation in social dilemmas with imperfect information. *arXiv preprint arXiv:1710.06975*, 2017a.

Alexander Peysakhovich and Adam Lerer. Prosocial learning agents solve generalized stag hunts better than selfish ones. *arXiv preprint arXiv:1709.02865*, 2017b.

Ariel Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica: Journal of the Econometric Society*, pages 97–109, 1982.

Thomas C Schelling. An essay on bargaining. *The American Economic Review*, 46(3):281–306, 1956.

Thomas C Schelling. The strategy of conflict. prospectus for a reorientation of game theory. *Journal of Conflict Resolution*, 2(3):203–264, 1958.

Duncan Snidal. Coordination versus prisoners' dilemma: Implications for international cooperation and regimes. *American Political Science Review*, 79(4):923–942, 1985.

Peter Stone and Manuela Veloso. Task decomposition, dynamic role assignment, and low-bandwidth communication for real-time strategic teamwork. *Artificial Intelligence*, 110(2):241–273, 1999.

Weixun Wang, Jianye Hao, Yixi Wang, and Matthew Taylor. Towards cooperation in sequential prisoner's dilemmas: a deep multiagent reinforcement learning approach. *arXiv preprint arXiv:1803.00162*, 2018.

Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Efficient communication in multi-agent reinforcement learning via variance based control. In *Advances in Neural Information Processing Systems*, pages 3235–3244, 2019.

Lin Zhou. The nash bargaining theory with non-convex problems. *Econometrica: Journal of the Econometric Society*, pages 681–685, 1997.