Cooperative Open-ended Learning Framework for Zero-shot Coordination

Yang Li [*1] **Shao Zhang** [*2] **Jichen Sun** [2] **Yali Du** [3] **Ying Wen** [2] **Xinbing Wang** [2] **Wei Pan** [1]

Abstract

Zero-shot coordination in cooperative artificial intelligence (AI) remains a significant challenge, which means effectively coordinating with a wide range of unseen partners. Previous algorithms have attempted to address this challenge by optimizing fixed objectives within a population to improve strategy or behaviour diversity. However, these approaches can result in a loss of learning and an inability to cooperate with certain strategies within the population, known as cooperative incompatibility. To address this issue, we propose the **C**ooperative **O**pen-ended **LE**arning (**COLE**) framework, which constructs open-ended objectives in cooperative games with two players from the perspective of graph theory to assess and identify the cooperative ability of each strategy. We further specify the framework and propose a practical algorithm that leverages knowledge from game theory and graph theory. Furthermore, an analysis of the learning process of the algorithm shows that it can efficiently overcome cooperative incompatibility. The experimental results in the Overcooked game environment demonstrate that our method outperforms current state-of-the-art methods when coordinating with different-level partners. Our demo is available at `https://sites.google.com/view/cole-2023/`.

1. Introduction

Zero-shot coordination (ZSC) is a major challenge of cooperative AI to train agents that have the ability to coordinate with a wide range of unseen partners (Legg & Hutter, 2007; Hu et al., 2020). The traditional method of self-play (SP) (Tesauro, 1994) involves iterative improvement of strategies by playing against oneself. While SP can converge to an equilibrium of the game (Fudenberg et al., 1998), the strategies often form specific behaviours and conventions to achieve higher payoffs (Hu et al., 2020). As a result, a fully converged SP strategy may not be adaptable to coordinating with unseen strategies (Lerer & Peysakhovich, 2018; Hu et al., 2020).

To overcome the limitations of SP, most ZSC methods focus on promoting strategic or behavioural diversity by introducing population-based training (PBT) to improve strategies' adaptive ability (Carroll et al., 2019; Canaan et al., 2022; Zhao et al., 2021; Lupu et al., 2021). PBT aims to improve cooperative outcomes with other strategies in the population to promote zero-shot coordination with unseen strategies. This is achieved by maintaining a set of strategies to break the conventions of SP (Tesauro, 1994) and optimizing the rewards for each pair in the population. Most state-of-the-art (SOTA) methods attempt to pre-train a diverse population (Strouse et al., 2021; Lupu et al., 2021) or introduce hand-crafted methods (Canaan et al., 2022; Zhao et al., 2021), which are used to master cooperative games by optimizing fixed objectives within the population. These methods have shown to be efficacious in addressing intricate cooperative tasks such as Overcooked (Carroll et al., 2019) and Hanabi (Bard et al., 2020).

However, when optimizing a fixed population-level objective, such as expected rewards within population (Strouse et al., 2021; Lupu et al., 2021; Zhao et al., 2021) , the coordination ability of strategies within the population may not be improved. Specifically, while overall performance may improve, the coordination ability within the population may not be promoted in a simultaneous manner. This phenomenon, which we term "*cooperative incompatibility*", highlights the importance of considering the trade-offs between overall performance and coordination ability when attempting to optimize a fixed population-level objective.

In addressing the problem of cooperative incompatibility, we reformulate cooperative tasks as Graphic-Form Games (GFGs). In GFGs, strategies are characterized as nodes, with the weight of the edges between nodes representing the mean cooperative payoffs of the two associated strategies. Additionally, by utilizing sub-graphs of GFGs referred to as preference Graphic-Form Games (P-GFGs), we are

[*]Equal contribution [1]The University of Manchester [2]Shanghai Jiao Tong University [3]King's College London. Correspondence to: Ying Wen <ying.wen@sjtu.edu.cn>, Wei Pan <wei.pan@manchester.ac.uk>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

able to further profile each node's maximum cooperative payoff within the graph, enabling us to evaluate cooperative incompatibility and identify strategies that fail to collaborate. Furthermore, we propose the Cooperative Open-ended LEarning (**COLE**) framework, which iteratively generates a new strategy that approximates the best response to the empirical gamescapes of P-GFGs. We have proved that the COLE framework can converge to the optimal strategy with a Q-sublinear rate when using in-degree centrality as the preference evaluation metric.

To propose COLE framework to address the phenomenon of cooperative incompatibility, we implement a practical algorithm COLE_{SV} by combining the **S**hapley **V**alue solution (Shapley, 1971) with our GFG. COLE_{SV} comprises a simulator, a solver, and a trainer, designed to master cooperative tasks with two players specifically. The solver, utilizing the development of the intuitive solution concept Shapley value, evaluates the adaptive ability of strategies and calculates the cooperative incompatibility distribution. The trainer aims to approximate the best responses to the cooperative incompatibility distribution mixture in the most recent population. To evaluate the performance of the COLE_{SV}, we conducted experiments in Overcooked, a two-player cooperative task environment (Carroll et al., 2019). We evaluated the adaptive ability of COLE_{SV} by testing its performance against different level partners - middle-level and expert unseen partners. The middle-level partner is a commonly used behavior cloning model (Carroll et al., 2019), and the expert partners are strategies of current methods, i.e., SP, PBT, FCP, and MEP, and our proposed COLE_{SV}. The results of the experiments showed that COLE_{SV} outperforms the recent SOTA methods in both evaluation protocols. Additionally, through the analysis of GFGs and P-GFGs, the learning process of COLE_{SV} revealed that the framework efficiently overcomes cooperative incompatibility. The contributions in this paper can be summarized as follows.

- We introduce the concept of Graphic-Form Games (GFGs) and Preference Graphic-Form Games (P-GFGs) to intuitively reformulate cooperative tasks, which allows for a more efficient evaluation and identification of cooperative incompatibility during learning.
- We develop the concept of graphic-form gamescapes to help understand the objective and present the COLE framework to iteratively approximate the best responses preferred by most others.
- We prove that the algorithm will converge to the optimal strategy, and the convergence rate will be Q-sublinear when using in-degree preference centrality. Empirical experiments in the game Overcooked verify the proposed algorithm's effectiveness compared to SOTA methods.

2. Related Works

Zero-shot coordination. The goal of zero-shot coordination (ZSC) is to train a strategy that can coordinate effectively with unseen partners (Hu et al., 2020). Self-play (Tesauro, 1994; Carroll et al., 2019) is a traditional method of training a cooperative strategy, which involves iterative improvement of strategies by playing against oneself, but develops conventions between players and does not cooperate with other unseen strategies (Lerer & Peysakhovich, 2018; Hu et al., 2020). Other-play (Hu et al., 2020) is proposed to break such conventions by adding permutations to one of the strategies. However, this approach may be reduced to self-play if the game or environment does not have symmetries or has unknown symmetries. Another approach is population-based training (PBT) (Jaderberg et al., 2017; Carroll et al., 2019), which trains strategies by interacting with each other in a population. However, PBT does not explicitly maintain diversity and thus fails to coordinate with unseen partners(Strouse et al., 2021).

Recent research has focused on training robust strategies that use diverse populations of strategies (Strouse et al., 2021; Lupu et al., 2021; Zhao et al., 2021) to achieve the goal of ZSC. Fictitious co-play (FCP) (Strouse et al., 2021) obtains a population of periodically saved checkpoints during self-play training with different seeds and then trains the best response to the pre-trained population. TrajeDi (Lupu et al., 2021) also maintains a pre-trained self-play population but encourages distinct behavior among the strategies. The maximum entropy population (MEP) (Zhao et al., 2021) method proposes population entropy rewards to enhance diversity during pre-training. It employs prioritized sampling to select challenging-to-collaborate partners to improve generalization to previously unseen policies. Furthermore, methods such as MAZE (Xue et al., 2022) and CG-MAS (Mahajan et al., 2022) have been proposed to improve generalization ability through coevolution and combinatorial generalization. In this paper, we propose a COLE framework that could dynamically identify strategies that fail to coordinate due to cooperative incompatibility and continually poses and optimizes objectives to overcome this challenge and improve adaptive capabilities.

Open-ended learning. Another related area of research is open-ended learning, which aims to continually discover and approach objectives (Srivastava et al., 2012; Team et al., 2021; Meier & Mujika, 2022). In MARL, most open-ended learning methods focus on zero-sum games, primarily posing adaptive objectives to expand the frontiers of strategies (Lanctot et al., 2017; Balduzzi et al., 2019; McAleer et al., 2020; Yang et al., 2021; Liu et al., 2021; McAleer et al., 2022). In the specific context of ZSC, the MAZE method (Xue et al., 2022) utilizes open-ended learning by maintaining two populations of strategies and partners and

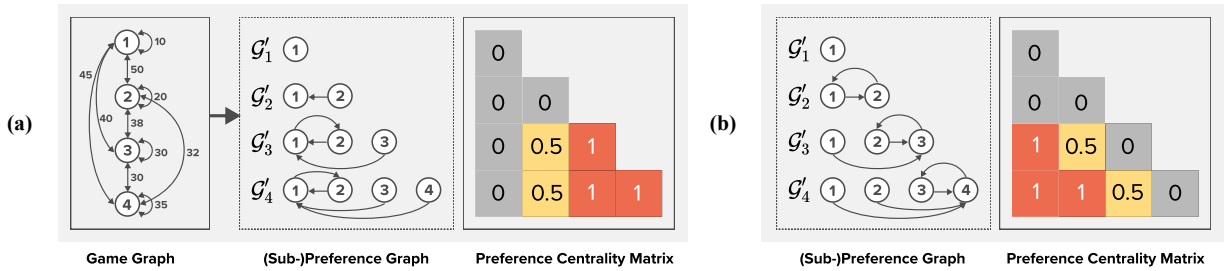


Figure 1. The Game Graph, (sub-) preference graph and corresponding preference centrality matrix. The (sub-) preference graphs are for all four iterations in the training process, and the corresponding preference in-degree centrality matrix is based on them. As can be observed in the $\mathcal{G}'_3$ and $\mathcal{G}'_4$, the newly updated strategies fail to be preferred by others and have centrality values of 1, despite an increase in the mean of rewards with all others. In *(b)*, we illustrate an ideal learning process in which a newly generated strategy can achieve higher outcomes with all previous strategies.

training them collaboratively throughout multiple generations. In each generation, MAZE pairs strategies and partners from the two populations and updates them together by optimizing a weighted sum of rewards and diversity. This method co-evolves the two populations of strategies and partners based on naive evaluations such as best or worst performance with strategies in partners. Our proposed method, COLE framework, combines GFGs and P-GFGs in open-ended learning to evaluate and identify the cooperative ability of strategies to solve cooperative incompatibility efficiently with theoretical guarantee.

3. Preliminaries

Normal-form Game. A two-player normal-form game is defined as a tuple $(N, \mathcal{A}, \mathbf{w})$, where $N = \{1, 2\}$ is a set of two players, indexed by i, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ is the joint action space, and $\mathbf{w} = (w_1, w_2)$ with $w_i : \mathcal{A} \to \mathbb{R}$ is a reward function for the player i. In a two-player common payoff game, two-player rewards are the same, meaning $w_1(a_1, a_2) = w_2(a_1, a_2)$ for $a_1, a_2 \in \mathcal{A}$.

Empirical Game-theoretic Analysis (EGTA), Empirical Game and Empirical Gamescape. EGTA is the study of finding meta-strategies based on experience with prior strategies (Walsh et al., 2002; Tuyls et al., 2018). An empirical game is built by discovering strategies and meta-reasoning about exploring the strategy space (Lanctot et al., 2017). Furthermore, empirical gamescapes (EGS) are introduced to represent strategies in functional form games geometrically (Balduzzi et al., 2019). Given a population $\mathcal{N}$ of n strategies, the empirical gamescapes is often defined as $\mathcal{G} := \{\text{convex mixture of rows of } \mathcal{M}\}$, where $\mathcal{M}$ is the empirical payoff table recording the expected outcomes for each joint strategy.

Cooperative Theoretic Concepts. We consider a set of players $\mathcal{N} = \{1, \ldots, n\}$, where a coalition is denoted as a subset of players $\mathcal{N}$, symbolized by $C \subseteq \mathcal{N}$. The player set $\mathcal{N}$ is also known as the grand coalition. A characteristic function game, denoted by G, consists of a pair (N, v), where $N = \{1, \ldots, n\}$ represents a finite, non-empty set of agents, and $v : 2^N \to \mathbb{R}$ is the characteristic function. This function assigns a real number $v(C)$ to each coalition $C \subseteq N$, with the number $v(C)$ typically considered as the coalition's value.

Shapley Value (Shapley, 1971) is one of the important solution concepts for characteristic function games (Chalkiadakis et al., 2011; Peleg & Sudhölter, 2007). The Shapley Value aims to distribute fairly the collective value, like the rewards and cost of the team across individuals by each player's contribution. Taking into account a coalition game $(\mathcal{N}, v)$ with a strategy set $\mathcal{N}$ and characteristic function v, the Shapley Value of a player $i \in \mathcal{N}$ could be obtained by

$$SV(i) = \frac{1}{n!} \sum_{\pi \in \Pi_{\mathcal{N}}} v(P_i^{\pi} \cup \{i\}) - v(P_i^{\pi}), \qquad (1)$$

where π is one of the one-to-one permutation mappings from $\mathcal{N}$ to itself in the permutation set Π and $\pi(i)$ is the position of player $i \in \mathcal{N}$ in permutation π. $P_i^{\pi} = \{j \in \mathcal{N} | \pi(j) < \pi(i)\}$ is the set of all predecessors of i in π.

Centrality in Graph Theory. Node Centrality is a graph theory concept that quantifies a node's relative importance or influence within a network. It is a measure of how central a node is to the overall network structure, with higher centrality values indicating greater importance. Degree Centrality is one of the simplest centrality concepts, based on the number of edges connected to a node (Freeman, 1978). In directed graphs, it can be further divided into in-degree (number of incoming edges) and out-degree (number of outgoing edges) Centrality.

PageRank (Page et al., 1999) is a centrality measure used to rank web pages in search engine results by estimating

the relative importance of each page in a hyperlink network. PageRank(WPG) (Xing & Ghorbani, 2004) is an extension of the original PageRank algorithm that considers edge weights in addition to the basic structure of the network. This modification allows the algorithm to handle networks where the strength of connections between nodes varies, such as in citation networks or social networks where the influence of nodes may differ. The formula of WPG is given as follows:

$$\sigma(u) = (1-d) + d \sum_{v \in B(u)} \sigma(v) \frac{I_u}{\sum_{p \in R(v)} I_p} \frac{O_u}{\sum_{p \in R(v)} O_p}, \tag{2}$$

where d is the damping factor set to 0.85, $B(u)$ is the set of nodes that point to u, $R(v)$ denotes the nodes to which v is linked, and I, O are the degrees of inward and outward of the node, respectively.

4. Cooperative Open-Ended Learning

In this section, we first introduce graphic-form games to intuitively reformulate cooperative games, then create an open-ended learning framework to solve cooperative incompatibility and further improve zero-shot adaptive ability.

4.1. Graphic-Form Games (GFGs)

It is important to evaluate cooperative incompatibility and identify those failed-to-collaborate strategies to conquer cooperative incompatibility. Therefore, we propose graphic-form games (GFGs) to reformulate normal-form cooperative games from the perspective of game theory and graph theory, which is the natural development of empirical games (Balduzzi et al., 2019). The definition of GFG is given below.

Definition 4.1 (Graphic-Form Game)**.** Given a set of parameterized strategies $\mathcal{N} = \{1, 2, \cdots, n\}$, a two-player graphic-form game (GFG) is a tuple $\mathcal{G} = (\mathcal{N}, \mathbf{E}, \mathbf{w})$, which could be represented as a directed weighted graph. $\mathcal{N}, \mathbf{E}, \mathbf{w}$ are the set of nodes, edges, and weights, respectively. Given an edge (i, j), $\mathbf{w}(i, j)$ represents the expected results of i playing with j. The graphic representation of GFG is called a game graph.

The payoff matrix of $\mathcal{G}$ is denoted as $\mathcal{M}$, where $\mathcal{M}(i, j) = \mathbf{w}(i, j), \forall i, j \in \mathcal{N}$. Our goal is to improve the upper bound of other strategies' outcomes in the cooperation within the population, which implies that the strategy should be preferred over other strategies.

Moreover, we propose preference graphic-form games (P-GFGs) as an efficient tool to analyze the current learning state, which can profile the degree of preference for each node in GFGs. Specifically, P-GFG is a subgraph of GFG, where each node only retains the out-edge with maximum weight among all out-edges except for its self-loop.

Definition 4.2 (Preference Graphic-Form Game)**.** Given a graphic-form game $(\mathcal{N}, \mathbf{E}, \mathbf{w})$, the Preference Graphic-Form Game (P-GFG) could be defined as $\mathcal{G}' = \{\mathcal{N}, \mathbf{E}', \mathbf{w}\}$, where $\mathbf{E}' = \{(i, j) | \arg\max_j \mathbf{w}(i, j), \forall j \in \{\mathcal{N} \backslash i\}, \forall i \in \mathcal{N}\}$ is the set of edges. The graphic representation of P-GFG is called a preference graph.

To deeply investigate the learning process, we further introduce the *sub-preference graphs* based on P-GFGs, which aim to reformulate previous learning states and analyze the learning behavior of the algorithm. Suppose that there is a set of sequentially generated strategies $\mathcal{N}_n = \{1, 2, \cdots, n\}$, where the index also represents the number of iterations for simplicity. For each previous iteration $i < n$, the sub-preference game form graph is denoted as $\{\mathcal{N}_i, \mathbf{E}'_i, \mathbf{w}_i\}$, where $\mathcal{N}_i = \{1, 2, \cdots, i\}$ is the set of strategies in iteration i, and $\mathbf{E}'_i, and\ \mathbf{w}_i$ are the corresponding edges and weights.

The semantics of the preference graph is that a strategy or node i prefers to play with the tailed node to achieve the highest results. In other words, the more in-edges one node has, the more cooperative ability this node can achieve. Ideally, if one strategy can adapt well to all others, all the other strategies in the preference graph will point to this strategy. To evaluate the adaptive ability of each node, the centrality concept is introduced into the preference graph to evaluate how a node is preferred.

Definition 4.3 (Preference Centrality)**.** Given a P-GFG $\{\mathcal{N}, E', \mathbf{w}\}$, preference centrality of $i \in \mathcal{N}$ is defined as,

$$\eta(i) = 1 - \mathrm{norm}(d_i),$$

where d_i is a graph centrality metric to evaluate how the node is preferred, and $\mathrm{norm}(\cdot) : \mathbb{R} \to [0, 1]$ is a normalization function.

Note that the d is a kind of centrality that could evaluate how much a node is preferred. A typical example of d is the centrality of degrees, which calculates how many edges point to the node.

Fig. 1 is an example of a common payoff game, showing the game graph, (sub-)preference graphs, and the preference centrality matrix for four sequentially generated strategies. Note that in the corresponding sub-preference graphs, the updated strategies fail to improve the outcome of others after the second iteration, and the preference centrality matrix also shows the same results. The example shows an existing cooperative incompatibility that presents as the value of η is kept at 1 in the matrix, meaning no nodes want to collaborate with the updated strategies. Ideally, all the other strategies should prefer latest strategy (Fig. 1 (b)) which means the monotonic improvement of cooperative ability.

Moreover, the analysis of the MEP algorithm, as shown in Fig. 2, discloses a cooperative incompatibility in the learning

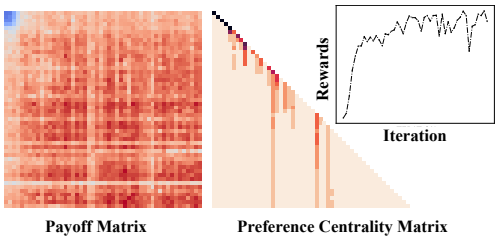


Figure 2. The payoff matrix of each strategy during training and the corresponding preference centrality matrix of the MEP algorithm in the Overcooked. The darker the color in the payoff matrix, the higher the rewards. The darker the color in the preference centrality matrix, the lower the centrality value, and the more other strategies prefer it.

process in Overcooked environment (Carroll et al., 2019). In the preference indegree centrality matrix, a strategy is preferred by more strategies if its color is darker. In the learning process of MEP, although the mean rewards are always improving (as shown in the upper-right of Fig. 2), serious cooperative incompatibility problems occur after a period of training, where more strategies prefer to play with some previous strategies with a darker color rather than new strategies to obtain higher rewards.

4.2. Cooperative Open-Ended Learning Framework

To tackle cooperative incompatibility by understanding the objective, we develop empirical gamescapes (Balduzzi et al., 2019) for GFGs, which geometrically represent strategies in graphic-form games. Given a GFG $\{\mathcal{N}, \mathbf{E}, \mathbf{w}\}$ and corresponding payoff matrix $\mathcal{M}$, the empirical gamescapse (EGS) is defined as

$$\bar{\mathcal{G}} := \{\text{convex mixture of rows of } \mathcal{M}\}\,. \tag{3}$$

However, learning directly with EGS to cooperate with these well-collaborated strategies is inefficient in improving adaptive ability. To conquer cooperative incompatibility, the natural idea is to learn with the mixture of cooperative incompatible distribution on the most recent population $\mathcal{N}$. Given a population $\mathcal{N}$, we present *cooperative incompatible solver* to assess how strategies collaborate, especially with those strategies that are difficult to collaborate with. The solver derives the cooperative incompatible distribution ϕ, where strategies that do not coordinate with others have higher probabilities.

We also optimize the cooperative incompatible mixture over the individual objective, which is the cumulative self-play rewards to improve the adaptive ability with expert partners. To simplify, we name it the individual and cooperative incompatible mixture (IPI mixture). We use an approximate oracle to approach the best response over the IPI mixture. Given strategy s_n, the oracle returns a new strategy s_{n+1} : $s_{n+1} = \text{oracle}(s_{n+1}, \mathcal{J}(s_n, \phi))$, with $\eta(s_{n+1}) = 0$, if possible. $\mathcal{J}$ is the objective function as follows,

$$\mathcal{J}(s_n, \phi) = \mathbb{E}_{p\sim\phi}\mathbf{w}(s_n, p) + \alpha\mathbf{w}(s_n, s_n), \tag{4}$$

where α is the balance hyperparameter. The objective consists of the cooperative compatible objective and the individual objective. The cooperative compatible objective aims to train the best response to those failed-to-collaborate strategies, and the individual objective aims to improve the adaptive ability with expert partners. We call the best response the local best-preferred strategy in the population if $\eta(s_{n+1}) = 0$.

However, arriving at the local best-preferred strategy with $\eta(s_{n+1}) = 0$ is hard or even impossible. Therefore, we seek to approximate the local best-preferred strategies by relaxing the local best strategy to the strategy whose preference centrality ranks top k. The approximate oracle could be rewritten as

$$s_{n+1} = \text{oracle}(s_n, \mathcal{J}(s_n, \phi)), with\ \mathcal{R}(\eta(s_{n+1})) > k, \tag{5}$$

where $\mathcal{R}$ is the ranking function.

We extend the approximated oracle to open-ended learning and propose COLE framework (Fig. 3). The COLE framework iteratively updates new strategies that approximate the local best-preferred strategies to the cooperative incompatible mixture and the individual objective. The simulator completes the payoff matrix with the newly generated strategy and others in the population. The solver aims to derive the cooperative incompatible distribution of the Game Graph builder and the cooperative-incompatible solver. The trainer uses the oracle to approximate the local best-preferred strategy to the cooperative incompatible mixture and individual objective and outputs a newly generated strategy added to the population for the next generation.

Although we relax the local best-preferred strategy to the strategy in the top k centrality in the constraint, COLE framework still converges to a local best-preferred strategy with zero preference centrality. Formally, the local best-preferred strategy convergence theorem is given as follows.

Theorem 4.4. *Let $s_0 \in \mathcal{S}$ be the initial strategy and $s_i = \text{oracle}(s_{i-1})$ for $i \in \mathbb{N}$. Under the assumption that $\lim_{n\to\infty} \mathcal{J}(s_n, \phi) \geq \mathcal{J}(s_{n-1}, \phi)$ holds, we can say that the sequence $\{s_i\}$ for $i \in \mathbb{N}$ converges to a local optimal strategy s^*, i.e., the local best-preferred strategy.*

Proof. See Appendix B. □

Besides, if we choose in-degree centrality as the preference centrality function, the convergence rate of COLE framework is Q-sublinear.

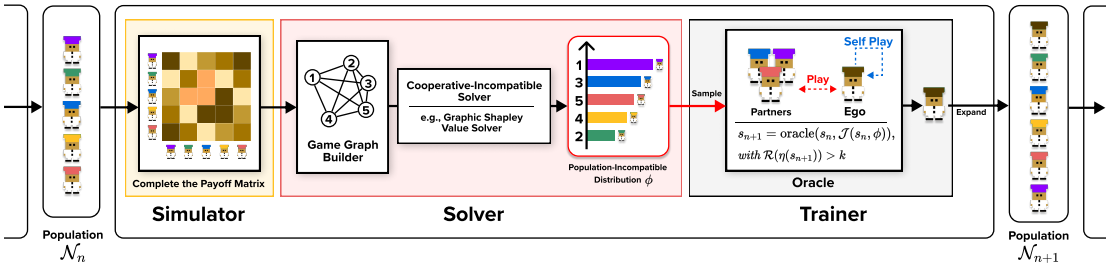


Figure 3. An overview of one generation in COLE framework: The solver derives the cooperative incompatible distribution ϕ using a cooperative incompatibility solver, which can be any algorithm that evaluates cooperative contribution. The trainer then approximates the relaxed best response by optimizing individual and cooperative compatible objectives. The oracle's training data is generated using partners selected based on the cooperative incompatibility distribution and the agent's strategy. Finally, the approximated strategy s_{n+1} is added to the population, and the next generation begins.

Corollary 4.5. *Let $\eta : \mathcal{G}' \to \mathbb{R}^n$ be a function that maps a P-GFG to its in-degree centrality, the convergence rate of the sequence $\{s_i\}$ is Q-sublinear concerning η.*

Proof. See Appendix C. □

5. Practical Algorithm

To address common-payoff games with two players, we implemented COLE$_{\text{SV}}$, where SV refers to *Shapley Value*, based on COLE framework that can overcome cooperative incompatibility and improve zero-shot coordination capabilities, focusing on the solver and trainer components. As shown in Fig. 3, at each generation, COLE$_{\text{SV}}$ inputs a population $\mathcal{N}$ and generates an approximate local best-preferred strategy added to $\mathcal{N}$ to expand the population. The simulator calculates the payoff matrix $\mathcal{M}$ for the input population $\mathcal{N}$. Each element $\mathcal{M}(i, j)$ for $i, j \in \mathcal{N}$ represents the cumulative rewards of the players i and j at both starting positions. The solver evaluates and identifies failed-to-collaborate strategies by calculating the incompatible cooperative distribution. To effectively evaluate the cooperative ability of each strategy with all others, we incorporate weighted PageRank (WPG) (Xing & Ghorbani, 2004) from graph theory into the Shapley Value to evaluate adaptability, particularly with failed-to-collaborate strategies. The trainer then approximates the local best-preferred strategy over the recent population.

5.1. Solver: Graphic Shapley Value

To approximate the local best-preferred strategies over the recent population and overcome cooperative incompatibility, we need to calculate the cooperative incompatible distribution as the mixture. In this paper, we combine the Shapley Value (Shapley, 1971) solution, an efficient single solution concept for cooperative games to assign the obtained team value across individuals, with our GFG to evaluate and identify the strategies that did not cooperate. To apply the Shapley Value, we define an additional characteristic function to evaluate the value of the coalition. Formally, given a coalition $C \subseteq \mathcal{N}$, we have the following:

$$v(C) = \mathbb{E}_{i \sim C, j \sim C} \sigma(i)\sigma(j)\mathbf{w}(i, j), \tag{6}$$

where σ is a mapping function that evaluates how badly a node performs on its game graph. We use the characteristic function to evaluate the coalition value of how it could cooperate with those hard-to-collaborate strategies.

We take the inverse of WPG (Xing & Ghorbani, 2004) on the game graph as the metric σ. WPG is proposed to assess the popularity of a node in a complex network, as given in Eq. 2. Therefore, the metric σ evaluates how unpopular a node is and is equal to the inverse of the WPG value.

Then we calculate the Shapley Value of each node by taking a characteristic function in equation 1, named the graphic Shapley Value. We utilize the Monte Carlo permutation sampling (Castro et al., 2009) to approximate the Shapley Value, which can reduce the computation complexity from exponential time to linear time. After inverting the probabilities of the graphic Shapley Value, we get the cooperative incompatible distribution ϕ, where strategies that fail to collaborate with others have higher probabilities. We provide the Graphic Shapley Value algorithm in Appendix D.

5.2. Trainer: Approximating local best-preferred Strategy

The trainer takes the cooperative incompatible distribution ϕ as input and samples its teammates to learn to approach the local best-preferred strategy on the IPI mixture.

Recall the oracle for $s_n : s_{n+1} = \text{oracle}(s_{n+1}, \mathcal{J}(s_n, \phi))$, with $\mathcal{R}(\eta(s_{n+1})) > k$. COLE$_{\text{SV}}$ aims to optimize the local best-preferred strategy over the IPI mixture. The $\mathcal{J}(s_n, \phi)$ is the joint objective that consists of individual and cooperative compatible objectives. The individual objective aims

Algorithm 1 COLE_{SV} Algorithm

Input: population $\mathcal{N}_0$, the sample times a, b of $\mathcal{J}_i, \mathcal{J}_c$, hyperparameters α, k
for $t = 1, 2, \cdots$, **do**
  // Step 1: Completing the payoff matrix
  $\mathcal{M}_n \leftarrow \text{Simulator}(\mathcal{N}_t)$
  // Step 2: Solving the cooperative incompatibility distribution
  $\phi = \text{Graphic Shapley Value}(\mathcal{N}_t)$ by Algorithm 2
  // Step 3: Approximate the local best-preferred strategy
  $\mathcal{J} = \sum_{p\sim\phi}^{b} \phi(p)\mathbf{w}(s_t, p) + \alpha \sum^{a} \mathbf{w}(s_t, s_t)$, where $s_t = \mathcal{N}_t(t)$, ϕ is updated each time by Eq 8
  $s_{t+1} = \text{oracle}(s_t, \mathcal{J})$ with $\mathcal{R}(\eta(s_{n+1})) > k$
  // Step 4: Expand the population
  $\mathcal{N}_{t+1} = \mathcal{N}_t \cup \{s_{t+1}\}$
end for

to improve the performance within itself and promote the adaptive ability with expert partners, formulated as follows: $\mathcal{J}_i(s_n) = \mathbf{w}(s_n, s_n)$, where s_n is the strategy named ego strategy that needs to optimize in generation n.

And the cooperative compatible objective aims to improve cooperative outcomes with those failed-to-collaborate strategies: $\mathcal{J}_c = \mathbb{E}_{p\sim\phi}\mathbf{w}(s_n, p)$, where the objective is the expected rewards of s_n with cooperative incompatible distribution-supported partners. $\mathbf{w}$ estimates and records the mean cumulative rewards of multiple trajectories and starting positions. The expectation can be approximated as:

$$\mathcal{J}_c = \sum_{p\sim\phi}^{b} \phi(p)\mathbf{w}(s_t, p), \tag{7}$$

where b is the number of sampling times.

To balance exploitation and exploration as the learning continues, we present the Sampled Upper Confidence Bound for Game Graph (SUCG) that combines the Upper Confidence Bound (UCB) and GFG to control the sampling for more strategies with higher probabilities or new strategies. Additionally, we view the SUCG value as the probability of sampling teammates instead of using the maximum item in typical UCB algorithms. Specifically, in the game graph, we keep the information on the times that a node has been visited. Therefore, the probability of each node considers both the Shapley Value and visiting times, denoted as $\hat{p}$. The SUCG for any node u in $\mathcal{N}$ could be calculated as follows:

$$\hat{\phi}(u) = \phi(u) + c\frac{\sqrt{\sum_{i\in\mathcal{N}} \mathbf{N}(i)}}{1 + \mathbf{N}(u)}, \tag{8}$$

where c is a hyperparameter that controls the degree of exploration and $\mathbf{N}(i)$ is the visit times of node i. SUCG

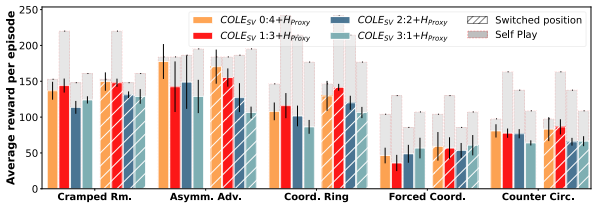


Figure 4. The result of the combining objectives' effectiveness evaluation. Mean episode rewards over 400 timesteps trajectories for COLE_{SV} s with different objective ratios 0:4, 1:3, 2:2, and 3:1, paired with the unseen middle-level partner H_{proxy}. The gray bars behind present the rewards of self-play.

could efficiently prevent COLE_{SV} from generating data with a few fixed strategies that did not cooperate, which could lead to a loss of adaptive ability.

We conclude the COLE_{SV} as Algorithm 1. Moreover, to verify the influence of different ratios of two objectives, we denote COLE_{SV} with different ratios as 0:4, 1:3, 2:2, and 3:1. Specifically, COLE_{SV} with $a : b$ represents different partner sampling ratios for the combining objective, where a is the corresponding times to generate data using self-play for the individual objective, and b is the number of sampling times in $\mathcal{J}_c$. For example, COLE_{SV} 1:3 represents that the ego agent is trained by using self-play once and with partners sampled from the cooperative incompatible distribution three times to generate data and update objectives.

6. Experiments

6.1. Environment and Experimental Setting

In this paper, we conduct a series of experiments in the Overcooked environment (Carroll et al., 2019; Charakorn et al., 2020; Knott et al., 2021). The details of the Overcooked environment can be found in Appendix E. We construct evaluations with different ratios between individual and cooperative compatible objectives, such as 0:4, 1:3, 2:2, and 3:1. These studies demonstrate the effectiveness of optimizing both individual and cooperative incompatible goals. We also compare our method with other methods, including self-play (Tesauro, 1994; Carroll et al., 2019), PBT (Jaderberg et al., 2017; Carroll et al., 2019), FCP (Strouse et al., 2021), and MEP (Zhao et al., 2021), all of which use PPO (Schulman et al., 2017) as the RL algorithm. To thoroughly assess the ZSC ability, we evaluated the algorithms with unseen middle-level and expert partners. We use the human proxy model H_{proxy} proposed by Carroll et al.(Carroll et al., 2019) as middle-level partners and the models trained with baselines and COLE_{SV} as expert partners. The rewards' mean is recorded as the performance of each method in collaborating with expert teammates. In the case study, we analyze the learning process of COLE_{SV}, which shows that our method overcomes cooperative incompatibility. Furthermore, we

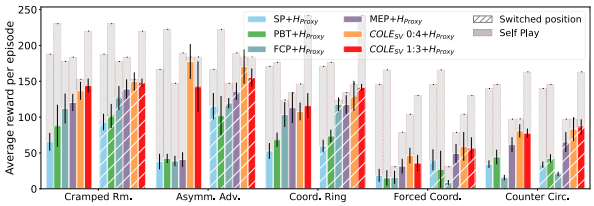


Figure 5. Performance with middle-level partners. The performance of COLE_{SV} with middle-level partners is presented in terms of mean episode rewards over 400 timesteps trajectories for different objective ratios of 0:4 and 1:3, when paired with the unseen middle-level partner H_{proxy}. The results include the mean and standard error over five different random seeds. The gray and hashed bars indicate the rewards obtained when playing with themselves and the performance when starting positions are switched.

visualize the trajectories with different ratios and play with expert teammates to analyze how the ratios affect the learned strategies. Appendix F and Appendix G give details of the implementation of COLE_{SV} and baselines.

6.2. Combining Objectives' Effectiveness Evaluation

This section evaluated the effectiveness of different objective ratios, including 0:4, 1:3, 2:2, and 3:1 of two objectives. We divided each training batch into four parts, the ratio indicating the proportion of data generated by self-play and data generated by playing with strategies from the cooperative incompatible distribution. We omitted the 4:0 ratio as it would result in the framework degenerating into self-play. Fig. 4 shows the mean rewards of episodes over 400 timesteps of gameplay when paired with the unseen middle-level partner H_{proxy} (Carroll et al., 2019). We found that COLE_{SV} with ratios 0:4 and 1:3 achieved better performance than the other ratios. In particular, COLE_{SV}, with a ratio of 1:3, outperformed the other methods in the Cramped Room, Coordination Ring, and Counter Circuit layouts. On the Forced Coordination layout, which is particularly challenging for cooperation due to the separated regions, all four ratios performed similarly on average across different starting positions. Interestingly, when paired with the middle-level partner, COLE_{SV} with only the cooperative compatible objective (ratio 0:4) performed better on the Asymmetric Advantages and Forced Coordination layouts. We discuss this phenomenon further in Section 6.3. The effectiveness evaluations indicate that combining individual and cooperatively compatible objectives is crucial to improving performance with unseen partners. In general, we choose the ratio of 1:3 as the best choice.

6.3. Evaluation with Different Levels of Partners

To thoroughly evaluate the zero-shot cooperative ability of all methods, we adopted two sets of evaluation protocols. The first protocol involves playing with a trained human model H_{proxy} trained in behavior cloning.However, due to the quality and quantity of human data used for behavior cloning to train the human model is limited, the capabilities of the human proxy model can only be classified as middle-level. Therefore, we use an additional evaluation protocol to coordinate with unseen expert partners. We selected the best models of our reproduced baselines and COLE_{SV} 0:4 and 1:3 as expert partners.

Table 1. Performance with expert partners. Mean episode rewards over 1 min trajectories for baselines and COLE_{SV} with ratio 0:4, 1:3. Each column represents a different expert group, in which the result is the mean reward for each model playing with all others.

Layout	Ratio	Baselines				COLEs
		SP	PBT	FCP	MEP	
Cramped Rm.	0:4	153.00	198.50	199.83	178.83	169.76
	1:3	165.67	209.83	207.17	196.83	**212.80**
Asymm.Adv.	0:4	108.17	164.83	175.50	179.83	**182.80**
	1:3	108.17	161.50	172.17	179.83	178.80
Coord. Ring	0:4	132.00	106.83	142.67	130.67	118.08
	1:3	133.33	158.83	144.00	124.67	**166.32**
Forced Coord.	0:4	58.33	61.33	50.50	79.33	46.40
	1:3	61.50	70.33	62.33	38.00	**86.40**
Counter Circ.	0:4	44.17	48.33	60.33	21.33	90.72
	1:3	65.67	64.00	46.50	76.67	**105.84**

Fig. 5 presents the performance of SP, PBT, MEP, and COLE_{SV} with 0:4 and 1:3 when cooperating with middle-level partners. We observed that different starting positions on the left and right in asymmetric layouts resulted in significant performance differences for the baselines. For example, in the Asymmetric Advantages, the cumulative rewards of all baselines in the left position were nearly one-third of those in the right position. On the contrary, COLE_{SV} performed well at the left and right positions.

As shown in Fig. 5, COLE_{SV} outperforms other methods in all five layouts when paired with the middle-level partner-human proxy model. Interestingly, COLE_{SV} 0:4 with only the cooperatively compatible objective achieves better performance than COLE_{SV} 1:3 on some layouts, such as Asymmetric Advantages. However, the self-play rewards of COLE_{SV} 0:4 are much lower than COLE_{SV} 1:3 and even other baselines. Furthermore, the performance with unseen experts of COLE_{SV} 0:4 as shown in Table 1, is sometimes lower than the baselines. We visualize the trajectories in the evaluation at the expert level and provide further analysis to explain this situation in Appendix H.

Table 1 presents the outcomes of each method when cooperating with expert partners. Each column in the table represents different expert groups, including four baselines and one COLE_{SV} with a ratio of 0:4 or 1:3. The last column, labeled "COLEs," represents the mean rewards of the corresponding COLE_{SV} when working with other baselines. The

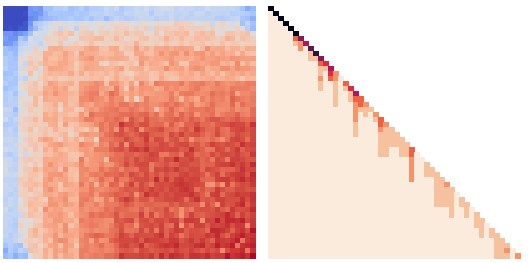


Figure 6. The learning process analysis of COLE_{SV} 1:3. The darker-colored element on the left represents higher rewards, while the darker-colored element on the right represents lower centrality. The clustering of darker-colored areas around the diagonal on the right indicates that the new strategy adopted in each generation is preferred by most strategies, thus overcoming the cooperative incompatibility.

table displays the mean cumulative rewards of each method when working with all other models in the expert group. The results indicate that COLE_{SV} 1:3 outperforms the baselines and COLE_{SV} 0:4, except in the layout of Asymmetric Advantages. In the Asymmetric Advantages, COLE_{SV} 0:4 only achieved a four-point victory over COLE_{SV} 1:3, which can be considered insignificant considering the margin of error. In the other four layouts, the rewards obtained by COLE_{SV} 1:3 while working with expert partners are significantly higher than those of COLE_{SV} 4:0 and the baselines.

Our results suggest that COLE_{SV} 1:3 has a stronger adaptive ability with different levels of partners. Furthermore, individual objectives are crucial in zero-shot coordination with expert partners. In conclusion, COLE_{SV} 1:3 is more robust and flexible in real-world scenarios when working with partners of different levels.

6.4. Effectively Conquer Cooperative Incompatibility

In our analysis of the learning process of COLE_{SV} 1:3 in the Overcooked environment, as shown in Fig. 6, we observe that the method effectively overcomes the problem of cooperative incompatibility. The figure on the left in Fig. 6 shows the payoff matrix of 50 uniformly sampled checkpoints during training, with the upper left corner representing the starting point of training. Darker red elements in the payoff matrix indicate higher rewards. The figure on the right displays the centrality matrix of preferences, which is calculated by analyzing the learning process. Unlike the payoff matrix, the darker elements in the centrality matrix indicate lower values, indicating that more strategies prefer them in the population. As shown in the figure, the darker areas cluster around the diagonal of the preference centrality matrix, indicating that most of the others prefer the updated strategy of each generation. Thus, we can conclude that our proposed COLE_{SV} effectively overcomes the problem of cooperative incompatibility.

7. Conclusion

In this paper, we propose graphic-form games and preference graphic-form games to intuitively reformulate cooperative games, which can efficiently evaluate and identify cooperative incompatibility. Furthermore, we develop empirical gamescapes for GFG to understand the objectives and present COLE framework to iteratively approximate the best response preferred by most others over the most recent population. Theoretically, we prove that COLE framework converges to the optimal strategy preferred by all others. Furthermore, if we choose the in-degree centrality as the preference centrality function, the convergence rate would be Q-sublinear. Empirically, our experiments on the Overcooked environment show that our algorithm COLE_{SV} outperformed SOTA ones and that COLE_{SV} efficiently overcame cooperative incompatibility.

Limitations and Future Work. Our practical algorithm, COLE_{SV}, incorporates the Shapley Value as a tool and develops the Graphic Shapley Value to scrutinize cooperative incompatibility. Despite our use of Monte Carlo permutation sampling to mitigate computational complexity, this remains a challenge. As such, future efforts should focus on enhancing the efficiency of the Graphic Shapley Value solver and investigating alternative solvers for evaluating cooperative incompatibility. Besides, our framework is restricted to two-player games. Therefore, future research should also aim to extend the COLE framework and develop corresponding algorithms for more complex multi-player games.

Acknowledgements

Yang Li is supported by the China Scholarship Council (CSC) Scholarship. The Shanghai Jiao Tong University team is supported by the National Key R&D Program of China (2022ZD0114804), National Natural Science Foundation of China (No.62106141), and Shanghai Sailing Program (21YF1421900). The authors also thank Xihuai Wang for his kind assistance and advice.

References

Balduzzi, D., Garnelo, M., Bachrach, Y., Czarnecki, W. M., Perolat, J., Jaderberg, M., and Graepel, T. Open-ended learning in symmetric zero-sum games, 2019.

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

Canaan, R., Gao, X., Togelius, J., Nealen, A., and Menzel, S. Generating and adapting to diverse ad-hoc partners in hanabi. *IEEE Transactions on Games*, pp. 1–1, 2022.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the shapley value based on sampling. *Comput. Oper. Res.*, 36:1726–1730, 2009.

Chalkiadakis, G., Elkind, E., and Wooldridge, M. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(6):1–168, 2011.

Charakorn, R., Manoonpong, P., and Dilokthanakul, N. Investigating partner diversification methods in cooperative multi-agent deep reinforcement learning. In *International Conference on Neural Information Processing*, pp. 395–402. Springer, 2020.

Freeman, L. C. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978. ISSN 0378-8733.

Fudenberg, D., Drew, F., Levine, D. K., and Levine, D. K. *The theory of learning in games*, volume 2. MIT press, 1998.

Hu, H., Lerer, A., Peysakhovich, A., and Foerster, J. N. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, 2020.

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. Population based training of neural networks. *arXiv preprint arXiv:1711.09846*, 2017.

Knott, P., Carroll, M., Devlin, S., Ciosek, K., Hofmann, K., Dragan, A. D., and Shah, R. Evaluating the robustness of collaborative agents. *arXiv preprint arXiv:2101.05507*, 2021.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., and Graepel, T. A unified game-theoretic approach to multiagent reinforcement learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Legg, S. and Hutter, M. Universal intelligence: A definition of machine intelligence. *Minds Mach.*, 17(4):391–444, dec 2007.

Lerer, A. and Peysakhovich, A. Learning social conventions in markov games. *arXiv preprint arXiv:1806.10071*, 2018.

Liu, X., Jia, H., Wen, Y., Hu, Y., Chen, Y., Fan, C., Hu, Z., and Yang, Y. Towards unifying behavioral and response diversity for open-ended learning in zero-sum games. *Advances in Neural Information Processing Systems*, 34: 941–952, 2021.

Lupu, A., Cui, B., Hu, H., and Foerster, J. Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning*, pp. 7204–7213. PMLR, 2021.

Mahajan, A., Samvelyan, M., Gupta, T., Ellis, B., Sun, M., Rocktäschel, T., and Whiteson, S. Generalization in cooperative multi-agent systems. *arXiv preprint arXiv:2202.00104*, 2022.

McAleer, S., Lanier, J. B., Fox, R., and Baldi, P. Pipeline psro: A scalable approach for finding approximate nash equilibria in large games. *Advances in neural information processing systems*, 33:20238–20248, 2020.

McAleer, S., Lanier, J., Wang, K., Baldi, P., Fox, R., and Sandholm, T. Self-play psro: Toward optimal populations in two-player zero-sum games. *arXiv preprint arXiv:2207.06541*, 2022.

Meier, R. and Mujika, A. Open-ended reinforcement learning with neural reward functions. *arXiv preprint arXiv:2202.08266*, 2022.

Page, L., Brin, S., Motwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Peleg, B. and Sudhölter, P. *Introduction to the theory of cooperative games*. Theory and Decision Library Series C. Springer Science+Business Media, United States, 2 edition, 2007.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shapley, L. S. Cores of convex games. *International journal of game theory*, 1:11–26, 1971.

Srivastava, R., Steunebrink, B., Stollenga, M., and Schmidhuber, J. Continually adding self-invented problems to the repertoire: First experiments with powerplay. *2012 IEEE International Conference on Development and Learning and Epigenetic Robotics, ICDL 2012*, pp. 1–6, 11 2012.

Strouse, D., McKee, K., Botvinick, M., Hughes, E., and Everett, R. Collaborating with humans without human data.

Advances in Neural Information Processing Systems, 34: 14502–14515, 2021.

Team, O. E. L., Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., et al. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*, 2021.

Tesauro, G. Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 6 (2):215–219, 1994.

Tuyls, K., Perolat, J., Lanctot, M., Leibo, J. Z., and Graepel, T. A generalised method for empirical game theoretic analysis. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, pp. 77–85, Richland, SC, 2018.

Walsh, W. E., Das, R., Tesauro, G., and Kephart, J. O. Analyzing complex strategic interactions in multi-agent systems. In *AAAI-02 Workshop on Game-Theoretic and Decision-Theoretic Agents*, pp. 109–118, 2002.

Xing, W. and Ghorbani, A. Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pp. 305–314, 2004. doi: 10.1109/DNSR.2004.1344743.

Xue, K., Wang, Y., Yuan, L., Guan, C., Qian, C., and Yu, Y. Heterogeneous multi-agent zero-shot coordination by coevolution. *arXiv preprint arXiv:2208.04957*, 2022.

Yang, Y., Luo, J., Wen, Y., Slumbers, O., Graves, D., Bou Ammar, H., Wang, J., and Taylor, M. E. Diverse auto-curriculum is critical for successful real-world multiagent learning systems. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, pp. 51–56, Richland, SC, 2021.

Zhao, R., Song, J., Haifeng, H., Gao, Y., Wu, Y., Sun, Z., and Wei, Y. Maximum entropy population based training for zero-shot human-ai coordination. *arXiv preprint arXiv:2112.11701*, 2021.

A. Reproducibility Statement

Our demo is provided in https://sites.google.com/view/cole-2023/.

Besides, more reproducibility information could be found in the appendix.

- The detailed pseudocode of the Graphic Shapley Value Solver is provided in Appendix D.
- In Appendix E, we introduce the details of the experimental environment - the Overcooked game.
- The implementation and hyperparameters used in experiments are in Appendix F.

B. Proofs of Theorem 4.4

Theorem 4.4. *Let $s_0 \in \mathcal{S}$ be the initial strategy and $s_i = \text{oracle}(s_{i-1})$ for $i \in \mathbb{N}$. Under the assumption that $\lim_{n\to\infty} \mathcal{J}(s_n, \phi) \geq \mathcal{J}(s_{n-1}, \phi)$ holds, we can say that the sequence $\{s_i\}$ for $i \in \mathbb{N}$ converges to a local optimal strategy s^*, i.e., the local best-preferred strategy.*

Proof. According to the definition of the local best-preferred strategy, the local optimal strategy is the node with zero preference centrality (η). Therefore, we need to prove that the value of η will converge to zero.

Let η_t denote the centrality value of the preference of the updated strategy s_t in generation t, where $0 \leq \eta \leq 1$. We assume that the algorithm makes some improvement in step t. With the assumption that not all optimization steps fail to improve, we can deduce that

$$\eta_t = \eta_{t-1} - \epsilon, \tag{9}$$

where ϵ is a positive value and $0 < \epsilon \leq \eta_{t-1}$. By further simplifying the equation, we have

$$\eta_t = \eta_{t-1} - \epsilon, \quad = \eta_{t-1} - \alpha_{t-1}\eta_{t-1}, \; = \beta_{t-1}\eta_{t-1}, \tag{10}$$

where $0 < \alpha_{t-1} \leq 1$ and $\beta_{t-1} = 1 - \alpha_{t-1}$.

Assuming that the centrality value of the preference in the initial step is $0 \leq \eta_0 \leq 1$, we can recursively calculate the following formula:

$$\eta_t = \beta_{t-1}\eta_{t-1}, \quad = \beta_{t-1}\beta_{t-2}\eta_{t-2}, \; = \cdots, \quad = \prod_{i=0}^{t-1} \beta_i \times \eta_0. \tag{11}$$

For any $\beta \in \beta_0, \cdots, \beta_{t-1}$, we have $\beta \geq 0$. In addition, we set η_t as a very small positive number if $\eta_t = 0$.

Therefore, by assuming that β is not always equal to zero, we can conclude that η_t will approach zero as outlined in equation 11. Through this proof, we have established that if the assumption that $\lim_{i\to\infty} \mathcal{J}(s_i) \geq \mathcal{J}(s_{i-1})$ holds, the sequence s_i for $i \in \mathbb{N}$ converges to a local optimal strategy s^*, also known as the local best-preferred strategy.

□

C. Proof of Corollary 4.5

Corollary 4.5. *Let $\eta : \mathcal{G}' \to \mathbb{R}^n$ be a function that maps a P-GFG to its in-degree centrality, the convergence rate of the sequence $\{s_i\}$ is Q-sublinear concerning η.*

Proof. In Theorem 4.4, we have proved that the strategies generated by the COLE framework will converge to the local best-preferred strategy. When we use the in-degree centrality function as η, the preference centrality function can be rewritten as:

$$\eta(i) = 1 - \frac{I_i}{n-1}, \tag{12}$$

where I_i is the in-degree of node i and n is the size of the strategy set $\mathcal{N}$. Therefore, we have

$$\begin{aligned}\lim_{t\to\infty}\frac{|\eta_{t+1}-0|}{|\eta_t-0|} &= \lim_{t\to\infty}\frac{\eta_{t+1}}{\eta_t}\\ &= \lim_{t\to\infty}\frac{1-\frac{k_{t+1}}{t}}{1-\frac{k_t}{t-1}}\\ &= \lim_{t\to\infty}\frac{t-1}{t}\frac{t-k_{t+1}}{t-k_t-1}\\ &= \lim_{t\to\infty}\frac{t-k_{t+1}}{t-k_t-1}\\ &=1\end{aligned} \tag{13}$$

Therefore, using the in-degree centrality, we can conclude that the COLE framework will converge to the local optimal strategy at a Q-sublinear rate. □

D. Graphic Shapley Value Solver Algorithm

Algorithm 2 gives the detailed steps of the graphic Shapley value solver in Section 5.2.

Algorithm 2 Graphic Shapley Value Solver Algorithm

Input:: population $\mathcal{N}$, the number of Monte Carlo permutation sampling k, the size of negative population
Initialize $\phi = \mathbf{0}_{|\mathcal{C}|}$
for $(1, 2, \cdots, k)$ **do**
  $\pi \longleftarrow$ *Uniformly sample from* $\Pi_{\mathcal{C}}$, where $\Pi_{\mathcal{C}}$ is permutation set
  for $i \in \mathcal{N}$ **do**
    // Obtain predecessors of player i in sampled permutation π
    $S_\pi(i) \longleftarrow \{j \in \mathcal{N} | \pi(j) < \pi(i)\}$
    // Update incompatibility weights
    $\phi_i \longleftarrow \phi_i + \frac{1}{k}(v(S_\pi(i) \cup \{i\}) - v(S_\pi(i)))$
  end for
end for
$\phi \longleftarrow \phi / \sum \phi$
$\phi \longleftarrow (1-\phi) / \sum (1-\phi)$
Output:: ϕ

E. Overcooked Environment

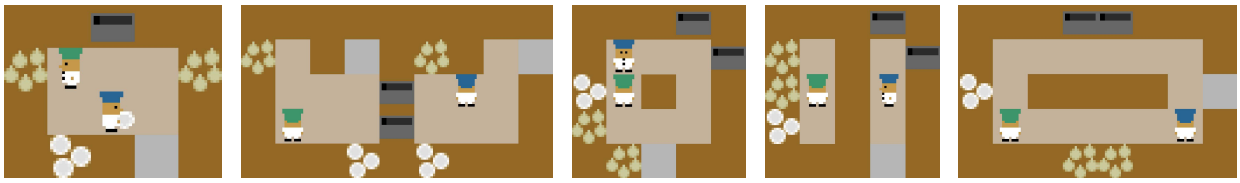
In this paper, we conduct a series of experiments in the Overcooked environment (Carroll et al., 2019; Charakorn et al., 2020; Knott et al., 2021), which is proposed for the coordination challenge, to verify the performance of COLE_{SV}. As a two-player common payoff game, each player controls one chef in a kitchen to cook and serve soup, which results in a reward of 20 for the team. We test our codes on five different layouts: Cramped Room, Asymmetric Advantages, Coordination Ring, Forced Coordination, and Counter Circuit.

The Overcooked environment that we used has five layouts, including **Cramped Room**, **Asymmetric Advantages**, **Coordination Ring**, **Forced Coordination**, and **Counter Circuit**. Screenshots of these layouts can be seen in Fig. 7.

The detailed introduction of five layouts is as follows.

(a) **Cramped Room**. The cramped room is a simple environment where two players are limited to a small room with only one pot (black box with gray bottom) and one serving spot (light gray square). Therefore, players are expected to fully utilize the pot and effectively deliver soup, even with basic coordination.

(b) **Asymmetric Advantages**. In this layout, two players are placed in two disconnected kitchens. As the name suggests, the positions of onions, pots, and serving spots are asymmetric. In the left kitchen, onions are far from the pots, while



(a) Cramped Room (b) Asymmetric Advantages (c) Coordination Ring (d) Forced Coordination (e) Counter Circuit

Figure 7. Overcooked environment layouts.

serving spots are near the middle area of the layout. However, in the right kitchen, onions are placed near the middle area and the serving areas are far from the pots.

(c) **Coordination Ring**. This ring-like layout requires both players to keep moving to prevent blocking each other, especially in the top-right and bottom-left corners where the onions and pots are located. For optimal cooperation, both pots should be utilized.

(d) **Forced Coordination**. The Forced Coordination is another layout that separates the two agents. There are no pots or serving spots on the left side, nor are there onions or pots on the right side. Therefore, two players must coordinate with each other to complete the task. The left player is expected to prepare onions and plates while the right player cooks and serves them.

(e) **Counter Circuit**. The Counter Circuit is another ring-like layout but larger in map size. In this layout, pots, onions, plates, and serving spots are placed in four different directions. Limited by the narrow aisles, players are easily blocked. Therefore, coordinating and performing the task is difficult in this environment. Players need to learn the advanced technique of putting onions in the middle area to pass them to the other quickly, which can further improve performance.

F. Experimental Details of COLE$_{\text{SV}}$

This paper utilizes Proximal Policy Optimization (PPO) (Schulman et al., 2017) as the oracle algorithm for our strategy set $\mathcal{N}$, which consists of convolutional neural network parameterized strategies. Each network is composed of 3 convolution layers with 25 filters and 3 fully-connected layers with 64 hidden neurons. To manage computational resources, we maintain a population size of 50 strategies. In instances where the population exceeds this limit, we randomly select one of the earliest 10 strategies for removal.

We run and evaluate all our experiments on Linux servers, which include two types of nodes: 1) 1-GPU node with NVIDIA GeForce 3090Ti 24G as GPU and AMD EPYC 7H12 64-Core Processor as CPU, 2) 2-GPUs node with GeForce RTX 3090 24G as GPU and AMD Ryzen Threadripper 3970X 32-Core Processor as CPU. On the Overcooked game environment, the COLE$_{\text{SV}}$ takes about one to two days on the 2-GPUs machine for one layout's training.

The hyperparameter setup is similar to those in PBT and MEP, which are given as follows.

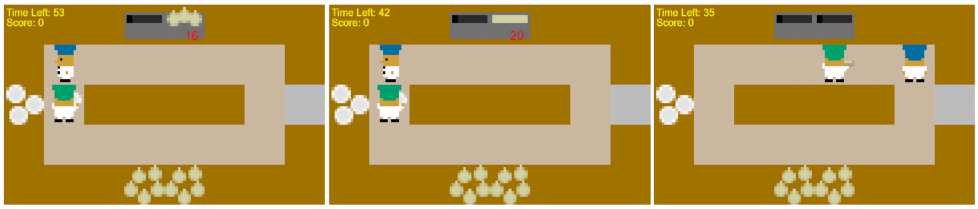
- The learning rate for each layout is 2e-3 , 1e-3 , 6e-4 , 8e-4 , and 8e-4.
- The gamma γ is 0.99.
- The lambda λ is 0.98.
- The PPO clipping factor is 0.05.
- The VF coefficient is 0.5.
- The maximum gradient norm is 0.1.
- The total training time steps for each PPO update is 48000, divided into 10 mini-batches.
- The total numbers of generations for each layout are 80, 60, 75, 70, and 70, respectively.
- For each generation, we update 10 times to approximate the best-preferred strategy.
- The α is 1.

G. Implementations of Baselines

In this part, we will introduce the detailed implementations of baselines. We train and evaluate the self-play and PBT based on the Human-Aware Reinforcement Learning repository (Carroll et al., 2019) [1] and used Proximal Policy Optimization (PPO) (Schulman et al., 2017) as the RL algorithm. We implement FCP according to the FCP paper (Strouse et al., 2021) and use PPO as the RL algorithm. The implementation is based on the Human-Aware Reinforcement Learning repository (the same used in the self-paly and PBT). The MEP agent is trained with population size as 5, following the MEP paper (Zhao et al., 2021) and used the original implementation[2].

H. Trajectory Visualization

We visualize the trajectories produced by COLE_{SV} 1:3 and 0:4 with middle-level and expert partners in Overcooked at https://sites.google.com/view/cole-2023/. Fig. 8 presents three screenshots of the COLE_{SV} 0:4 model (blue player) that collaborates with one of the expert partners, the PBT model (green player). The case illustrates the importance of the individual objective in zero-shot coordination with expert partners. Frame A is a screenshot taken at 53s when the two players start to impede each other. The PBT model has taken the plate and wants to load and serve the dish. The blue player wants to take the plate but does not know how to change the objective to allow the green player to load the dish. After blocking for about 11s, the blue player starts to move and lets the green player go to the pots (Frame B). However, the process is not smooth and takes 7s to reach Frame C. This phenomenon does not occur in COLE_{SV} 1:3 coordination with expert partners, which shows that including individual objectives might improve the cooperative ability with expert partners.



Frame A : time left 53s **Frame B : time left 42s** **Frame C : time left 35s**

Figure 8. Trajectory snapshots of the COLE_{SV} 0:4 model (blue) with one of the expert partners - PBT model (green).

[1] https://github.com/HumanCompatibleAI/human_aware_rl/tree/neurips2019.

[2] The code of MEP original implementation: https://github.com/ruizhaogit/maximum_entropy_population_based_training.