The Importance of Credo in Multiagent Learning

David Radke
dtradke@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Kate Larson
kate.larson@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Tim Brecht
brecht@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

ABSTRACT

We propose a model for multi-objective optimization, a *credo*, for agents in a system that are configured into multiple groups (i.e., teams). Our model of credo regulates how agents optimize their behavior for the groups they belong to. We evaluate credo in the context of challenging social dilemmas with reinforcement learning agents. Our results indicate that the interests of teammates, or the entire system, are not required to be fully aligned for achieving globally beneficial outcomes. We identify two scenarios without full common interest that achieve high equality and significantly higher mean population rewards compared to when the interests of all agents are aligned.

KEYWORDS

Reinforcement Learning; Multiagent Learning; Coordination

1 INTRODUCTION

Humans have evolved the inherent ability to cooperate and organize into teams. Some hypothesize that this has significantly supported our path of achieving higher intelligence [45, 57]. Organizational Psychology and Biology have studied the structures and behavioral conditions under which people tend to be most efficient [55]. People tend to organize themselves into "teams-of-teams" within a larger system that are not in zero-sum competition, improving self identification and clarity of goals within a smaller group [34]. Today, these teams are present at various levels of complexity in order to survive, compete in sports, or complete tasks.

Wayne Gretzky, a former ice hockey player known as *The Great One*, describes a successful team as requiring "each and every [player] helping each other and pulling in the same direction". This statement, however, raises a number of questions. Do players on successful teams only optimize for the goals of their team? Is this strategy the best way for teams to achieve success? If not, under which conditions does optimizing for an alternative goal help or hinder team success, and can incentives for individual goals actually promote behavior that is beneficial to the team? In this work, we analyze how the performance of teams are impacted when agents may have different preferences they optimize their behavior for.

Multiagent Reinforcement Learning (MARL) has achieved impressive results in competitive two-team zero-sum settings such as capture the flag [28], hide-and-seek [8], and Robot Soccer [30], where teammates optimize for the goals of their team. The recent call to make cooperation central to the development of AI places emphasis on understanding the mechanisms behind teamwork beyond just competition [14, 15] and to adapt findings from Organizational Psychology [5]. In MARL, agents learning to cooperate often build common interest by sharing exogenous rewards [1, 7]; however, purely pro-social agents may not be possible when considering agents designed by different manufacturers or hybrid AI/human populations. Agents in these settings may have some self-interest for personal goals; therefore, it is important to understand how and when cooperation can be supported in systems where agents may optimize for multiple objectives.

We introduce agent *credo*, a model which regulates how agents optimize for multiple objectives in the presence of teams. The noun credo, defined as "the beliefs or aims which guide someone's actions" [53], fits appropriately to describe our model of how agents aim to optimize for goals. We analyze credo in mixed-motive social dilemmas popular in recent MARL research on cooperation [31, 59]. We discover multiple situations in which, despite some selfish preferences, certain credo configurations significantly outperform populations where the interests of all agents are aligned. This work makes the following contributions:

- We formally define a model of credo for multiagent teams.
- We study how the incentive structures of social dilemmas depend on agents' credo and environmental variables.
- With learning agents, we empirically evaluate the impact of credo and find that a population can achieve over 30% higher rewards if agents partially optimize for personal or fragmentary goals instead of goals shared by the entire population.

2 RELATED WORK

Human and Animal Teamwork: The notion of separate entities working together to form a *team* is observed across various organisms and levels of complexity. Animals such as ants, hyenas, and whales have also been shown to display similar teamwork characteristics which increase their outputs beyond any individual [3, 4]. Humans have been shown to evolve with an inherent bias towards teamwork and cooperation [25, 58]. This ability to cooperate is hypothesized to have significantly contributed to our development of our intelligence and ability to learn complex skills [47, 57]. Humans have developed systems of "teams-of-teams" often observed in industry, military, and sports teams that must be understood by examining between-team behavior [19].

Multiteam Systems (MTSs) [34, 63], a sub-field of Organizational Psychology, specializes in understanding characteristics of these systems which lead to a larger well-functioning organism [17, 32]. MTSs research relies on human user studies which often leads to conflicting results regarding the importance of component team

boundary status, connections between teams, goal type, and within-team alignment [64]. A common theme across many research areas in MTSs is the idea of social identification, or how agents perceive their goals [42, 55]. Team members may need to balance tendencies for their own personal goals with the goals of their team or the entire system [11, 61]. Our model of credo formalizes how RL agents manage tension between different goals, partly inspired by the population structure and partially shared objectives of MTSs.

Teamwork in AI: In AI, the concept of multiple non-conflicting teams that compose a larger system has been primarily explored for task completion [23, 56]. This work built on years of research studying ideas of within-team coordination, such as how agents share plans [24], beliefs [40, 41], and intentions [23]. Tambe [56] defines a model for a system of teams that can work together towards a common goal; however, that work lacked the ability to generalize. These early models assume teammates have full common interest.

Recent work on teams in AI has focused on ad hoc teamwork [20, 33, 36, 54, 60], teams in competition [48, 49] or coordination problems [29]. For teams in competition, building common interest among teammates often translates to fully sharing the rewards of the team [8, 9]. In social dilemmas, reward sharing is often used to align the interests of two [1], a subset [7, 43], or an entire population of agents [35]. Reward *gifting* allows agents to incentivize others, though manifests extra reward and violates budget-balance [26, 62]. These approaches have become standard in mixed-motive AI research to provide agents both an extrinsic reward from their environment and intrinsic reward from their peers [27, 29, 39, 60]. However, recent work for single group task completion has found their best results when agents do not only optimize for their group's preference, but also consider their individual preferences [20].

We expand the study of multiagent team structures and multi-goal optimization to define credo, a model of how agents optimize for multiple objectives. The model of teams we implement in our analysis is inspired by early work in both AI and MTSs. We define credo to regulate the amount of utility common interest among teammates and between teams. We base our analysis around similar concepts to MTSs using learning agents, such as the impact of interaction types, where cooperation happens, and under which conditions agents learn to cooperate.

3 MODEL OF CREDO WITH TEAMS

We model our base environment as a stochastic game $\mathcal{G} = \langle N, S, \{A\}_{i \in N}, \{R\}_{i \in N}, P, \gamma, \Sigma \rangle$. N is our set of all agents that learn online from experience and S is the state space, observable by all agents, where s_i is a state observed by agent i. $A = A_1 \times \ldots \times A_N$ is the joint action space for all agents where A_i is the action space of agent i. $R = R_1 \times \ldots \times R_N$ is the joint reward space for all agents where R_i is the reward function of agent i defined as $R_i : S \times A \times S \mapsto \mathbb{R}$, a real-numbered reward for taking an action in an initial state and resulting in the next state. $P : S \times A \mapsto \Delta(S)$ represents the transition function which maps a state and joint action into a next state with some probability and γ represents the discount factor so that $0 \leq \gamma < 1$. Σ represents the policy space of all agents, and the policy of agent i is represented as $\pi_i : S \mapsto A_i$ which specifies an action that the agent should take in an observed state.[1]

[1]We can also allow for randomized policies.

We introduce agent credo, a model to regulate how much an agent optimizes for different reward components it has access to. In this paper, we use "common interest" in the utility sense, referring to when agents share exogenous rewards instead of sharing views. A *team* is a subset of agents which have some degree of common interest for team-level goals. Given a population, multiple teams with different preferences and interests may co-exist that are not in zero-sum competition. The collection of all teams is referred to as a team *structure*. We denote the set of all teams as $\mathcal{T}$, the teams agent i belongs to as $\mathcal{T}_i$, and a specific team as $T_i \in \mathcal{T}_i$.

Our goal is to relax the modelling assumption that teammates are bound through full common interest [8, 13, 28, 43] to study how different credos impact a system of learning agents. For example, an agent may optimize their policy for the performance of one or multiple teams, while also being somewhat oriented towards its own personal goals. We represent these guiding principles by decomposing the reward any agent i may receive from the environment into three components: their individual exogenous reward $IR_i = R_i$, the rewards i receives from each team they belong to $TR_i^{T_i} \forall T_i \in \mathcal{T}_i$, and the reward i receives from the system of N agents SR_i. $TR_i^{T_i}$ and SR_i can be implemented with any function to aggregate and distribute rewards amongst multiple agents.

We define credo to be a vector of *parameters*, $\mathbf{cr}_i$, where the sum of all parameters is 1. The credo of an agent is represented $\mathbf{cr}_i = \langle \psi_i, \phi_i^{T_1}, \ldots, \phi_i^{T_{|\mathcal{T}|}}, \omega_i \rangle$, where ψ is the credo parameter for i's individual reward IR_i, $\phi_i^{T_i}$ is the credo parameter for the reward $TR_i^{T_i}$ from team $T_i \in \mathcal{T}_i$, and ω_i is the credo parameter for the reward i receives from the system SR_i. The parameter notation is organized by increasing order of group size, so that $\mathbf{cr}_i = \langle \text{self}, \ldots, \text{teams}, \ldots, \text{system} \rangle$, where $|\text{self}| < |\text{teams}| \leq |\text{system}|$. Agent i's credo-based reward function $R_i^{\mathbf{cr}}$ is calculated as:

$$R_i^{\mathbf{cr}} = \psi_i IR_i + \sum_{T_i \in \mathcal{T}_i} \phi_i^{T_i} TR_i^{T_i} + \omega_i SR_i, \quad (1)$$

The environment in our analysis consists of a stochastic game with a model of team structure $\langle \mathcal{G}, \mathcal{T} \rangle$. A particular case is where agents belong to exactly one team, which we study in the rest of this paper. Formally, $\mathcal{T}$ is a partition of the population into disjoint teams, $\mathcal{T} = \{T_i | T_i \subseteq N, \cup T = N, T_i \cap T_j = \emptyset \forall i, j\}$. This team structure simplifies the credo vector for each agent to be $\mathbf{cr}_i = \langle \psi_i, \phi_i, \omega_i \rangle$ where ϕ_i is the credo parameter for i's team.

Any function can be used to calculate IR_i, $TR_i^{T_i}$, or SR_i in our model. We implement functions to be consistent with building common interest between learning agents, similar to past work [8, 28, 60]. IR_i is defined to be the agent's normal exogenous reward R_i. Their team reward is defined as $TR_i^{T_i} : S \times A_i \times S \mapsto \mathbb{R}$, so that:

$$TR_i^{T_i} = \frac{\sum_{j \in T_i} R_j(S, A_j, S)}{|T_i|},$$

where teammates share their rewards equally. The system-wide reward is defined as $SR_i : S \times A_i \times S \mapsto \mathbb{R}$ so that:

$$SR_i = \frac{\sum_{j \in N} R_j(S, A_j, S)}{|N|},$$

the mean reward of all N agents. The final credo-based reward for agent i, R_i^{cr}, is calculated using Equation 1 and these functions.

As is standard in many MARL problems, agents are trained to independently maximize their rewards. In particular, at time t each agent i selects some action a_i which together form a joint action a^t. This action results in a transition from state s^t to state s^{t+1}, according to the transition function P, and provides each agent i with reward $R_{i,t}(s^t, a^t, s^{t+1})$. Agents seek to maximize their sum of discounted future rewards, $V_i = \sum_{t=0}^{\infty} \gamma^t R_{i,t}$. Our model replaces R_i with R_i^{cr}, reconfiguring the learning problem so agents must learn behavior that maximizes their sum of discounted future credo-based rewards according to the team structure and environment.

4 SOCIAL DILEMMAS

Social dilemmas present tension between short-term individual incentives and the long-term collective interest, where agents must cooperate for higher rewards. All agents prefer the benefits of a cooperative outcome; however, the short-term benefits of self-interested behavior outweigh those of cooperative behavior, making social dilemmas an excellent testbed for credo. Social dilemmas and mechanisms for agents or people to overcome them has been widely studied across game theory [12, 50], economics [10, 38], psychology [18, 21], and more recently, AI [1, 16, 31]. For our analysis, we consider intertemporal social dilemmas with active provision, defined as when cooperation carries an explicit cost [27]. We implement our model of credo and teams in the Iterated Prisoner's Dilemma (IPD) [46] and Cleanup Markov game [59].

4.1 Iterated Prisoner's Dilemma (IPD)

In the one-shot Prisoner's Dilemma, two agents interact and each must decide to either cooperate (C) with or defect (D) on each other. We assume there is a cost (c) and a benefit (b) to cooperating where $b > c > 0$ with payoffs shown in Table 1 in the appendix. If an agent cooperates, it incurs the cost c. If both agents cooperate, they both benefit, each receiving a reward of $b - c$. If one agent cooperates but the other defects, then we assume that the cooperating agent incurs the cost c, but the defecting agent reaps the benefit b (e.g., by stealing the contribution of the cooperator). If neither cooperate, neither benefit nor incur a cost, leading to a reward of zero for both. The unique Nash Equilibrium is obtained when both agents defect, since if one agent cooperates, the other agent is strictly better off defecting and receiving b, instead of $b - c$.

The degree to which agents optimize for different groups may change how these interactions are viewed. The traditional assumption is that each agent acts individually and is solely focused on their own individual reward. However, if agents have certain credos, they may optimize for entire groups they belong to instead of themselves which may have a considerable impact on how agents learn. We can view the standard Prisoners' Dilemma in Table 1 as an instance of the game where the credo of both agents is fully self-focused ($\text{cr}_i = \langle 1, 0, 0 \rangle$). If both agents were entirely system-focused ($\text{cr}_i = \langle 0, 0, 1 \rangle$), or on the same team and team-focused ($\text{cr}_i = \langle 0, 1, 0 \rangle$), agents would wish to take actions which maximize their shared utility, resulting in mutual cooperation.

In the IPD, this game is played repeatedly which adds a temporal component and allows agents to learn a policy over time. Instead

of just two agents, we work with a population of agents that are divided into teams *a priori*. At each timestep, agents are randomly paired with another agent, a *counterpart*, that may or may not be a teammate. Agents are informed as to what team their counterpart belongs to through a numerical signal s_i, though additional identity information is not shared. Agents must decide to cooperate with or defect on this counterpart for a single round of the Prisoner's Dilemma. Their payoff for the interaction is their reward defined by their credo, R_i^{cr}, based on their own and other agents' interactions. Agents update their strategies (i.e., learn) using their direct observation s_i, what action they chose a_i, and their credo reward R_i^{cr}. Since only the team information of the counterpart is shared, the strategies of all agents on team T_i ultimately affects how agents learn to play any member of T_i. Further implementation details can be found in Appendix A.1.

4.1.1 Equilibrium Analysis with Credo. We are interested in understanding the conditions under which credo may help or hinder cooperation. Thus, as a first step we investigate the impact of credo on the *stage game* of the IPD with teams. To provide a clear comparison with the standard IPD, we take an ex-ante approach, where agents are aware of their imminent interaction and the existence of other teams, but not the actual team membership of their counterpart. Refer to Appendix B for further details and the complete derivation of our equilibrium analysis.

Assume a pair of agents, i, j, have been selected to interact at some iteration of the IPD and agent i knows j will be a teammate with probability v and a non-teammate with probability $(1 - v)$. Let $\sigma_{T_i} = (\sigma_{ji}, 1 - \sigma_{ji})$ represent j's strategy profile when $j \in T_i$, where σ_{ji} is the probability for cooperation (C). Likewise, let $\sigma_{T_j} = (\sigma_{jj}, 1 - \sigma_{jj})$ be j's strategy profile when $j \in T_j$, any other team.

For the sake of our analysis, we make the assumption that all agents have the same credo. We calculate the expected values of cooperation and defection in situations where agents are fully self-focused ($\text{cr}_i = \langle 1, 0, 0 \rangle$), team-focused ($\text{cr}_i = \langle 0, 1, 0 \rangle$), and system-focused ($\text{cr}_i = \langle 0, 0, 1 \rangle$). We then calculate the conditions in which agent i has the incentive to cooperate as when the expected value of cooperation dependant on credo is greater than the expected value of defection. After algebraic simplification, we determine agent i is better off cooperating whenever:

$$\phi_i \left(v - \frac{2c}{b+c} \right) + \omega_i \left(\frac{b-c}{2} \right) \geq \psi_i c. \tag{2}$$

Note that this is independent of the strategy profile of their counterpart, σ_T. Whenever cooperation is the dominant strategy in a stage game, it will be supported in the repeated game.

Figure 1 shows the expected reward value of cooperation minus the expected reward value of defection, solving Equation 2 for $|\mathcal{T}| = 5$ teams. Each triangle shows the results for the linear combination of agent credo composed of self- (ψ; right axis), team- (ϕ; left axis), and system-focused (ω; bottom axis) parameters (increments of 0.02). The colors indicate which action has higher expected value, defection (blue) or cooperation (red), computed as the difference of their expected values (white represents equality).

Each row of plots represents different values of v, the likelihood of being paired with a teammate. The remaining probability $1 - v$ is spread across the $|\mathcal{T}| - 1$ teams uniformly. With five teams,
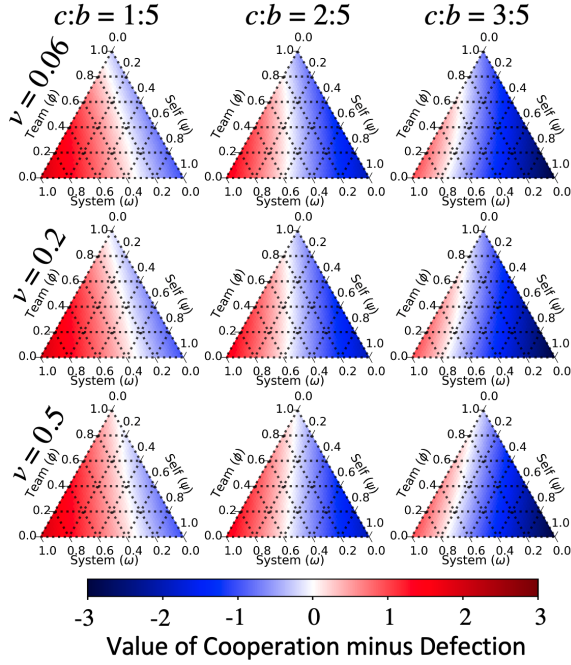
Figure 1: Impact of teammate pairing probability ν and the cost of cooperation c (benefit $b = 5$) on action incentives with credo. Red corresponds with cooperation being incentivized and blue corresponds with defection.

these values of ν represent when the chance of a counterpart being from another team is four times more likely than their own team ($\nu = 0.06$), being from any of the five teams has equal probability ($\nu = 0.2$), and being from the same team is four times more likely than another team ($\nu = 0.5$). Each column of plots represents a different cost of cooperation so that $c \in \{1, 2, 3\}$ with the benefit fixed to $b = 5$. For our entire analysis, we increase the cost and fix the benefit since we are interested in the ratio between the cost and benefit of cooperation instead of their absolute values.

We observe less overall incentive to cooperate as the cost c increases (i.e., darker blue and has more area inside the triangles). This pattern resembles findings observed in human behavior, where the amount of cooperation depends on the size of the benefit compared to the cost [51]. Another observation is that defection is incentivized in the presence of any amount of self-focus (right axes), with the exception of one environment ($c = 1$ and $\nu = 0.5$). Even in this scenario, defection becomes quickly incentivized as self-focus increases to $\psi_i = 0.2$. The experiments in Sections 5.1 show that learning agents are able to develop globally beneficial cooperative behavior in multiple settings where defection is incentivized.

4.2 Cleanup Markov Game

The Cleanup domain [59] is a temporally and spatially extended Markov game representing a social dilemma. Cleanup allows us to examine if our findings generalize to more complex environments since agents must learn a cooperative policy through movement and decision actions instead of choosing an explicit *cooperation*

action like in the IPD. Active provision is represented in Cleanup by agents choosing actions with no associated environmental reward that are necessary before any agent achieves a reward.

Appendix A.2 contains a screenshot of the Cleanup environment with $|\mathcal{T}| = 3$ teams. At each timestep, agents choose among nine actions: 5 movement (up, down, left, right, or stay), 2 turning (left or right), and a cleaning or punishing beam. One half of the map contains an aquifer (or river) and the other an apple orchard. Waste accumulates in the river with some probability at each timestep which must be cleaned by agents. Apples spawn in the orchard with a function that accounts for the cleanliness of the river. Agents receive a reward of +1 for consuming an apple by moving on top of them. The dilemma exists in agents needing to spend time cleaning the river (to spawn new apples) and receiving no exogenous reward versus just staying in the orchard and enjoying the fruits of another's labor. Agents have the incentive to only pick apples; however, if all agents attempt policy, no apples grow and no agents get any reward. A successful society will balance the temptation to free-ride with the public obligation to clean the river.

5 EMPIRICAL EVALUATION

In the rest of this paper, we present the setup and results of experiments in both environments using learning agents. While our model does not require it, we assume that for all teams $T_i, T_j \in \mathcal{T}$, $|T_i| = |T_j|$ (i.e., given a team model, the teams are the same size). This avoids complications that might arise with agent interactions if teams were of significantly different sizes and to be consistent across domains, leaving other interactions and different sized teams for future work. We initialize $\mathbf{cr}_i$ to be the same for all agents a *a priori* and do not change $\mathbf{cr}_i$ within an experiment. Since fully self-focused and system-focused credos have agents working as individuals (i.e., non-team setting) and a full group (i.e., cooperative setting), they serve as benchmarks against which we can compare the performance of other credos with teams.

5.1 IPD Evaluation

In the IPD, each experiment lasts 1.0×10^6 episodes. We configure $N = 25$ Deep Q-Learning [37] (DQN) agents into $|\mathcal{T}| = 5$ teams of equal size. While our team model allows for an arbitrary number of teams of any size, this work is concerned with the relationship between agent credo and environmental conditions.

An episode is defined by a set of agent interactions where each agent is paired with another agent and plays an instance of the Prisoner's Dilemma. Agent pairings are assigned based on ν, the probability of being paired with a teammate and agents are unable to explicitly modify who they interact with, a challenging scenario for cooperation without additional infrastructure [2]. Each experiment is repeated five times with different random seeds. Further implementation details are provided in Appendix A.1.

5.1.1 IPD: Full-Focus Credo. We start by analyzing how the behavior of agents is impacted by the extreme cases of full self-focus ($\mathbf{cr}_i = \langle 1, 0, 0 \rangle$), team-focus ($\mathbf{cr}_i = \langle 0, 1, 0 \rangle$), or system-focus ($\mathbf{cr}_i = \langle 0, 0, 1 \rangle$) credo. Figure 2 shows our results where the x-axis of each plot shows the time and the y-axis shows the percent of actions where agents chose to cooperate, averaged over 2,000 episode windows. We set $c = 1$, $b = 5$, and $\nu = 0.2$ so counterparts have
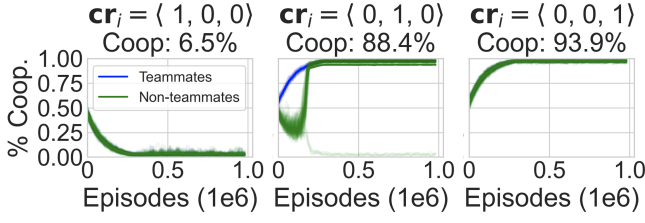
Figure 2: *IPD:* **Full self-, team-, and system-focused agents when** $c = 1$, $b = 5$, $v = 0.2$, **and** $|\mathcal{T}| = 5$ **of five agents each.**

equal probability of being selected from any team (since $|\mathcal{T}| = 5$). Each line represents the behavior of one agent when paired with a teammate (blue) or a non-teammate (green).

When all agents are fully self-focused (left), $\mathbf{cr}_i = \langle 1, 0, 0 \rangle \forall i \in N$, they immediately learn defection towards all other agents (blue overlapped by green). When agents are team-focused (middle), $\mathbf{cr}_i = \langle 0, 1, 0 \rangle$, defection is incentivized as shown in Figure 1. However, we find agents quickly identify and cooperate with their teammates and almost every agent simultaneously develops stable prosocial policies with non-teammates despite not sharing rewards. We hypothesize this is due to a combination of reduced reward variance for actions in specific states and interactions with teammates providing a strong positive feedback signal favoring cooperation. In the IPD, fully system-focused DQN agents will achieve the upper bound of performance since they maximize the collective rewards of all agents, equivalent to choosing cooperation. In the right plot when agents are fully system-focused, agents do learn to cooperate with every agent regardless of team (blue overlapped by green). While other work requires strong assumptions of behavior to steer agents towards cooperation [1], these results indicate that full common interest may not be required to promote cooperation across an entire population with teams.

5.1.2 IPD: Multi-Focus Credo. Next, we experiment with settings where agents can simultaneously partially optimize for their own, their team's, or the system's goals through their credo. We use the same environmental settings as Figure 1, so that $v \in \{0.06, 0.2, 0.5\}$, $c \in \{1, 2, 3\}$, and $b = 5$ to understand how credo and environmental parameters impact how agents learn. We evaluate the case where all agents have the same credo parameters, that is, $\mathbf{cr}_i = \mathbf{cr}_j \forall i, j \in N$.

Figure 3 shows our results for various combinations of credo with 0.2 step increments, teammate pairing probability v, and the cost of cooperation c. This creates nine different environments, each with 21 combinations of credo represented by the intersections of dotted lines from the three axes of each triangle of Figure 3. In the IPD, mutual cooperation yields the highest mean population reward. Thus, the hexagonal area around each intersection point is colored according to the global mean percent of actions which were to cooperate from blue (less cooperation) to red (more cooperation). RL agents are able to condition their policy on the information of their counterpart's team, allowing us to observe how they learn behavior towards different groups. For each environment, we plot the total cooperation, cooperation with teammates (In-Team), and cooperation with non-teammates (Out-Team).

Agents achieve high cooperation when they have full system-focus (left corners) and learn defection as agents become self-focused (right corners). Despite the incentive to defect in eight of nine environments, fully team-focused agents learn cooperation in five environments, similar to the behavior in Figure 2. This cooperation is robust if full team-focus can not be achieved, such as when self-focus $\psi = 0.2$ and $c = 1$. In these settings, the rate of cooperation is higher when agents have high team-focus compared to high system-focus, though decreases as when $c = 3$. Unlike previous implementations of teams which assume agents have full common interest [8, 13, 28], our results show teammates are not required to be fully aligned to achieve good results.

This shows that teams of highly team-focused agents have the ability to support global cooperation despite incentives to defect or some self-focus, relaxing the assumption that teammates are bound together with full common interest. Contrary to Gretzky's belief in Section 1, our results indicate teams still achieve good results despite some self-focus. To understand this significance, consider situations where pure-common interest among teammates may not be guaranteed or all agents are unable to be controlled. Shown next, results in the Cleanup domain actually improve beyond full system-focus with certain credo parameters.

5.2 Cleanup Evaluation

Using the Cleanup evaluation domain, we experiment with $N = 6$ agents learning with Proximal Policy Optimization (PPO) [52] split into $|\mathcal{T}| = 3$ teams of two agents each. Past work that has used the Cleanup domain typically uses $N = 5$ agents for a time of 1.6×10^8 environment steps (each episode is 1,000) [27, 60]. We increase the population to $N = 6$ agents to allow for three equal sized teams and calculate metrics over the last 25% of timesteps, similar to the IPD evaluation. Agent observability is limited to a 15×15 RGB window centered on the agent's current location. Teammates appear as the same color and optimize their own $R_i^{\mathbf{cr}}$ after each timestep. Each experiment is completed eight times with different random seeds. Further experiment details are found in Appendix A.2.

5.2.1 Cleanup: Reward. While success in the IPD relies on agents choosing to cooperate in direct interactions, success in Cleanup requires agents to coordinate and form an effective joint policy to clean the river enough for apples to grow. Since agents must learn their own coordinated policy by taking actions in states with higher complexity, we can no longer assume the full-common interest scenario (full system-focus) will achieve the upper bound on performance. Figure 4 shows the mean credo-based population reward per-episode $\overline{R^{\mathbf{cr}}}$ for all 21 credo configurations in Cleanup, removing the subscript i when referencing all agents. The mean population reward gives insight into how well agents learn to solve the dilemma. Each hexagon corresponds with a combination of credo, centered at the intersections of three dotted lines from each axis (self, team, and system). Hexagons are colored according to the mean population reward from low (white) to high (red).

Fully self-focused agents fail to solve the dilemma, receiving the lowest mean population reward of any scenario. Previous work has found that the highest rewards in Cleanup are obtained when agents optimize for reward signals from all agents (i.e., system-focus) [22, 35, 60]. However, we find that some self- or team-focus
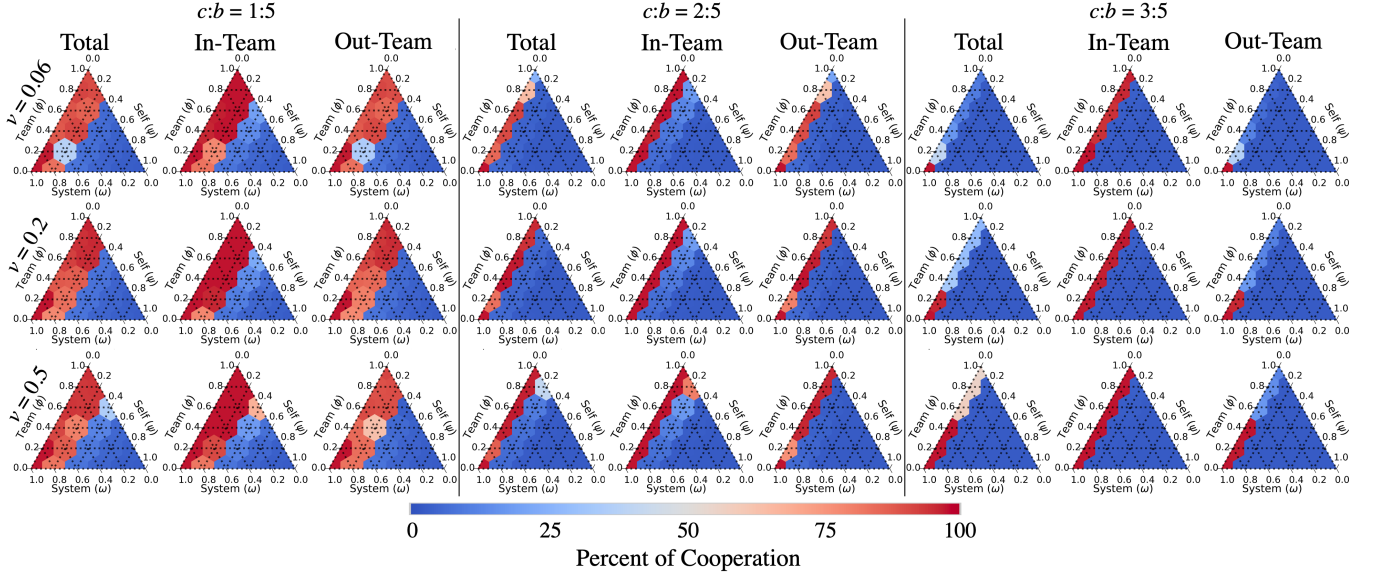
Figure 3: *IPD:* Percent of total cooperation, cooperation with teammates (In-Team), and cooperation with non-teammates (Out-Team) when all agents follow the same credo. We experiment with different probabilities of being paired with a teammate $\nu \in \{0.06, 0.2, 0.5\}$, costs of cooperation $c \in \{1, 2, 3\}$, and fix the benefit $b = 5$.
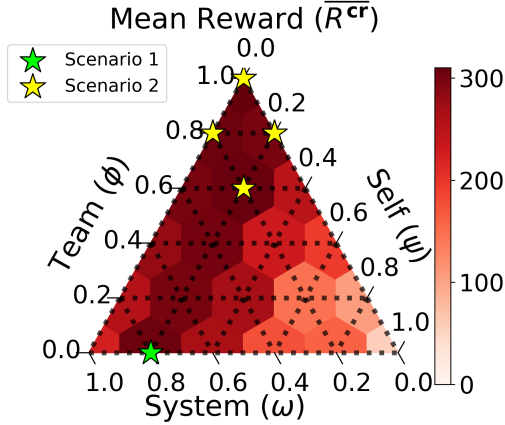

Figure 4: *Cleanup:* Mean population reward for every credo in our evaluation. These experiments have $|\mathcal{T}| = 3$ teams of two agents each. The scenarios with the highest reward often have agents with slight self-focus. We identify two types of credo scenarios that achieve the highest reward, when credo has slight self-focus paired with high system-focus (green star) and when team-focus is high (yellow stars).

improves the mean reward significantly over the system-focused setting. We divide the five highest-reward environments into two scenarios shown in Figure 4. First, when agents with high system-focus also have slight self-focus (Figure 4, green star), and second, when agents have high team-focus relative to their other credo parameters (Figure 4, yellow stars). These scenarios achieve at

least 30% higher mean population reward per-episode than the full common interest setting (system-focused; left corner).

Scenario 1: The first scenario we examine is when highly system-focused agents have slight self-focus, $\mathbf{cr}_i = \langle 0.2, 0.0, 0.8 \rangle$ (green star). Agents with this credo achieve 33% higher reward per-episode compared to a population with full common interest. This result is comparable to a similar finding in past work [20], where a cooperative group performs best when agents have some selfish preferences of how to complete a task. This suggests that, despite using an entirely different domain, some amount of self-focus may consistently lead to high global performance and is worthy of more exploration.

Scenario 2: With the introduction of teams-focus, we more closely examine another credo scenario that contains four of the top five experiments with high mean rewards. The yellow stars in Figure 4 show experiments when agents have high team-focus relative to their other credo parameters, specifically $\mathbf{cr}_i \in \{\langle 0.0, 1.0, 0.0 \rangle, \langle 0.2, 0.8, 0.0 \rangle, \langle 0.0, 0.8, 0.2 \rangle, \langle 0.2, 0.6, 0.2 \rangle\}$. As discussed in Scenario 1, agents with high system-focus experience a decrease in rewards when they are *too* system-focused. In Scenario 2, high team-focused agents achieve high rewards regardless if self-focus is zero or 0.2, echoing our result in the IPD that teammates are not required to have full common interest. These insights may be useful when attempting to influence credo in settings where agents are unable to guarantee their amount of self-focus or team commitment.

5.2.2 Cleanup: Division of Labor. We find that agents in the highest-reward experiments often learn to divide labor and specialize to either clean the river or pick apples. This ability to coordinate with other teams and fill roles significantly impacts the global reward. We observe the best division of labor strategy when two agents clean the river (i.e., cleaners) and four agents pick apples (i.e., pickers). In the following analysis, each line in Figures 5 and 6 represent
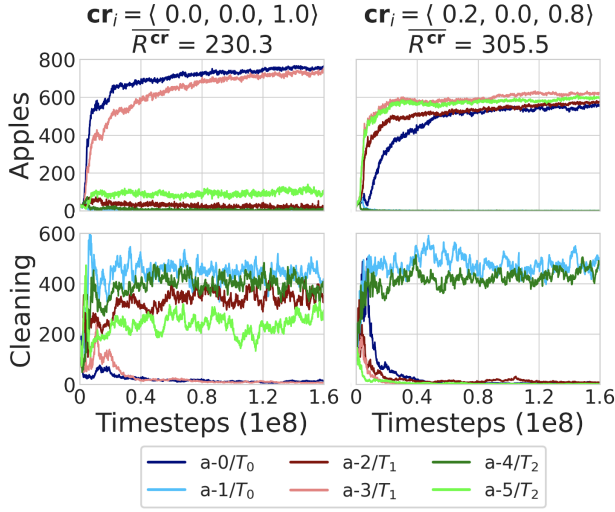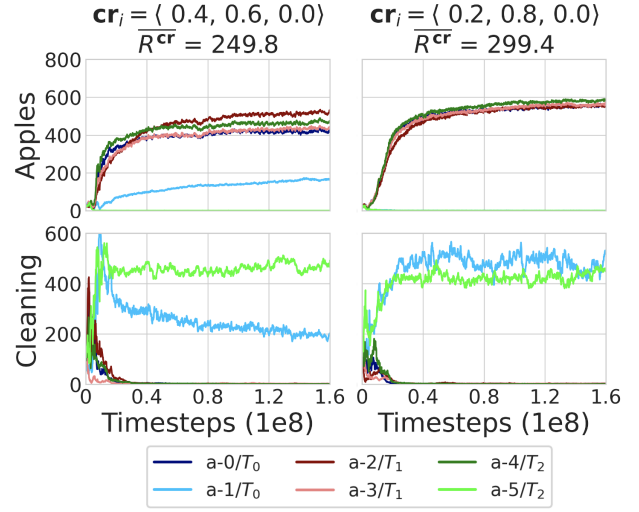
Figure 5: *Cleanup:* (Scenario 1) high system-focus with slight self-focus. Agents are labeled so that "a-0/T_0" represents agent #0 belonging to team #0. Each column of plots shows (left) fully system-focused agents and (right; green star in Figure 4) when agents become slightly self-focused. Better division of labor strategies are learned when self-focus increases from zero to 0.2 by enticing four agents to pick apples instead of just two, leading to 33% higher reward.

Figure 6: *Cleanup:* (Scenario 2) high team-focus achieves high rewards despite a some self-focus. Each column of plots shows when team-focus is 0.6 (left) and 0.8 (right, a yellow star in Figure 4), offset with self-focus. As team-focus increases, two teams coordinate so at least one teammate cleans the river, leading to better division of labor which is maintained even when agents become fully team-focused.

the behavior of a single agent. For example, a line labeled "a-0/T_0" represents agent #0 belonging to team #0. Teammates appear as different shades of the same color (T_0 blue, T_1 red, and T_2 green). In adjacent trials, agents with the same label (i.e., a-0/T_0) may learn different behavior, making aggregating policies from all eight trials difficult. Therefore, we present figures from one trial representing the most commonly learned behavior in each setting.

Scenario 1: We first analyze the division of labor in Scenario 1 (green star in Figure 4). Figure 5 shows the number of apples picked (top) and cleaning actions taken (bottom) by all six. The left plots show when agents are fully system-focused ($\mathbf{cr}_i = \langle 0.0, 0.0, 1.0 \rangle$) and the right plots shows when agents become slightly self-focused ($\mathbf{cr}_i = \langle 0.2, 0.0, 0.8 \rangle$). The full system-focused population evolves into four cleaning agents and two apple pickers, with each cleaner receiving rewards from both pickers regardless of team membership. This amount of reward suppresses any desire for cleaning agents to learn to pick apples, causing the population to reach a local minimum. The two apple pickers pick over 700 apples each resulting in a mean population reward of $R^{\mathbf{cr}} = 230.3$. However, increasing the self-focus to $\phi_i = 0.2$ (right plots) provides enough individual incentive to for four agents to pick apples and collect 600 apples each for $R^{\mathbf{cr}} = 305.5$. Due to high system-focus, the two cleaning agents receive enough reward from all four pickers to incentivize them to continue cleaning, and the entire system achieves 33% higher reward by escaping the previous local minimum.

Scenario 2: We now analyze when agents have high team-focus compared to their other credo parameters, indicated by yellow stars in Figure 4. Of the four experiments in this scenario, we choose $\mathbf{cr}_i = \langle 0.2, 0.8, 0.0 \rangle$ and compare with $\mathbf{cr}_i = \langle 0.4, 0.6, 0.0 \rangle$. Since self-

and team-focus change by one increment (of 0.2) and system-focus is zero, this is the simplest possible comparison.

The columns of Figure 6 represent when agents increase team-focus from 0.6 (left) to 0.8 (right), with the remaining credo being self-focus. When team-focus is 0.6 (left), only one team (T_2, green) learns to divide into the different roles of one river cleaner and one apple picker. While a-0 on T_0 (dark blue) fully learns to pick apples, their teammate (a-1) does not fully learn the role of river cleaner. This agent attempts to also pick apples and free ride on the cleaning of a-5. T_1 does not commit either agent to clean the river, resulting in fewer than two full river cleaners overall. This hinders population reward, since fewer than two total cleaners is unable to generate enough apples to support the remaining apple pickers. Thus, the four main apple pickers only collect an average of just over 400 apples each for a mean population reward of $R^{\mathbf{cr}} = 249.8$.

In the right column when agents have higher team-focus ($\mathbf{cr}_i = \langle 0.2, 0.8, 0.0 \rangle$, yellow star scenario), two teams learn to divide into one river cleaner and one apple picker, ensuring two agents are always cleaning. This produces enough apples for four pickers to collect about 600 each and both cleaners receive enough shared reward to overcome the incentive to free ride. As a result, the population earns $R^{\mathbf{cr}} = 299.4$, which is 20% more reward than when $\mathbf{cr}_i = \langle 0.4, 0.6, 0.0 \rangle$ (left) and 30% more reward than the full common interest setting $\mathbf{cr}_i = \langle 0.0, 0.0, 1.0 \rangle$ (Figure 5 left). This division of labor is consistently learned when team-focus is high (yellow stars).

Overall, our results show specific combinations of credo support globally beneficial behavior among a population of teams. We expand a result from [20] to social dilemmas showing some selfishness improves group performance. Furthermore, we identify that
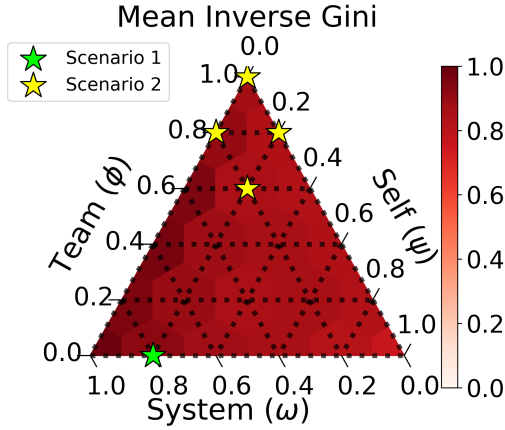
Figure 7: *Cleanup:* **Inverse Gini index for every credo in our evaluation. These experiments have** $|\mathcal{T}| = 3$ **teams of two agents each. Despite drastically different rewards, the credos which achieve high rewards also have high equality.**

agent specialization within their component teams results in high team-focus achieving more reward than fully system-focused credo.

5.2.3 Cleanup: Equality. McKee et al. [35] identify that analyzing an aggregated reward signal such as mean population reward may overlook settings with high reward inequality that may be detrimental to a system's stability. Since agents do not receive any exogenous reward for cleaning the river, it is important to consider the implications and potential side effects of credo and teams on population equality, or how evenly reward is distributed among a population of agents. We model population reward equality as the inverse Gini index, similar to past work [35]:

$$Equality = 1 - \frac{\sum_{i=0}^{N} \sum_{j=0}^{N} |R_i^{\mathbf{cr}} - R_j^{\mathbf{cr}}|}{2|N|^2 \overline{R^{\mathbf{cr}}}}, \qquad (3)$$

where values closer to 1 represent more equality. Figure 7 shows our results for equality (dark red is full equality). The full system-focused case, by definition, has perfect equality since all agents share rewards equally. Scenario 1 still has high system-focus and also achieves high equality. In Scenario 2, we observe the two agents that learn to clean the river always emerge from two teams learning to divide their labor. Since each team has at least one apple picker agent, and agents have high team-focus to share rewards with cleaners, the population maintains high equality. Both scenarios achieve more equality than fully self-focused agents $\mathbf{cr}_i = \langle 1.0, 0.0, 0.0 \rangle$ while obtaining significantly higher reward.

6 DISCUSSION AND FUTURE WORK

We proposed a model of credo that regulates how agents optimize their behavior for different groups they belong to (i.e., self, teams, or system). Our analysis serves as a proof of concept for exploring how agents simultaneously optimize for multiple objectives and learn to cooperate and coordinate. Our main contributions are two-fold. First, we show how agents are able to form cooperative policies that are robust to some amount of self-focus. Second, we find that agents

achieve high population reward with high team-focus compared to other credo parameters. Teams are also not required to have full common interest to achieve high reward, unlike previous definitions of teams [8, 13, 28]. This has significant implications in settings where the amount of agents' self-focus may be unknown, or full alignment of a group's interests may not be possible or desired.

A key takeaway of this work is that full cooperation in a system of individual learners may not achieve the highest reward, despite several recent studies using this scenario as a basis for comparison [22, 35, 60]. This is contrary to the spirit of Gretzky's observation mentioned in the introduction: a certain amount of personal striving is beneficial for the system overall. We observed that fully cooperative systems did not take full advantage of the efficiencies of labor division and task specialization, which did arise when agents had some self- or team-focus. This tended to be particularly problematic as the number of agents sharing rewards increased, indicating a correlation between team structure, credo, and learning role-specialization. Similar to a recent single-agent finding [6], we hypothesize this is due to increased complexity in credit assignment as the reward-sharing group increases in size.

It might be possible to mitigate credit-assignment issues by actively tuning credo to regulate the amount of reward shared with teammates or the overall system. If credo is tuned to emphasize self-focus, agents may gain a stronger direct feedback signal from their actions, guiding them towards better policies. On the other hand, fully self-focused agents may become more cooperative if their credo is adjusted to encompass teams or the system as a whole. In general, this may be viewed as a form of meta-learning where some credo-regulating policy shapes the environment in which agents learn. Tuning credo could be approached in two ways: centralized or decentralized. Using a centralized credo-tuner has the advantages of striving for global goals such as egalitarian or utilitarian ideas of equity, diversity, or productivity, though at a cost of potentially significant overhead. On the other hand, a fully decentralized credo-tuning model may make analysing systems and the resulting equilibria even more challenging.

While past MARL work without teams tries to prevent specialization [22, 35], the areas of Team Forming and Coalition Structure Generation often construct teams specifically with specialized agents to fill roles [5]. Successful human teams, such as in sports, are specifically composed of players with diverse abilities [44]. The introduction of teams should change how specialization is viewed in MARL settings, such as agents learning emergent roles from only their credo-based reward. If rewards can not be accessed, such as with hybrid AI/human teams, performance incentives may alter the inherent goals of agents. Furthermore, credo may help us understand which features of a system lead to inequality, create underlying incentives, or how to achieve alternative societal goals.

We see many interesting directions for this work. For example, analyzing situations when teams or teammates have different credo or if agents are able to dynamically tune or learn credo over time as mentioned above. Another possible direction for future work is studying how teammates could influence the credo of their teammates or other teams and which processes achieve the highest behavioural influence. We hope that this work inspires future directions studying multiagent teams, multi-objective optimization, and incentives to improve system performance.

ACKNOWLEDGMENTS

REFERENCES

[1] Nicolas Anastassacos, Julian Garcia, S. Hailes, and Mirco Musolesi. 2021. Cooperation and Reputation Dynamics with Reinforcement Learning. In *AAMAS*.

[2] Nicolas Anastassacos, S. Hailes, and Mirco Musolesi. 2020. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. In *AAAI*.

[3] C. Anderson and N. Franks. 2001. Teams in animal societies. *Behavioral Ecology* 12 (2001), 534–540.

[4] C. Anderson and N. Franks. 2003. Teamwork in Animals, Robots, and Humans. *Advances in The Study of Behavior* 33 (2003), 1–48.

[5] E. Andrejczuk, J. A. Rodriguez-Aguilar, and C. Sierra. 2016. A concise review on multiagent teams: contributions and research opportunities. *Multi-Agent Systems and Agreement Technologies* (2016).

[6] Dilip Arumugam, Peter Henderson, and Pierre-Luc Bacon. 2020. An information-theoretic perspective on credit assignment in reinforcement learning. *Workshop on Biological and Artificial Reinforcement Learning (NeurIPS 2020)* (2020).

[7] Bowen Baker. 2020. Emergent Reciprocity and Team Formation from Randomized Uncertain Social Preferences. *NeurIPS* (2020).

[8] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch. 2019. Emergent Tool Use From Multi-Agent Autocurricula. In *ICLR*.

[9] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Christopher Hesse, R. Józefowicz, Scott Gray, Catherine Olsson, Jakub W. Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, J. Schneider, S. Sidor, Ilya Sutskever, Jie Tang, F. Wolski, and Susan Zhang. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *ArXiv* abs/1912.06680 (2019).

[10] G. Brady. 1993. Governing the Commons: The Evolution of Institutions for Collective Action. *Southern Economic Journal* 60 (1993), 249–251.

[11] D. R. Carter, R. Asencio, Hayley M. Trainer, Leslie A. DeChurch, R. Kanfer, and S. Zaccaro. 2019. Best Practices for Researchers Working in Multiteam Systems.

[12] D. Challet and Yi-Cheng Zhang. 1997. Emergence of cooperation and organization in an evolutionary game. *Physica A-statistical Mechanics and Its Applications* 246 (1997), 407–418.

[13] Stephen Chung. 2021. MAP Propagation Algorithm: Faster Learning with a Team of Reinforcement Learning Agents. *Advances in Neural Information Processing Systems* 34 (2021).

[14] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground. *Nature* 593 (2021), 33–36.

[15] A. Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, K. Larson, and T. Graepel. 2020. Open Problems in Cooperative AI. *ArXiv* abs/2012.08630 (2020).

[16] Panayiotis Danassis, Zeki Doruk Erden, and B. Faltings. 2021. Improved Cooperation by Exploiting a Common Signal. In *AAMAS*.

[17] Robert Davison, J. Hollenbeck, Christopher M. Barnes, Dustin J. Sleesman, and D. R. Ilgen. 2012. Coordinated action in multiteam systems. *The Journal of applied psychology* 97 4 (2012), 808–24.

[18] Robyn M. Dawes and David M. Messick. 2000. Social Dilemmas. *International Journal of Psychology* 35, 2 (2000), 111–116.

[19] Leslie A. DeChurch and S. Zaccaro. 2010. Perspectives: Teams Won't Solve This Problem. *Human Factors: The Journal of Human Factors and Ergonomics Society* 52 (2010), 329 – 334.

[20] Ishan Durugkar, E. Liebman, and P. Stone. 2020. Balancing Individual Preferences and Shared Objectives in Multiagent Reinforcement Learning. In *IJCAI*.

[21] E. Fehr and Ivo Schurtenberger. 2018. Normative foundations of human cooperation. *Nature Human Behaviour* 2 (2018), 458–468.

[22] Ian Gemp, Kevin R McKee, Richard Everett, Edgar A Duéñez-Guzmán, Yoram Bachrach, David Balduzzi, and Andrea Tacchetti. 2022. D3C: Reducing the Price of Anarchy in Multi-Agent Learning. *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems* (2022).

[23] B. Grosz and S. Kraus. 1996. Collaborative Plans for Complex Group Action. *Artif. Intell.* 86 (1996), 269–357.

[24] B. J. Grosz and C. L. Sidner. 1990. Plans for Discourse. In *Intentions in Communication*, P. R. Cohen, J. Morgan, and M. E. Pollack (Eds.). MIT Press, Cambridge, MA, 417–444.

[25] E. Herrmann, J. Call, M. V. Hernández-Lloreda, B. Hare, and M. Tomasello. 2007. Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis. *Science* 317 (2007), 1360 – 1366.

[26] David Earl Hostallero, Daewoo Kim, Sangwoo Moon, Kyunghwan Son, Wan Ju Kang, and Yung Yi. 2020. Inducing cooperation through reward reshaping based on peer evaluations in deep multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*. 520–528.

[27] Edward Hughes, Joel Z. Leibo, Matthew Phillips, K. Tuyls, Edgar A. Duéñez-Guzmán, A. Castañeda, Iain Dunning, Tina Zhu, Kevin R. McKee, R. Koster, Heather Roff, and T. Graepel. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *NeurIPS*.

[28] Max Jaderberg, Wojciech Czarnecki, Iain Dunning, Luke Marris, Guy Lever, A. Castañeda, Charlie Beattie, Neil C. Rabinowitz, Ari S. Morcos, Avraham Ruderman, Nicolas Sonnerat, Tim Green, Louise Deason, Joel Z. Leibo, D. Silver, D. Hassabis, K. Kavukcuoglu, and T. Graepel. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* 364 (2019), 859 – 865.

[29] Natasha Jaques, A. Lazaridou, Edward Hughes, Çaglar Gülçehre, Pedro A. Ortega, D. Strouse, Joel Z. Leibo, and N. D. Freitas. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *ICML*.

[30] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and Eiichi Osawa. 1997. RoboCup: The Robot World Cup Initiative. In *AGENTS '97*.

[31] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *AAMAS-17*.

[32] Margaret M Luciano, Leslie A. DeChurch, and J. Mathieu. 2018. Multiteam Systems: A Structural Framework and Meso-Theory of System Functioning. *Journal of Management* 44 (2018), 1065 – 1096.

[33] W. Macke, R. Mirsky, and P. Stone. 2021. Expected Value of Communication for Planning in Ad Hoc Teamwork. In *AAAI-21*.

[34] J. E. Mathieu, M. A. Marks, and S. J. Zaccaro. 2001. Multiteam Systems. *Handbook of Industrial, Work, and Ordanizational Psychology* 2 (2001).

[35] Kevin R. McKee, I. Gemp, Brian McWilliams, Edgar A. Duéñez-Guzmán, Edward Hughes, and Joel Z. Leibo. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. *AAMAS* (2020).

[36] Reuth Mirsky, William Macke, A. Wang, Harel Yedidsion, and P. Stone. 2020. A Penny for Your Thoughts: The Value of Communication in Ad Hoc Teamwork. In *IJCAI*.

[37] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (2015), 529–533.

[38] E. Ostrom. 1990. Governing the Commons: The Evolution of Institutions for Collective Action. *Natural Resources Journal* 32 (1990), 415.

[39] Alexander Peysakhovich and Adam Lerer. 2018. Prosocial Learning Agents Solve Generalized Stag Hunts Better than Selfish Ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18).* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2043–2044.

[40] M. Pollack. 1986. A Model of Plan Inference that Distinguishes between the Beliefs of Actors and observers. In *ACL*.

[41] Martha E. Pollack. 1990. Plans As Complex Mental Attitudes. In *Intentions in Communication*. MIT Press, 77–103.

[42] J. Porck, Fadel K Matta, J. Hollenbeck, Jo K. Oh, Klodiana Lanaj, and S. Lee. 2019. Social Identification in Multiteam Systems: The Role of Depletion and Task Complexity. *Academy of Management Journal* 62 (2019), 1137–1162.

[43] David Radke, Kate Larson, and Tim Brecht. 2022. Exploring the Benefits of Teams in Multiagent Learning. In *IJCAI*.

[44] D. T. Radke, D. L. Radke, T. Brecht, and A. Pawelczyk. 2021. Passing and Pressure Metrics in Ice Hockey. *Workshop of AI for Sports Analytics* (2021).

[45] David G. Rand and M. Nowak. 2013. Human cooperation. *Trends in Cognitive Sciences* 17 (2013), 413–425.

[46] A. Rapoport. 1974. Prisoner's Dilemma — Recollections and Observations.

[47] Peter Richerson. 1998. *The Evolution of Human Ultra-sociality* (2.01 ed.). University of California, Davis, Davis, California.

[48] Heechang Ryu, Hayong Shin, and Jinkyoo Park. 2020. Multi-Agent Actor-Critic with Hierarchical Graph Attention Network. In *AAAI*.

[49] Heechang Ryu, Hayong Shin, and Jinkyoo Park. 2021. Cooperative and Competitive Biases for Multi-Agent Reinforcement Learning. In *AAMAS*.

[50] F. Santos, F. C. Santos, and J. Pacheco. 2018. Social norm complexity and past reputations in the evolution of cooperation. *Nature* 555 (2018), 242–245.

[51] E. Schnell, R. Schimmelpfennig, and M. Muthukrishna. 2021. The Size of the Stag Determines the Level of Cooperation. *bioRxiv* (2021).

[52] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* (2017).

[53] J. A. Simpson and E. S. C. Weiner. 1989. *The Oxford English Dictionary*. Oxford University Press.

[54] P. Stone, G. Kaminka, S. Kraus, and J. Rosenschein. 2010. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. In *AAAI*.

[55] E. Sundstrom, K. P. D. Meuse, and D. Futrell. 1990. Work teams: Applications and effectiveness. *American Psychologist* 45 (1990), 120–133.

[56] Milind Tambe. 1997. Towards Flexible Teamwork. *J. Artif. Intell. Res.* 7 (1997), 83–124.

[57] M. Tomasello, A. Melis, C. Tennie, Emily Wyman, and E. Herrmann. 2012. Two Key Steps in the Evolution of Human Cooperation. *Current Anthropology* 53 (2012), 673 – 692.

[58] M. Tomasello and Amrisha Vaish. 2013. Origins of human cooperation and morality. *Annual review of psychology* 64 (2013), 231–55.

[59] Eugene Vinitsky, Natasha Jaques, Joel Leibo, Antonio Castenada, and Edward Hughes. 2019. An Open Source Implementation of Sequential Social Dilemma Games. https://github.com/eugenevinitsky/sequential_social_dilemma_games/issues/182. GitHub repository.

[60] J. X. Wang, E. Hughes, C. Fernando, W. M. Czarnecki, E. A. Duéñez-Guzmán, and J. Z. Leibo. 2019. Evolving Intrinsic Motivations for Altruistic Behavior. In *AAMAS-19*. 683–692.

[61] Julia Wijnmaalen, H. Voordijk, S. Rietjens, and G. Dewulf. 2019. Intergroup behavior in military multiteam systems. *Human Relations* 72 (2019), 1081 – 1104.

[62] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. 2020. Learning to incentivize other learning agents. *Advances in Neural Information Processing Systems* 33 (2020), 15208–15219.

[63] S. Zaccaro, M. Marks, and Leslie A. DeChurch. 2012. Multiteam Systems: An Introduction.

[64] Stephen J. Zaccaro, Samantha Dubrow, Elisa M. Torres, and Lauren N.P. Campbell. 2020. Multiteam Systems: An Integrated Review and Comparison of Different Forms. *Annual Review of Organizational Psychology and Organizational Behavior* 7, 1 (2020), 479–503.

	Cooperate	Defect
Cooperate	$b - c, b - c$	$-c, b$
Defect	$b, -c$	$0, 0$

Table 1: An example of the Prisoner's Dilemma with the costs (c) and benefits (b) of cooperating ($b > c > 0$).

A ENVIRONMENT CONFIGURATIONS

A.1 Iterated Prisoner's Dilemma (IPD)

A.1.1 Matching Algorithm. At each instance of the IPD, agent i is given a counterpart, j to play the game in Table 1. Agent pairings are assigned subject to v, the probability that the counterpart is a teammate. In our analysis, we experiment with three different environments with $|\mathcal{T}| = 5$ teams of five agents each. When $v = 0.06$ agents are four times more likely to receive a counterpart from another team. When $v = 0.2$, a counterpart from any of the five teams has equal likelihood. When $v = 0.5$, agents are four times more likely to receive a counterpart from their own team. Each episode, we construct N pairings by matching each agent $i \in N$ to a counterpart $j \in N \setminus \{i\}$ sampled subject to v. Each agent observes the team their counterpart belongs to through a numerical signal $s_i \in S$, but not their actual individual identity.

A.1.2 IPD RL Algorithm. For each round of the IPD, agent i is paired with another agent j chosen randomly from the population, subject to the probability of being paired with a teammate v. The two agents play one iteration of the game shown in Table 1. Each agent observes the team their counterpart belongs to instead their actual identity. In particular, for a pair of agents, i and j, their states s_i and s_j are defined as $s_i = T_j$ and $s_j = T_i$. Given each s, the two agents simultaneously choose actions a_i and a_j, either to cooperate or defect. They do not observe the action of their counterpart, but instead receive rewards $R_i^{\mathbf{cr}}$ and $R_j^{\mathbf{cr}}$ based on every interaction and their credo. Each i (and also j with their information) stores the tuple $\langle s_i, a_i, R_i^{\mathbf{cr}} \rangle$ in their replay buffer to train their policy after each episode using Deep Q-Learning [37] by sampling a random batch of 32 interactions. Each agent's internal neural network consists of an input layer of size $|\mathcal{T}| = 5$, two hidden layers of 200 nodes each with hyperbolic tangent activation functions, and a two-action output layer with a linear activation function. Our agents use a learning rate of 1×10^{-4} and discount factor of 0.99 with ϵ-exploration.

A.2 Cleanup Markov Game

A.2.1 Cleanup RL Algorithm. The typical environmental setup for Cleanup implemented in past work uses five agents. However, five agents are unable to create multiple teams with the same size. Therefore, we instantiate $N = 6$ agents to create $|\mathcal{T}| = 3$ teams of two agents each. We deploy the default Proximal Policy Optimization (PPO) [52] learning algorithm from the Cleanup repository [59]. PPO is a policy gradient algorithm which constrains the space of policy updates to avoid large policy updates for smoother training and has been previously shown as a good algorithm for agents to learn in Cleanup. Agents are only able to observe the a 15×15 box centered at their location and update their policies using the environment's default batch size of at least 16,000 timesteps and
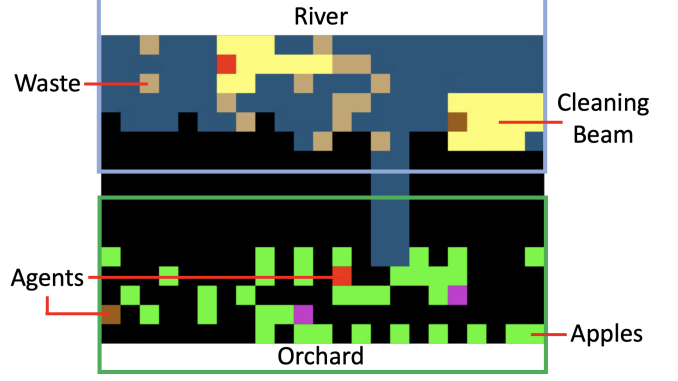

Figure 8: Screenshot of the Cleanup environment with three teams of two agents.

a maximum of 100,000 timesteps. Each episode executes for 1,000 timesteps and we run experiments for 1.6×10^8 timesteps. The learning rate decreases linearly from 1.2×10^{-3} to 1.2×10^{-5} over the first 2×10^7 timesteps and remains static afterwards.

A.2.2 Environment Parameters. An example of Cleanup with three teams is shown in Figure 8. The social dilemma dynamics of Cleanup rely on the existence of waste and apples. The functions which govern the creation of these features can be easily modified; however, we evaluate our model of teams primarily using the default parameters of the environment which include a waste regeneration rate, waste regeneration threshold, and apple generation rate. Once less than 40% of the river (aquifer) grid-cells contain waste, waste regenerates at each clean cell with a probability of 50% at each timestep. Below this waste regeneration, apples spawn at each location in the orchard with a linear probability ranging from 0% when waste is 40% of the river up to 5% when waste makes up 0% of the river.

B EQUILIBRIUM ANALYSIS

When calculating the expected values of cooperation and defection with different credo, the fully self-focused and system-focused values are simply calculated using Table 1 of the main text. Team-focused credo becomes more complex since it is a mixture of the mixed-motive and common interest game depending on the probability of being paired with a teammate v. We show the derivation for team-focused agents and continue with the final equilibrium with credo below.

B.1 Team-Focused Agents

The expected utility of choosing to cooperate (C) or defect (D) for an agent with team-focused credo can be derived based on Table 1, v, and the strategy of j, σ_j. First we show the derivation for a fully team-focused agent i's expected utility for choosing C subject to j's strategy:

$$\mathbb{E}(C,\sigma)_{\mathcal{T}} = v\left[\sigma_j(b-c) + (1-\sigma_j)\frac{b-c}{2}\right] +$$
$$(1-v)\left[\sigma_j(b-c) + (1-\sigma_j)(-c)\right]$$
$$= v\left[\frac{2\sigma_j(b-c)}{2} + \frac{b-c}{2} - \frac{\sigma_j(b-c)}{2}\right] +$$
$$(1-v)\left[\sigma_j b - \sigma_j c - c + \sigma_j c\right]$$
$$= v\left[\frac{\sigma_j b - \sigma_j c}{2} + \frac{b-c}{2}\right] + (1-v)\left[\sigma_j b - c\right]$$
$$= v\left[\frac{(b-c)(\sigma_j+1)}{2}\right] + (1-v)\left[\sigma_j b - c\right]$$
$$= \frac{v(b-c)(\sigma_j+1)}{2} + (1-v)(\sigma_j b - c).$$

Now we show the derivation for a team-focused agent i's expected utility for choosing D subject to j's strategy:

$$\mathbb{E}(D,\sigma)_{\mathcal{T}} = v\left[\sigma_j\frac{(b-c)}{2}\right] + (1-v)\left[\sigma_j b\right]$$
$$= \frac{v\sigma_j(b-c)}{2} + (1-v)\sigma_j b$$

The terms for playing defection with a counterpart who mutually defects $(1-\sigma_j)$ is zero, and therefore omitted above. Next, we show how the final equilibrium is derived using our parameters which define credo.

B.2 Equilibrium with Credo

Our credo vector defines how self-focused, team-focused, or system-focused an agent is while it learns in our environment. We can calculate and derive when an agent has the incentive to cooperate in the Prisoner's Dilemma stage-game as:

$$\psi\mathbb{E}(C,\sigma)_{\mathcal{I}} + \phi\mathbb{E}(C,\sigma)_{\mathcal{T}} + \omega\mathbb{E}(C,\sigma)_{\mathcal{S}} \geq$$
$$\psi\mathbb{E}(D,\sigma)_{\mathcal{I}} + \phi\mathbb{E}(D,\sigma)_{\mathcal{T}} + \omega\mathbb{E}(D,\sigma)_{\mathcal{S}}.$$

Expanding each term with the derivations above and in the main text, we get:

$$\psi\left[\sigma_j(b-c) + (1-\sigma_j)(-c)\right] +$$
$$\phi\left[\frac{v(b-c)(\sigma_j+1)}{2} + (1-v)(\sigma_j b - c)\right] +$$
$$\omega\left[\sigma_j(b-c) + (1-\sigma_j)\left(\frac{b-c}{2}\right)\right] \geq \psi\left[\sigma_j(b)\right] +$$
$$\phi\left[\frac{v\sigma_j(b-c)}{2} + (1-v)\sigma_j b\right] + \omega\left[\sigma_j\left(\frac{b-c}{2}\right)\right].$$

We expand and simplify:

$$\psi\left[\sigma_j b - \sigma_j c - c + \sigma_j c\right] + \phi\left[\frac{v(b-c)(\sigma_j+1)}{2} - c + vc\right] +$$
$$\omega\left[\sigma_j(b-c) + \frac{b-c}{2} - \sigma_j\left(\frac{b-c}{2}\right)\right] \geq \psi\left[\sigma_j(b)\right] +$$
$$\phi\left[\frac{v\sigma_j(b-c)}{2}\right] + \omega\left[\sigma_j\left(\frac{b-c}{2}\right)\right].$$

We can subtract everything on the right and be left with zero.

$$\psi\left[-c\right] + \phi\left[\frac{v(b-c)}{2} - c + vc\right] +$$
$$\omega\left[\sigma_j(b-c) + \frac{b-c}{2} - \sigma_j(b-c)\right] \geq 0$$
$$\psi\left[-c\right] + \phi\left[\frac{v(b-c)}{2} - c + vc\right] + \omega\left[\frac{b-c}{2}\right] \geq 0$$

The self- and system-focus terms are now fully simplified leaving the team-focus derivation remaining. We move $\phi\left[2c\right]$ to the other side of the inequality and simplify further.

$$-\psi c + \phi\left[v(b-c) + 2vc\right] + \omega\left[\frac{b-c}{2}\right] \geq \phi\left[2c\right]$$
$$-\psi c + \phi\left[v - \frac{2c}{b+c}\right] + \omega\left[\frac{b-c}{2}\right] \geq 0$$
$$\phi\left(v - \frac{2c}{b+c}\right) + \omega\left(\frac{b-c}{2}\right) \geq \psi c$$

This last step represents the final derivation shown as Equation 2 of the main text. This equilibrium signifies under which conditions an agent has more incentive to cooperate than to defect.