Journal of Artificial Intelligence Research 1 (2024) 1139-1185 Submitted 01/2024; published 07/2024

Tackling Cooperative Incompatibility for Zero-Shot Human-AI Coordination

Yang Li YANG.LI-4@MANCHESTER.AC.UK
The University of Manchester

Shao Zhang SHAOZHANG@SJTU.EDU.CN
Jichen Sun SUNJICHEN@SJTU.EDU.CN
Wenhao Zhang WENHAO_ZHANG@SJTU.EDU.CN
Shanghai Jiao Tong University

Yali Du YALI.DU@KCL.AC.UK
King's College London

Ying Wen YING.WEN@SJTU.EDU.CN
(Corresponding author)
Xinbing Wang XWANG8@SJTU.EDU.CN
Shanghai Jiao Tong University

Wei Pan WEI.PAN@MANCHESTER.AC.UK
(Corresponding author)
The University of Manchester

Abstract

Securing coordination between AI agent and teammates (human players or AI agents) in contexts involving unfamiliar humans continues to pose a significant challenge in Zero-Shot Coordination. The issue of cooperative incompatibility becomes particularly prominent when an AI agent is unsuccessful in synchronizing with certain previously unknown partners. Traditional algorithms have aimed to collaborate with partners by optimizing fixed objectives within a population, fostering diversity in strategies and behaviors. However, these techniques may lead to learning loss and an inability to cooperate with specific strategies within the population, a phenomenon named cooperative incompatibility in learning. In order to solve cooperative incompatibility in learning and effectively address the problem in the context of ZSC, we introduce the **C**ooperative **O**pen-ended **LE**arning (**COLE**) framework, which formulates open-ended objectives in cooperative games with two players using perspectives of graph theory to evaluate and pinpoint the cooperative capacity of each strategy. We present two practical algorithms, specifically COLE_{SV} and COLE_{R}, which incorporate insights from game theory and graph theory. We also show that COLE could effectively overcome the cooperative incompatibility from theoretical and empirical analysis. Subsequently, we created an online Overcooked human-AI experiment platform, the COLE platform, which enables easy customization of questionnaires, model weights, and other aspects. Utilizing the COLE platform, we enlist 130 participants for human experiments. Our findings reveal a preference for our approach over state-of-the-art methods using a variety of subjective metrics. Moreover, objective experimental outcomes in the Overcooked game environment indicate that our method surpasses existing ones when coordinating with previously unencountered AI agents and the human proxy model. Our code and demo are publicly available at `https://sites.google.com/view/cole-2023`.

1. Introduction

Significant advancements in artificial intelligence (AI) research have led to groundbreaking solutions that outperform humans in various tasks (Silver et al., 2018; Kirillov et al., 2023; Peng et al., 2017; Li et al., 2023b). However, in the real-world scenario, the advanced AI need to collaborate with human (Dafoe et al., 2021; Zhang et al., 2023, 2024). Establishing efficient collaboration between AI and unseen partners (either human players or AI agents), a concept known as zero-shot coordination (ZSC), continues to pose a significant challenge (Legg & Hutter, 2007; Hu et al., 2020; De Peuter & Kaski, 2022). The significance of zero-shot human-AI coordination becomes evident in various real-world applications, including manufacturing (Li et al., 2023a), autonomous vehicles (Aoki et al., 2021), and assistant robots (de Berardinis et al., 2020). The issue of *"cooperative incompatibility"* becomes particularly prominent when an AI agent is unsuccessful in synchronizing with certain previously unknown partners. As exemplified by multiplayer video games like Honor of King (Wei et al., 2022) and Overcooked (Carroll et al., 2020), the constant requirement for players to collaborate with unseen partners is evident. However, AI agents, which are trained via the maximization of rewards, tend to exhibit deterministic strategies and differ from those of human players or other AI agents, leading to a failure in coordinating with some unseen players (Ye et al., 2020; Gao et al., 2023; Yu et al., 2023).

One of the conventional methods to solve zero-shot human-AI coordination is self-play (Tesauro, 1994), which involves an iterative strategy refinement process through self-competition. Although SP can achieve equilibrium in a game (Fudenberg & Levine, 1998), strategies often develop specific behaviors and conventions to secure higher payoffs (Hu et al., 2020). As a result, a fully-converged SP strategy may face difficulties adapting to coordination with previously unencountered strategies and humans (Lerer & Peysakhovich, 2018; Hu et al., 2020; Zhao et al., 2021). To address SP's limitations, current zero-shot coordination (ZSC) approaches concentrate on enhancing strategic or behavioral diversity by incorporating population-based training (PBT) to improve adaptability (Carroll et al., 2020; Canaan et al., 2022; Zhao et al., 2021; Lupu et al., 2021; Charakorn et al., 2023). PBT aims to boost cooperative performance with other strategies in the population, fostering zero-shot coordination with unknown strategies, which is achieved by maintaining a set of strategies to disrupt SP conventions (Tesauro, 1994) and optimizing the rewards for each pair within the population. Most state-of-the-art (SOTA) methods focus on pre-training diverse populations (Strouse et al., 2021; Lupu et al., 2021) or introducing handcrafted techniques (Canaan et al., 2022; Zhao et al., 2021) to excel at cooperative games by optimizing fixed objectives within the population. And a recent work LIPO (Charakorn et al., 2023) pursues solutions compatible with their partner agents but incompatible with others in the population. These approaches have effectively tackled complex cooperative tasks, such as Overcooked (Carroll et al., 2020) and Hanabi (Bard et al., 2020).

However, optimizing a fixed population-level objective, such as maximizing expected rewards within the population (Strouse et al., 2021; Lupu et al., 2021; Zhao et al., 2021) or maximizing self-reward, but minimizing the reward with other agents (Charakorn et al., 2023), does not guarantee improved coordination capabilities for strategies within the population. Specifically, although overall performance may improve, simultaneous promotion of coordination abilities within the population may not occur. This phenomenon, which we

refer to as "*cooperative incompatibility in learning*", underscores the need to carefully weigh the trade-offs between overall performance and coordination ability when optimizing a fixed population-level objective.

To more effectively describe and formulate the cooperative incompatibility problem, we introduce Graphic-Form Games (GFGs) to reframe cooperative tasks from the perspective of norm-form games. In a GFG, strategies are depicted as nodes, with edge weights between nodes representing the corresponding cooperative utility of the connected strategies. Additionally, we derive Preference GFGs (P-GFGs) to profile each node's preferred counterpart in the population, where "prefer" signifies that a node achieves a higher score when cooperating with the preferred node rather than its neighbors. Using (P-)GFGs, cooperative incompatibility in a learning algorithm can be assessed based on the extent to which others prefer the updated strategy.

To address cooperative incompatibility, we propose the Cooperative Open-ended LEarning (**COLE**) framework, which iteratively generates a new strategy that approximates the best response to empirical gamescapes within P-GFGs. We have shown that COLE framework converges to the local best-preferred strategy at a Q-sublinear rate when using in-degree centrality as the preference evaluation metric. We propose two practical algorithms, COLE_{SV} and COLE_{R}, both of which comprise of a simulator, a solver, and a trainer that are incorporated to excel in two-player cooperative game Overcooked (Carroll et al., 2020). The primary distinguishing factor between COLE_{SV} and COLE_{R} is the solver component. While COLE_{SV} employs the intuitive concept of the Shapley value to evaluate the adaptability of strategies and ascertain the cooperative incompatibility distribution, COLE_{R} takes a different approach. It supplants the Shapley value with the average payoffs associated with nodes in the population as a measure of cooperative capability. The trainer aims to approximate the best responses to the cooperative incompatibility distribution mixture found in the population.

The prevailing literature of zero-shot human-AI collaboration largely concentrates on the quantitative assessment, indicated in the body of works (Carroll et al., 2020; Strouse et al., 2021; Yu et al., 2023), and often prioritizing metrics such as episode rewards and frequencies. Some methods conducted human-AI experiments to verify the zero-shot coordination performance with human players (Strouse et al., 2021; Lou et al., 2023). However, while these investigations provide valuable insights, most of these methods overlook more comprehensive aspects of the subject and lack a unified evaluation framework. These oversights accentuate the existing deficiencies in the contemporary experimental environments, which necessitates the urgency to introduce our proposed experimental framework. For the robust evaluation of human-AI collaboration performance, we devised an online Overcooked human-AI experimental pipeline, which is a cooperative task environment for two players as referenced in (Carroll et al., 2020). Our evaluation approach incorporates a broad spectrum of subjective metrics that assess individual games involving AI agents while implementing a comprehensive comparison along the entire participation. Primarily, these encompassing metrics appraise aspects such as intention, contribution, and teamwork during collaboration with AI agents. Besides, they extend the evaluation to encompass fluency, preference, and understanding across all the collaborated agents. Moreover, we developed all experimental components including ethic agreements, instructions, questionnaires, model weights into a unified experimental platform which we dubbed as the COLE-Platform. To the best of

our knowledge, our open-source human-AI experiment pipeline is the first comprehensive human-AI experimentation pipeline for zero-shot human-AI coordination evaluation including turnkey experimental procedures and scale design.

In this paper, we present the findings of a human-agent study involving 130 participants conducted on the COLE platform to assess the performance of our proposed algorithm. Participants were tasked with evaluating the AI agent's performance based on three criteria: the human's comprehension of AI intentions, AI's contribution to the task, and the effectiveness of human-AI collaboration. Furthermore, each participant compared the COLE platform with another randomly assigned AI agent, ranking the two agents in collaborative fluency, personal preference, and mutual understanding. Our findings demonstrate a clear preference for our algorithm over baseline approaches in various evaluation metrics. In addition to objective evaluations, we also examined zero-shot human-AI coordination performance by testing our algorithm with previously unencountered baseline agents and a human proxy model. This approach helps to further demonstrate the adaptability and effectiveness of our algorithm in novel collaborative scenarios. The human proxy model is a widely used behavior cloning model (Carroll et al., 2020). The results of the experiment demonstrate that COLE_{SV} surpasses the recent SOTA methods in both evaluation protocols. Furthermore, analysis of GFGs and P-GFGs during the learning process of COLE_{SV} reveals that the framework effectively overcomes cooperative incompatibility.

This work represents an extension of our conference paper (Li et al., 2023c). Significant enhancements incorporated in this study include extending the concept of cooperative incompatibility to include zero-shot human-AI coordination, introducing a comprehensive human-AI experimental pipeline, providing new theoretical analysis to investigate COLE, proposing an extra practical algorithm, and executing a broader set of experiments. The primary contributions in this paper can be summarized as follows.

- We introduce 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 propose 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 local best-preferred strategy, and the convergence rate will be Q-sublinear when using in-degree preference centrality.

- To the best of our knowledge, we propose the first comprehensive human-AI experimentation pipeline for zero-shot human-AI coordination evaluation including turnkey experimental procedures and scale design. And we conducted a human-AI experiment involving 130 participants, and the results demonstrate a preference for our COLE_{SV} over baselines using various subjective metrics. Additional objective experiments confirm the effectiveness of our proposed algorithm compared to SOTAs.

The remainder of this paper is structured as follows: Sections 2 and 3 present related works and preliminaries, while Section 4 formalizes GFGs, and related concepts and introduces the COLE framework. Section 5 describes two proposed practical algorithms COLE_{R} and COLE_{SV}. The COLE human-AI experiment pipeline is detailed in Section 6, and Section 7 outlines the experimental settings and results. Finally, we summarize our findings, limitations,

and future works in Section 9. Additionally, the Appendix offers supplementary information and in-depth proofs, including visual overviews of the human-AI experiment platform.

2. Related Works

Zero-Shot Human-AI Coordination. The zero-shot human-AI coordination is closely related to zero-shot coordination (ZSC) (Hu et al., 2020; Wang et al., 2024). ZSC aims to train a strategy to coordinate effectively with unseen partners (Hu et al., 2020). Self-play (Tesauro, 1994; Carroll et al., 2020; Wang et al., 2023) 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., 2020), which trains strategies by interacting with each other in a population. However, PBT does not explicitly maintain diversity and therefore does not coordinate with unseen partners(Strouse et al., 2021).

To achieve the goal of ZSC, 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). 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. The most recent work of LIPO (Charakorn et al., 2023) pursues solutions compatible with their partner agents but incompatible with others in the population. LIPO is also based on the framework of the PBT algorithm, whose objective function consists of two fixed teams: maximizing the rewards when it pairs with itself, but minimizing the rewards when it pairs with other agents in the population. Our previous research (Li et al., 2023c) emphasized the collaboration with unfamiliar AI agents through the continuous formulation and optimization of objectives. In this study, we expand the scope of COLE framework to facilitate cooperation with actual human players, which is validated by human players on the human-AI experimental platform we have developed.

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; Wen et al., 2024). 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., 2017a; 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) uses open-ended learning by maintaining two populations of

strategies and partners and 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 GFG and P-GFG in open-ended learning to evaluate and identify the cooperative ability of strategies to solve cooperative incompatibility efficiently with theoretical guarantee.

Coordination Graphs. Another domain closely associated with our research is the field of coordination graphs (CGs) (Guestrin et al., 2002). CGs offer an influential framework for modeling intricate interactions and dependencies among multiple agents, facilitating the representation and examination of cooperative problem-solving scenarios. Typically, CGs comprise a vertex (agent) set and an edge set, with each state potentially possessing its distinct coordination graph(Guestrin et al., 2002). Deep CGs (Böhmer et al., 2020) refine the joint value function of all agents per coordination graph, factoring payoffs between pairs of agents to strike a flexible balance between representational capacity and generalization. In ad hoc teamwork, Deep CGs have been employed by Rahman et al. (2021), albeit under full observability. The Generalized Policy Learning (GPL) approach (Rahman et al., 2021) leverages Deep CGs in ad hoc teamwork to learn agent models and joint-action value models in the face of varying team compositions. Moreover, the Partially Observable GPL (PO-GPL) method (Rahman et al., 2022) broadens the application of GPL by extending it from the fully observable problem domain to the partially observable one. In contrast to CGs and subsequent approaches, our graph-based games conceptualize the cooperative task from the standpoint of the normal-form game rather than the stochastic game.

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 $a_1, a_2 \in \mathcal{A}$.

Open-Ended Meta-Games and Analysis. A meta-game is characterized by the tuple $(\mathcal{N}, \mathcal{S}, \mathcal{M})$, where $\mathcal{N}$ represents the set of players, $\mathcal{S}$ is a set of policies (like a set of RL models), and $\mathcal{M}$ designates the meta-game payoff matrix. In common payoff cooperative meta-game with two players, the meta-game payoff is $\mathcal{M} = \left\{\phi(S_1, S_2) : (S_1, S_2) \in \mathcal{S}^1 \times \mathcal{S}^2\right\}$, where $\phi : \mathcal{S}^1 \times \mathcal{S}^2 \to \mathbb{R}$ is the utility function. The meta-game payoff is a symmetric matrix. The primary distinction between a meta-game and a normal-form game lies in their respective actions. In the case of a meta-game, a strategy such as a DRL model constitutes an action, as opposed to the atomic actions (e.g., up and down) typical of normal-form games. Open-ended meta-game refers to the continual learning process, wherein new strategies are persistently generated and incorporated into the strategy sets throughout the training phase.

Empirical Game-Theoretic Analysis (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., 2017a). Furthermore, empirical gamescapes (EGS) are defined as the convex hull of the payoff vectors of all strategies (Balduzzi et al., 2019). Given a population $\mathcal{N}$ of n strategies, the empirical gamescapes is often defined as

$$\begin{aligned}\mathcal{G} &= \{\text{convex mixture of rows of } \mathcal{M}\} \\ &= \left\{ \sum_i \alpha_i \cdot \boldsymbol{m}_i : \alpha \geq \mathbf{1}, \alpha^{\mathbf{T}}\mathbf{1} = \mathbf{1}, \mathbf{m_i} = \mathcal{M}_{[\mathbf{i},:]} \right\}\end{aligned} \tag{1}$$

where $\mathcal{M}$ is the empirical payoff matrix with the expected outcomes for each joint strategy.

Cooperative Theoretic Concepts. In the present study, our main focus is on characteristic function games, a particular class of cooperative games in which a coalition's generated value is represented by a characteristic function (Chalkiadakis et al., 2011). 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.

In addition, in this study, we examine transferable utility games (TU games), which are based on the underlying assumption that the coalitional value $v(C)$ can be divided among the members of C in any way agreed upon by the members of the coalition. This assumption allows for greater flexibility in analyzing cooperative problem solving and the distribution of the resulting payoffs.

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

$$SV(i) = \frac{1}{n!} \sum_{\pi \in \Pi_{\mathcal{N}}} v(P_i^\pi \cup \{i\}) - v(P_i^\pi), \tag{2}$$

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.

Graph Theoretic Concepts. In graph theory, a weighted directed graph or network can be represented as $D = (V, E, w)$, consisting of a set of vertices V, a set of directed edges E, and a weight function $w : E \to \mathbb{R}^+$ that assigns positive real numbers to each edge. If the graph has the property that $(v, u) \in E$ entails $(u, v) \in E$ and $w(v, u) = w(u, v)$ for all $(u, v) \in E$, it can be considered an undirected weighted graph, also known as a weight graph. An unweighted graph is a special case of a weighted graph, where the weight function assigns a value of 1 to every edge $(u, v) \in E$.

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. Weighted 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:

$$\sigma(u) = (1-d) + d \sum_{v \in B(u)} \sigma(v) \frac{I_u}{\sum_{p \in R(v)} I_p} \frac{O_u}{\sum_{p \in R(v)} O_p}, \tag{3}$$

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 introduce the cooperative open-ended learning framework designed to address the cooperative incompatibility problem and enhance zero-shot human-AI coordination. In Subsection 4.1, we propose graphic-form games and related concepts to model the coordination relationships in a population of strategies. We then present motivating examples in Subsection 4.2 to illustrate the importance of studying the cooperative incompatibility problem. This is followed by a conceptual definition of cooperative incompatibility in Section 4.3. Finally, we propose the cooperative open-ended learning framework in Section 4.4.

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)**.** Let a collection of strategies $\mathcal{N} = \{1, 2, \cdots, n\}$ be given, where each strategy could be a parameterized network or a human player. A two-player graphic-form game (GFG) can be defined as a triplet $\mathcal{G} = (\mathcal{N}, \mathbf{E}, \mathbf{w})$, which can be represented as a weighted directed graph. Here, $\mathcal{N}$, $\mathbf{E}$, and $\mathbf{w}$ represent the sets of nodes, edges, and weights, respectively. For a given edge (i, j), $\mathbf{w}(i, j)$ denotes the anticipated outcome when strategy i competes against strategy j. The visual representation of a GFG is referred to as a game graph.

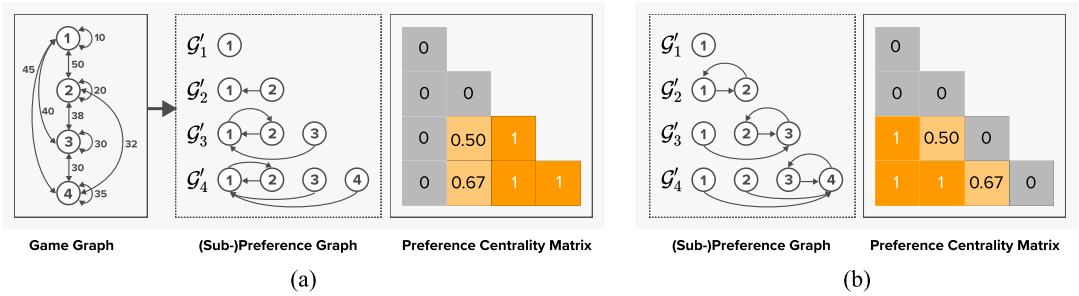


Figure 1: The Game Graph, (sub-) preference graph and corresponding preference centrality matrix. The (sub-) preference graphs are for all four iterations in the training process, and the corresponding preference in-degree centrality matrix is based on them. As it can be observed in $\mathcal{G}'_3$ and $\mathcal{G}'_4$ in *(a)*, 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 than all previous strategies.

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. Given a GFG $(\mathcal{N}, \mathbf{E}, \mathbf{w})$, the 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. The game graph delineates the interaction of players within a weighted directed graph, in which the weight signifies the payoff or utility derived from two participating agents. However, the preference graph pertains to the depiction of each agent identifying the partner with whom they can achieve the most significant utility in the population. This is the underlying principle behind the term "preference": an agent demonstrates a strong tendency to cooperate with a particular endpoint agent to maximize their utility.

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

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

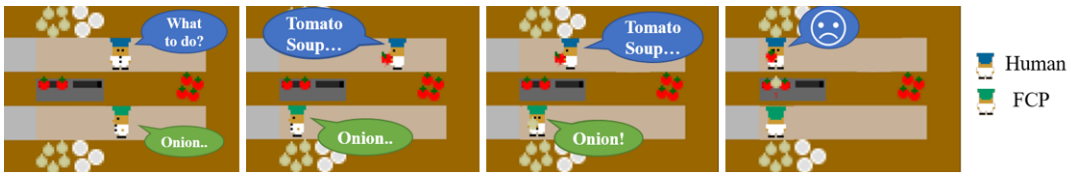
where d_i is a graph centrality metric to evaluate how the node is preferred, and norm := $\mathbb{R} \to [0, 1]$ is a normalization function.

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. In this work, our primary choice for the graph centrality metric is in-degree centrality. Besides, the degree centrality values are normalized (norm) by dividing by the maximum possible degree $n - 1$ where n is the number of nodes.

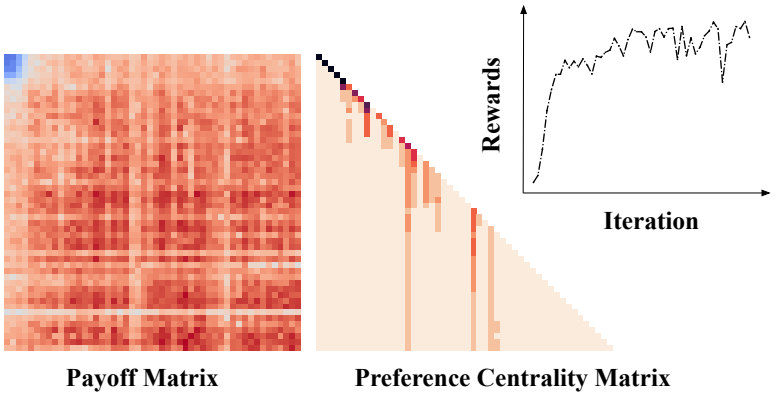
Fig. 1 provides examples of the learning processes of two algorithms, one with and one without a cooperative incompatibility issue. In Fig. 1(a), illustrative of the algorithm possessing this cooperative incompatibility, the updated strategy is observed to fall short in elevating the cooperative utility achieved with other strategies after the second generation. There are no edges pointing towards the node which validates this further, and is confirmed by the preference centrality matrix, where the strategy at hand is marked with a $\eta = 1$. This suggests that no nodes wish to collaborate with the updated strategies. On the other hand, Fig. 1(b), devoid of cooperative incompatibility, portrays a different scenario. In this depiction, every other strategy within the preference graph directs towards the most recent strategy, signifying the continuous enhancement of the cooperative capability.

4.2 Motivating Examples

The challenge of training AI agents to effectively coordinate with unfamiliar AI counterparts or human players constitutes a substantial impediment. Predominantly trained through reward maximization methods, AI agents typically demonstrate deterministic strategies which often diverge from those employed by human players or other AI agents, thereby resulting in a coordination failure with previously unseen players (Ye et al., 2020; Gao et al., 2023; Yu et al., 2023). Fig. 2(a), proposed by HSP (Yu et al., 2023), illustrates a motivating example in the Overcooked game of how an agent player could collaborate with some players and fail to collaborate with other players. In this example of the Overcooked game, chefs could accomplish two recipes: one calls for three onions and the other demands three tomatoes. The FCP algorithm, however, tends to settle into a distinctive pattern of cooking solely onion soup (Yu et al., 2023). Consequently, this FCP agent can achieve high scores when paired with similarly styled human players. Nevertheless, its performance declines significantly when paired up with a player who prefers the tomato soup style, resulting in a disruption of the human player's tomato soup cooking plan, thereby failing to complete the required recipe. Further instances of miscoordination can be observed in real-world scenarios, such as autonomous driving, where the question may arise whether the policy should be explorative or conservative.



(a) The illustration showcases an instance of an FCP agent's collaboration failure with specified human participants. This figure is referenced from HSP (Yu et al., 2023).



(b) Cooperative incompatibility issue in MEP training process.

Figure 2: Motivating examples. Fig. 2(a) features an analysis conducted by Hidden-utility Self-Play (HSP) method (Yu et al., 2023). The FCP agent converges to a fixed pattern of exclusively preparing onion soup, thereby failing to establish coordination with a human participant who prefers making tomato soup. Fig. 2(b) shows a training process of the MEP algorithm with several comparative incompatibility problems. The payoff matrix of each strategy during training and the corresponding preference centrality matrix of the MEP algorithm in the Overcooked. A deeper shade of red in the payoff matrix signifies higher utility. The darker the color in the preference centrality matrix, the lower the centrality value, and the more other strategies prefer it.

Nevertheless, miscoordination issues exist in the training of some population-based ZSC algorithms. An analysis of the MEP algorithm (Zhao et al., 2021), as depicted in Fig. 2(b), reveals a cooperative incompatibility evident during the learning process in the Overcooked environment (Carroll et al., 2020). The figure on the left represents the payoff matrix, where a deeper shade of red signifies higher utility. In the accompanying preference indegree centrality matrix, a darker color is indicative of a strategy being more preferred by others. Throughout the MEP learning process, despite a consistent improvement in mean rewards (as depicted in the upper-right of Fig. 2(b)), significant cooperative incompatibility problems arise post a certain training duration. At this juncture, many strategies show an inclination to engage with earlier strategies, represented by darker shades, over newer strategies, in an effort to secure higher rewards. Therefore, addressing this collaboration incompatibility is crucial to enhance the adaptability of AI agents in coordinating with unseen partners.

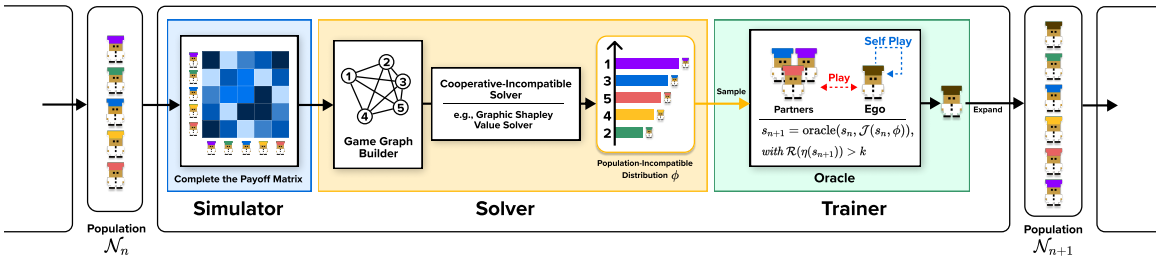


Figure 3: An overview of one generation in COLE framework: The solver derives the cooperative incompatible distribution ϕ using a cooperative incompatibility solver, which can be any algorithm that evaluates cooperative contribution. The trainer then approximates the relaxed best response by optimizing 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.

4.3 Cooperative Incompatibility

Fig. 1 provides examples of cooperative incompatibility, specifically concerning human-AI coordination and ZSC algorithm learning. Consequently, this section will delve deeper into a conceptual understanding of cooperative incompatibility and formally delineate this issue within the learning process, utilizing the tools previously introduced such as P-GFGs and preference centrality. The conceptual definition of cooperative incompatibility is as follows:

Definition 4.3 (Cooperative Incompatibility)**.** Cooperative incompatibility pertains to the occurrence wherein a participant, either an AI agent or a human player, is unable to align with certain specific partners, who may equally be AI agents or human players.

Cooperative Incompatibility in the learning of ZSC algorithms. A cooperative incompatibility issue arises in a population-based learning algorithm when the new strategy s_t, produced by the algorithm at step t, is unable to boost cooperative utility in conjunction with existing strategies in the population. This is to say, the preference centrality $\eta(t)$ of s_t is greater than 0. Ideally, an algorithm devoid of cooperative incompatibility issues will maintain a preference centrality of 0 at each step. This implies that the updated strategy enhances the collaborative utility of strategies within the population, prompting them to prefer cooperation with the updated strategy.

In this study, our aim is to resolve the cooperative incompatibility in the learning process of population-based algorithms, ultimately paving the way to tackle cooperative incompatibility issues in AI-AI and AI-human coordination.

4.4 Cooperative Open-Ended Learning Framework

To address the problem of cooperative incompatibility, we refine our comprehension of the objective by formulating empirical gamescapes (Balduzzi et al., 2019) from zero-sum games

to our studied shared payoff GFGs. This provides us with a geometric representation of player strategies within a GFG $\{\mathcal{N}, \mathbf{E}, \mathbf{w}\}$, thereby capturing the diversity and adaptive capacity of cooperative strategic behaviour. However, it is not efficient to directly learn how to extend the EGS in common payoff games to cooperate effectively with unseen partners.

To conquer cooperative incompatibility, the natural idea is to learn with the mixture of cooperative incompatible strategies on the most recent population $\mathcal{N}$ to improve gamescape. 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. In general, approximate oracles often employ techniques such as approximation methods (e.g., function approximation or neural networks), reinforcement learning (RL) algorithms, and optimization techniques to iteratively improve the agent's policy until it converges to a satisfactory solution (McMahan et al., 2003; Balduzzi et al., 2019). In this study, the approximate oracle functions as a solver utilizing RL algorithms to efficiently learn and approximate the best-response policy. Given strategy s_n, the oracle returns a new strategy s_{n+1} : $s_{n+1} = \text{oracle}(s_n, \mathcal{J}(s_n, \phi_n))$, with $\eta(s_{n+1}) = 0$, if possible. $\mathcal{J}$ is the objective function as follows,

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

Here, α represents the balancing hyperparameter, while ϕ_n denotes the cooperative incompatibility distribution calculated by the solver at generation n, in which strategies that fail to coordinate with others are assigned higher probabilities. 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 best-preferred strategy if $\eta(s_{n+1}) = 0$.

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

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

where $\mathcal{R}(\cdot)$ serves as the ranking function, wherein a lower preference centrality value corresponds to higher ranks. ϕ_n refers to the distribution at generation n. For simplicity, we may forego the use of the subscript n in the remainder of the paper. As illustrated by the equation, the terminal condition for a single generation of oracle training has been moderated to satisfy the criterion wherein the rank of preference centrality of optimized strategy resides within the top k. Instead of calculating best-preferred strategy, the approximate oracle aims to output best-k-preferred strategy to the IPI mixture.

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

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 at derive the cooperative incompatible distribution of the Game Graph builder and the cooperative incompatible solver. The trainer uses the oracle to approximate the best-preferred strategy to the cooperative incompatible mixture and individual objective and outputs a newly generated strategy which is added to the population for the next generation.

Although we relax the 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 convergence theorem of the local best preferred strategy is given as follows.

Theorem 4.4. *Let $s_0 \in \mathcal{S}$ be the initial strategy and $s_i = \text{oracle}(s_{i-1}, \mathcal{J}(s_{i-1}, \phi_{i-1}))$ for $i \in \mathbb{N}$. Under the effective functioning of the approximated oracle as characterized by Eq. 5, we can say that the sequence $\{s_i\}$ for $i \in \mathbb{N}$ could converge to a local optimal strategy s^*, i.e., the local best-preferred strategy.*

Proof. See Appendix A. □

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

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 B. □

5. Practical Algorithm

In an endeavour to tackle cooperative incompatibility issues in common-payoff games with two players, we have developed two practical algorithms, COLE_{R} and COLE_{SV}. These are based on the COLE framework, and are specifically designed to reconcile cooperative incompatibility and augment zero-shot coordination capabilities. As depicted in Fig. 3, these algorithms, at each generation, accept an input population $\mathcal{N}$ and from it, derive a local best-preferred strategy that is then appended to $\mathcal{N}$ in order to extend the population. The process of generating this strategy involves the collaboration of the simulator, solver, and trainer modules. While COLE_{SV} and COLE_{R} utilize a common simulator and trainer, they each employ a distinct solver. Computation of the payoff matrix $\mathcal{M}$ for the input population $\mathcal{N}$ is carried out by the simulator, with each element $\mathcal{M}(i, j)$ where $i, j \in \mathcal{N}$ symbolizing the cumulative rewards of players i and j at both respective starting positions. The solver's role involves assessing and identifying strategies that have failed to collaborate effectively, achieved through the calculation of an incompatible cooperative distribution. A Graphic Shapley Value solver is integrated into COLE_{SV} enabling it to measure the cooperative efficacy of each strategy relative to all others. This is achieved by implementing the weighted PageRank (WPG) (Xing & Ghorbani, 2004) from graph theory into the Shapley Value. This technique allows for the evaluation of adaptability, particularly in relation to those strategies that have failed to collaborate effectively. As a contrast, COLE_{R} incorporates a simplified

Algorithm 1 Practical Algorithms

Input:population $\mathcal{N}_0$, the sample times a, b of $\mathcal{J}_i, \mathcal{J}_c$, hyperparameters α, k, solver flag $FLAG$
for $t = 1, 2, \cdots$, **do**
  /* Step 1: Completing the payoff matrix */
  $\mathcal{M}_n \leftarrow$ Simulator($\mathcal{N}_t$)
  /* Step 2: Solving the cooperative incompatibility distribution */
  if FLAG is "SV" **then**
    /* Selecting Graphic Shapley Value Solver */
    $\phi =$ Graphic Shapley Value($\mathcal{N}_t$) by Algorithm 2
  else
    if FLAG is "R" **then**
      /* Selecting Reward Solver */
      $\phi =$ Reward Solver($\mathcal{N}_t$)
    end if
  end if
  /* Step 3: Approximate the 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
  $s_{t+1} =$ oracle$(s_t, \mathcal{J}_t)$ 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

solver which computes the cooperative ability by determining the payoff with other members of the population. Subsequently, the trainer approximates the strategy that is most preferred against the recently updated population.

5.1 Solvers

Graphic Shapley Value Solver for COLE$_{\mathbf{SV}}$. The graphic Shapley value solver is proposed to calculate the cooperative incompatible distribution as a mixture to approximate the best-preferred strategies in the recent population and overcome cooperative incompatibility. Specifically, we combine the Shapley Value (Shapley, 1971) solution, an efficient single solution concept for cooperative games to assign the obtained team value between 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: $v(C) = \mathbb{E}_{i\sim C, j\sim C}\sigma(i)\sigma(j)\mathbf{w}(i, j)$, 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. The formula

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 $\phi = \mathbf{0}_{|\mathcal{N}|}$
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

of WPG is given as follows:

$$\hat{\sigma}(u) = (1 - d) + d \sum_{v \in B(u)} \hat{\sigma}(v) \frac{I_u}{\sum_{p \in R(v)} I_p} \frac{O_u}{\sum_{p \in R(v)} O_p}, \tag{6}$$

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. Therefore, the metric σ evaluates how unpopular a node is and equals the inverse of the WPG value $\hat{\sigma}$.

Then we calculate the Shapley Value of each node by taking a characteristic function in equation 2, named the graphic Shapley Value. We use Monte Carlo permutation sampling (Castro et al., 2009) to approximate the Shapley Value, which can reduce the computational complexity from exponential 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. The details are given in Algorithm 2.

Reward Solver for COLE$_{\mathbf{R}}$. COLE$_{\mathrm{R}}$ substitutes the Shapley value for WPG weighted average rewards within the population. Specifically, for each player $i \in \mathcal{N}$, we have

$$\phi_i = \sum_{j \in \mathcal{N}} \sigma(i)\sigma(j)\mathbf{w}(i, j), \tag{7}$$

where $\mathbf{w}(i, j)$ is the weight in the game graph, i.e., average payoffs of player i and j. σ is calculated as same as the calculation of graphic Shapley value solver for COLE$_{\mathrm{SV}}$. The remainder of the reward solver processes are identical to the graphic Shapley value solver.

5.2 Trainer: Approximating Best-Preferred Strategy

The trainer inputs the cooperative incompatible distribution ϕ and samples its teammates to learn to approach the best-preferred strategy against the IPI mixture.

Recall the oracle for s_n : $s_{n+1} = \text{oracle}(s_n, \mathcal{J}(s_n, \phi_n))$, with $\mathcal{R}(\eta(s_{n+1})) > k$. COLE$_{\text{SV}}$ aims to optimize the best-preferred strategy over the IPI mixture. $\mathcal{J}(s_n, \phi_n)$ is the joint objective that consists of individual and cooperative compatible objectives. The individual objective aims to improve performance within itself and promote adaptive ability with expert partners, formulated as follows: $\mathcal{J}_i(s_n) = \mathbf{w}(s_n, s_n)$, where s_n is the ego strategy that needs to be optimized 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 follows: $\mathcal{J}_c = \sum_1^b \phi(p^i)\mathbf{w}(s_t, p^i)$, 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) which 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 could efficiently prevent COLE$_{\text{SV}}$ from generating data with a few fixed strategies that did not cooperate, which could lead to loss of adaptive ability.

We conclude COLE$_{\text{SV}}$ as Algorithm 1. Furthermore, to verify the influence of different ratios of two objectives, we denote COLE$_{\text{SV}}$ with different ratios as 0:4, 1:3, 2:2, and 3:1. Specifically, COLE$_{\text{SV}}$ 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$_{\text{SV}}$ 1:3 trains using self-play once and sampling from the cooperative incompatible distribution as partners three times to generate data and update the objectives.

6. Human-AI Experiment Pipeline

In this section, we present our Overcooked human-AI experiment pipeline, explicitly crafted for a comprehensive and streamlined assessment of the cooperative performance of AI agents, working in collaboration with novice human players. Fig. 4 illustrates the proposed experimental framework integrating human and AI participants, encompassing six crucial sequential stages. The crux of this pipeline lies in its human-AI evaluation design, integrating a broad array of subjective metrics. These metrics scrutinize individual games involving AI agents, along with providing a comprehensive comparative analysis across the entire participant pool. These encompassing metrics primarily evaluate vital aspects such as intentionality, contribution, and teamwork within human-AI collaborations. Moreover, these metrics facili-

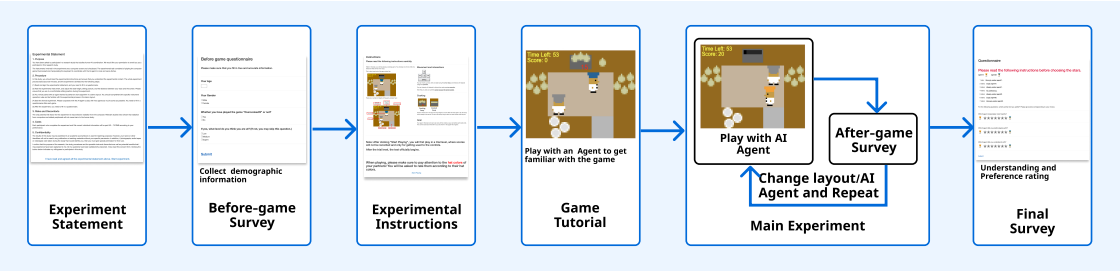


Figure 4: The illustrated figure characterizes the conceptual architecture of the proposed human-AI experimental pipeline, structured around six critical stages. **(Stage 1) Experiment Statement** delineates the nature of the experiment, associated risks, and ethical considerations among other relevant aspects. **(Stage 2) Before-game Survey** principally focuses on the acquisition of participant information. Delving deeper, **(Stage 3) Experimental Instructions** dispenses extensive procedural guidelines for the experiment. Subsequently, participants are invited to engage in a sequence of trial games to acquaint themselves with the experimental procedure in **(Stage 4) Game Tutorial**. Proceeding to **(Stage 5) Main Experiment**, it entails various rounds with a diverse array of differentiated AI agents. Post each round, participants are obliged to fill out a survey. The entire experimental process culminates with a comprehensive evaluation of the collaborative AI counterparts in **(Stage 6) Final Survey**. The pipeline is integrated into one platform designed for seamless interaction. Researchers have the flexibility to tailor the pipeline according to their needs, while participants benefit from a user-friendly interface that enables them to complete the stages with ease.

tate an extension of this evaluation to include parameters such as fluency, preference, and comprehension across all collaborating agents. Besides, all six steps are integrated into one platform, named COLE-Platform, designed for seamless interaction. Researchers have the flexibility to tailor the pipeline according to their needs, while participants benefit from a user-friendly interface that enables them to complete the stages with ease.

As shown in Fig. 4, the initial stage entails the formulation of an experiment statement, where the ethical considerations and potential risks associated with the experiment are thoroughly assessed, alongside the information of pertinent experimental details. Subsequently, participants are required to complete a preliminary survey, which seeks to gather fundamental information like their familiarity with the game and their age for more nuanced analyses. The next stage comprises detailed experimental instructions and a game tutorial, equipping participants with the necessary knowledge and skills to play the game and allowing them some initial exposure to the game for familiarization purposes. Then, the process transitions into various game stages, encompassing a game tutorial and the main experiment, which comprises multiple rounds against a range of distinct AI agents. After each round, participants are required to complete a survey. The experiment is concluded with a final survey. Comprehensive data, encompassing player trajectories and AI network parameters,

are meticulously recorded and stored in the respective database. In Appendix E, screenshots from different phases of the pipeline can be found (Figs. 14-19).

The remainder of this section is devoted to introducing the Overcooked game, an overview of the Human-AI Experiment Platform, and a discussion on the design of the experimental scale.

6.1 Overcooked Environment

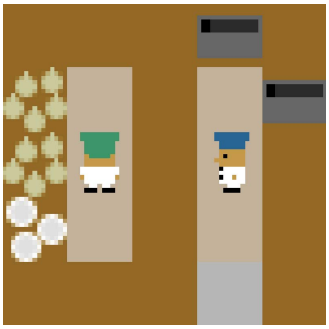
Our paper implements the platform in the Overcooked environment (Carroll et al., 2020; Charakorn et al., 2020; Knott et al., 2021), a simulation environment for reinforcement learning derived from the Overcooked!2 video game (Carroll et al., 2020). The Overcooked environment features a two-player collaborative game structure with shared rewards, where each player assumes the role of a chef in a kitchen, working together to prepare and serve soup for a team reward of 20 points. The environment consists of five distinct layouts: Cramped Room (Cramped Rm.), Asymmetric Advantages (Asymm. Adv.), Coordination Ring (Coord. Ring), Forced Coordination (Forced Coord.), and Counter Circuit (Counter Circ.). Visual representations of these layouts can be found in Fig. 5. The detailed introduction of five layouts is as follows.

Forced Coordination. The Forced Coordination environment is designed to necessitate cooperation between the two players, as they are situated in separate, non-overlapping sections of the kitchen. Furthermore, the available equipment is distributed between these two areas, with ingredients and plates located in the left section and pots and the serving area in the right section. Consequently, the players must work together and coordinate their actions to complete a recipe and earn rewards successfully.

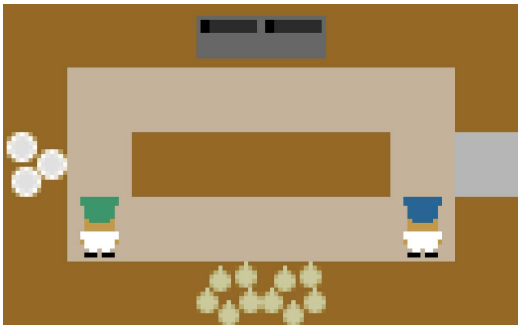
Counter Circuit. The Counter Circuit layout features a ring-shaped kitchen with a central, elongated table and a circular path between the table and the operational area. In this configuration, pots, onions, plates, and serving spots are positioned in four distinct directions within the operational area. Although the layout does not explicitly require cooperation, players may find themselves obstructed by narrow aisles, prompting the need for coordination to maximize rewards. One example of an advanced technique players can learn is to place onions in the middle area for quick and efficient passing, thereby enhancing overall performance.

Asymmetric Advantages. In the Asymmetric Advantages layout, players are divided into two separate areas, but each player can independently complete the cooking process in their respective areas without cooperation. However, the asymmetrical arrangement of the left and right sides encourages collaboration to achieve higher rewards. Specifically, two pots are placed in the central area, accessible to both players. The areas for serving and ingredients, however, are completely distinct. The serving pot is placed near the middle on the left side and far from the middle on the right side, with the ingredients area arranged oppositely. Players can minimize their walking time and improve overall efficiency by learning how to collaborate effectively.

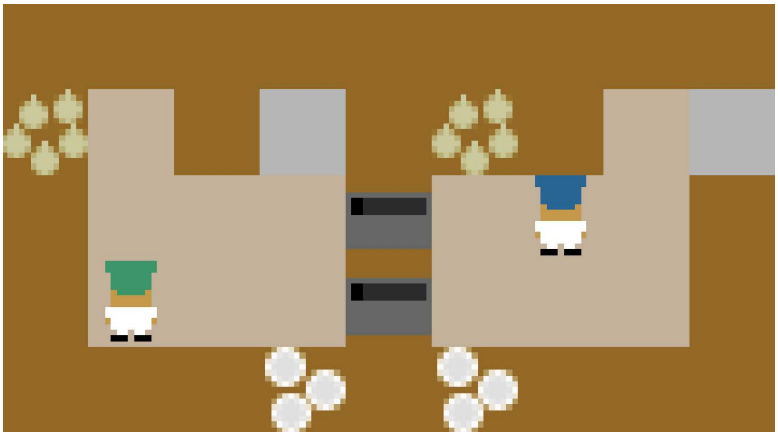
Cramped Room. The Cramped Room layout presents a simplistic environment in which two players share an open room with a single pot at the top and a serving area at the bottom right. In this setup, players could score high even without extensive coordination.



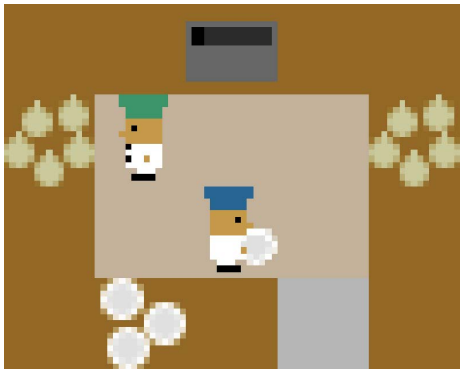
(a) Forced Coord.



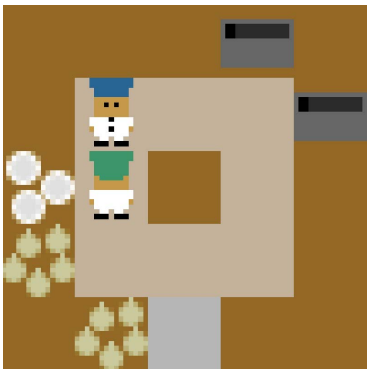
(b) Counter Circ.



(c) Asymm. Adv.



(d) Cramped Rm.



(e) Coord. Ring

Figure 5: Overcooked environment layouts. The Cramped Rm., Asymm. Adv., and Coord. Ring layouts are more conducive to higher rewards when players cooperate with different partners. On the other hand, the Forced Coord., Counter Circ., and Asymm. Adv. layouts offer distinct designs that serve as ideal testbeds to explore and foster cooperation between players.

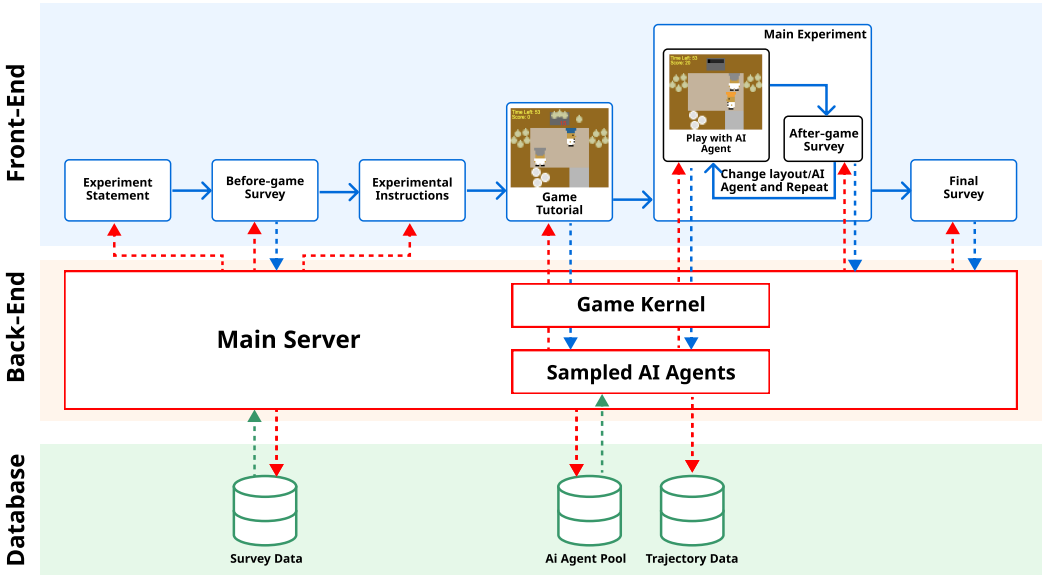


Figure 6: This figure outlines the structure of our human-AI experimental platform, including front-end, back-end, and database components. The front-end, where participants interact, starts with a pre-game process (experiment statement, before-game survey, and experimental instructions) and moves into game stages involving a game tutorial and main experiment with multiple rounds against varying AI agents. A survey follows each round, and a final survey concludes the experiment. All data, including player paths and AI weights, are stored in the database.

Coordination Ring. The Coordination Ring layout is another ring-shaped kitchen, similar to the Counter Circuit. However, this layout is considerably smaller than Counter Circuit, with a close arrangement that makes it easier for players to complete soups. The ingredients, serving area, and plates are all in the bottom left corner, while the two pots are in the top right. As a result, this layout allows more easily achieving high rewards.

In summary, the Cramped Rm., Asymm. Adv., and Coord. Ring layouts are more conducive to achieving higher rewards when players cooperate with different partners. On the other hand, the Forced Coord., Counter Circ., and Asymm. Adv. layouts, due to their unique designs, provide more suitable testbeds for investigating and promoting cooperation among players.

6.2 Details on Human-AI Experiment Platform

We delve into the details of our Human-AI Experiment Platform, offering a comprehensive overview of its components and functionality. To better understand the system's user interface, layout, and functionality, a visual representation of the Human-AI Experiment Platform can be found in Appendix E.

As illustrated in Fig. 6, our human-AI experimental platform is composed of three primary elements: the front-end interface, the back-end server, and the database. The front-end

interface is tasked with both rendering the game and recording real-time human keyboard input. Subsequently, at each frame, it engages in communication with the back-end server, transmitting the current game state and keyboard input as serialized data. On receiving this data, the back-end server initially processes it into an observation for the AI agent. Following this, it alters the game environment in response to the actions taken by both the AI agent and the human player. This newly altered game state is then conveyed back to the front-end interface for the human player's perusal. Simultaneously, the server captures the trajectories of both the human player and the AI agent during each game, storing this information for future analysis. Once the online experiments conclude, these captured trajectories are employed to conclude objective and subjective metrics.

Pre-experiment Pages. Before beginning the experiment, participants are presented with a page containing terms and conditions (e.g., experiment statement) and must decide if they agree to proceed. We provide the experimental statement used in our experiments as reported in Section 7, which includes information on Experimental Purpose, Procedure, Risks and Discomforts, Costs, and Confidentiality. If participants agree to the terms and conditions, they will be asked to complete a form with personal information, such as age, sex, level of skills in the game, etc. These data are used to facilitate a more in-depth analysis of the experiment results and will not be used for any purpose other than the experiment itself.

The instruction page then familiarizes participants with the mechanics of the game and the experiment process. Our example instruction page (Fig. 15) offers details on game settings, world objects, and game controls through both text and images. Following the instruction page, human players need to participate in a trial game to further practice how to play the game. Both the instruction page and trial game contribute to participants gaining a preliminary understanding of the experiment.

Gaming and Questionnaire Pages. The next component is the core of the Human-AI experiment, where participants play with randomly sampled AI teammates and complete questionnaires to evaluate the performance of these AI partners. Specifically, in our initial platform setup (more details are provided in Section 7.6.1), participants play with two different AI teammates on five Overcooked layouts, resulting in 10 games per participant. After each game, they are required to complete a questionnaire that assesses their cooperative gameplay with the AI partner. The final questionnaire primarily focuses on coordination with the AI agent with whom they collaborated and the game they just played. Upon completing all games, participants are directed to the final questionnaire page, where they are asked to provide preliminary feedback on AI agents based on their overall performance, considering factors such as fluency, legibility, and reliability. In contrast to individual game questionnaires, the purpose of this feedback is to analyze and compare the capacities of different agents.

The Human-AI Experiment Platform is designed for seamless customization, enabling users to easily modify various aspects of the system, such as statements, instructions, questionnaires, and game settings. Most customizations can be achieved simply by editing configuration files, streamlined for users to adapt the platform to their specific needs.

Table 1: Evaluation statements for the after-game and final questionnaires consist of three statements each. After-game statements aim to assess the cooperative experience with the AI agent in a single game, while final statements require participants to compare the two AI partners and rate their performance after all games have been completed.

Type	Index	Scale statement
After-game	Q1	The agent and I have good teamwork.
	Q2	The agent is contributing to the success of the team.
	Q3	I understand the agent's intentions.
Final	Q1	Which agent cooperates more fluently?
	Q2	Which agent did you prefer playing with?
	Q3	Which agent did you understand with?

6.3 Experimental Scale

In our effort to thoroughly examine the performance of human-AI coordination in a zero-shot scenario, we advocate for a systemically designed experimental scale. This scale incorporates an expansive range of subjective metrics that appraise individual games involving AI agents while providing a holistic comparison across multiple coordinated players. The metrics utilized extend beyond the superficial, capturing elements such as intention, contribution, and team dynamics during human-AI interactions. Furthermore, the evaluation process encompasses additional aspects such as fluency, participant preferences, and comprehension across all collaborated agents. More specifically, participants are requested to participate in after-game surveys and final-game surveys to assess the performance of AI agent collaboration. Each of these surveys comprises three questions that use a 7-point Likert scale, illustrated in Table 1. The selected survey questions for our experiment draw heavily from a pool of analogous questions found in (Hoffman, 2019). For the after-game surveys, the 7-point Likert scales are arranged with 1 denoting "strongly disagree" and 7 indicating "strongly agree". In the final experiment questionnaire, ratings from 1 to 7 reflect preferences from "strongly favoring the first agent" to "strongly favoring the second agent", for instance, a rating of 4 symbolizes an absence of preference. The rating system adopts a star-assigning format, with the rules clearly stated at the onset of each questionnaire.

7. Experiments

We carry out a series of experiments involving AI agents, human proxies, and human players to assess the performance of COLE_{SV} compared to the baseline methods when collaborating with partners in zero-shot settings. Our experiments focus on the following research questions (RQ):

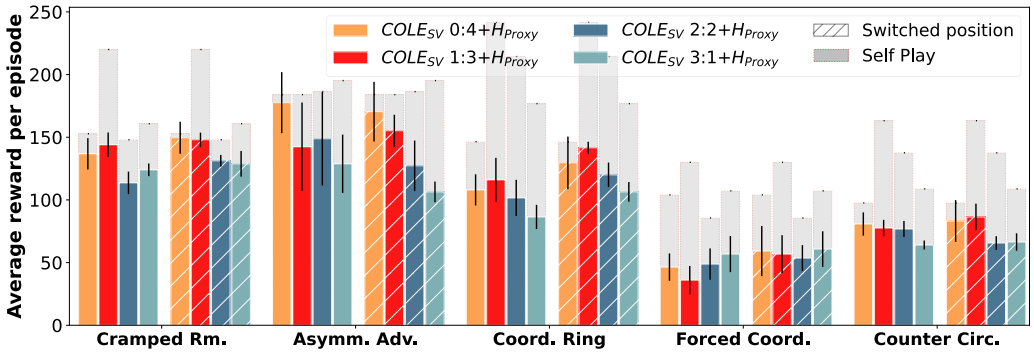


Figure 7: The result of the combining objectives' effectiveness evaluation. Mean episode rewards over 400 timesteps trajectories for COLE_{SV} s with different objective ratios 0:4, 1:3, 2:2, and 3:1, paired with the unseen human proxy partner H_{proxy}. The ratios 0:4, 1:3, 2:2, and 3:1 denote varying proportions between individual and cooperative compatible objectives. The gray bars behind present the rewards of self-play.

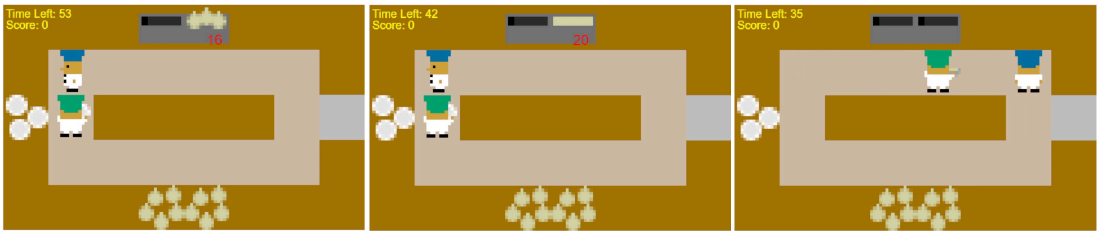
- RQ1: What is the optimal balance between individual and socially compatible objectives? (Section 7.1)
- RQ2: How do the COLE_{SV} and baseline agents perform when cooperating with agents of varying skill levels in a zero-shot setting? (Section 7.2)
- RQ3: Does COLE_{SV} effectively address the issue of cooperative incompatibility? (Section 7.3)
- RQ4: How do each modules within COLE_{SV} contribute to its overall performance? (Section 7.4)
- RQ5: Which solver exhibits superior performance? (Section 7.5)
- RQ6: How do the COLE_{SV} and baseline agents perform when working with humans in a zero-shot setting, and which algorithm do human users prefer? (Section 7.6)

7.1 Evaluation of Combining Objectives' Effectiveness

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.

7.1.1 Experimental Setting

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.



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).

7.1.2 Results

Fig. 7 shows the mean rewards of episodes over 400 time steps of gameplay when paired with the unseen human proxy partner H_{proxy} (Carroll et al., 2020). 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, COLE_{SV} with only the cooperative compatible objective (ratio 0:4) performed better on the Asymmetric Advantages and Forced Coordination layouts when paired with the human proxy partner. Effectiveness evaluations indicate that the combination of 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.

We further visualize the trajectories produced by COLE_{SV} 1:3 and 0:4 with human proxy and expert partners in Overcooked on our demo page. 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 objects 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.

7.2 Evaluation with Human Proxy and AI Agents

To thoroughly assess the ZSC ability, we evaluated the algorithms with unseen human proxy and expert partners. We compare our method with other methods, including self-play (Tesauro, 1994; Carroll et al., 2020), PBT (Jaderberg et al., 2017; Carroll et al., 2020), FCP (Strouse et al., 2021), and MEP (Zhao et al., 2021), all of which use PPO (Schulman

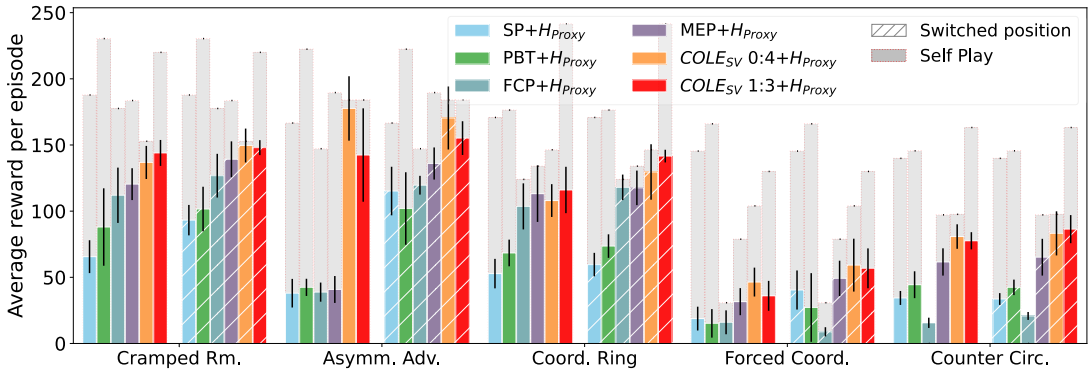


Figure 9: Performance with the human proxy partner. The performance of COLE_{SV} with human proxy partners is presented in terms of mean episode rewards over 400 timesteps trajectories for different objective ratios of 0:4 and 1:3, when paired with the unseen human proxy partner H_{proxy}. The ratios 0:4 and 1:3 denote varying proportions between individual and cooperative compatible objectives. The results include the mean and standard error over five different random seeds. The gray bars indicate the rewards obtained when playing with themselves; the hashed bars indicate the performance when starting positions are switched.

et al., 2017) as the RL algorithm. We use the human proxy model H_{proxy} proposed in (Carroll et al., 2020) as human proxy partners and the models trained with baselines and COLE_{SV} as expert partners.

7.2.1 Experimental Setting

We adopted two sets of evaluation protocols for the evaluation. The first protocol involves playing with a trained human model H_{proxy} trained in behavior cloning. 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 models are limited. 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. The mean of the rewards is recorded as the performance of each method in collaborating with expert teammates. Appendix C and Appendix D give details of the implementation of COLE_{SV} and baselines.

7.2.2 Results with Human Proxy and AI Agents

Fig. 9 presents the performance of SP, PBT, MEP, and COLE_{SV} with 0:4 and 1:3 when cooperating with human proxy 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.

Table 2: Performance with expert partners. Mean episode rewards over 1 min trajectories for baselines and COLE_{SV} with ratio 0:4, 1:3. The ratios 0:4 and 1:3 denote varying proportions between individual and cooperative compatible objectives. 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**

As shown in Fig. 9, COLE_{SV} outperforms other methods in all five layouts when paired with the 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. The objective function of COLE_{SV} 0:4 consists of only cooperative compatible term $\mathbb{E}_{p\sim\phi}\mathbf{w}(s_n, p)$. We believe that focusing solely on cooperative compatible objectives may lead to learning stagnation when collaborating with low-performing partners sampled from distribution ϕ. Consequently, the self-play rewards may be limited. Furthermore, the performance with unseen experts of COLE_{SV} 0:4 as shown in Table 2, is sometimes lower than the baselines.

Table 2 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 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

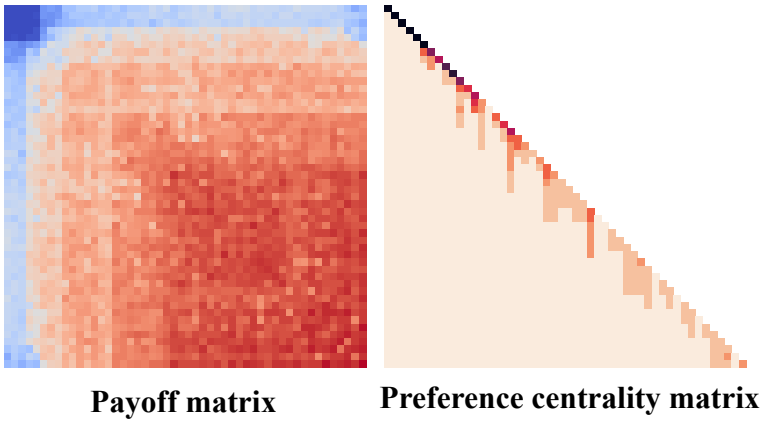


Figure 10: The learning process analysis of COLE_{SV} 1:3. A deeper shade of red in the payoff matrix signifies higher utility, while the darker-colored element on the right represents lower centrality. Clustering of darker-colored areas around the diagonal on the right indicates that the new strategy adopted in each generation is preferred by most strategies, thus overcoming the cooperative incompatibility.

by COLE_{SV} 1:3 while working with expert partners are significantly higher than those of COLE_{SV} 4:0 and the baselines.

Our findings indicate that COLE_{SV} 1:3 exhibits superior adaptive capacity when dealing with partners of expert levels, emphasising individual objectives is key to achieving zero-shot coordination with expert partners. To summarize, COLE_{SV} 1:3 manifests enhanced robustness and versatility in real-world environments, making it apt for collaboration with partners across a spectrum of expertise levels. Hence, in the forthcoming experiments and human-AI interaction studies, we will implement COLE_{SV} 1:3 as our definitive agent, tailored to adapt to human players of diverse skill levels.

7.3 Effectively Conquer Cooperative Incompatibility

To We analyze the learning process of COLE_{SV}, which shows that our method overcomes cooperative incompatibility.

In our analysis of the learning process of COLE_{SV} 1:3 in the Overcooked environment, as shown in Fig. 10, we observe that the method effectively overcomes the problem of cooperative incompatibility. The figure on the left in Fig. 10 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

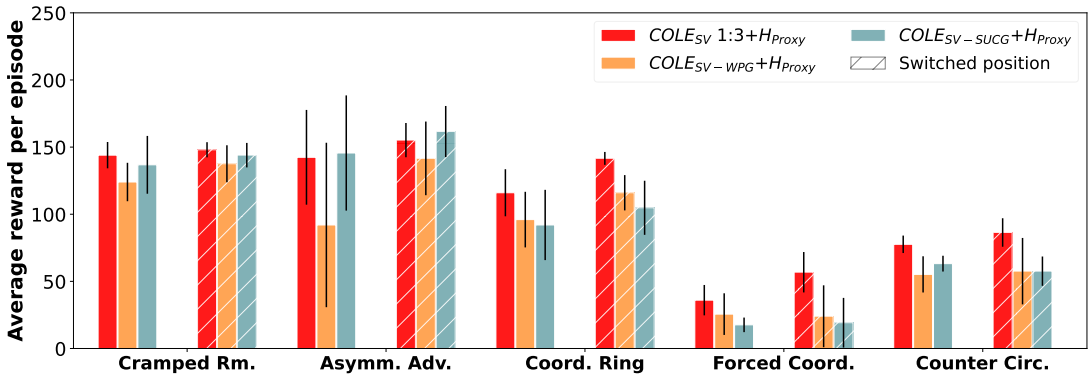


Figure 11: Performance comparison of module effectiveness in COLE_{SV} performance. The performance is presented in terms of mean episode rewards over 400 timesteps trajectories, when paired with the unseen human proxy partner H_{proxy}. $\text{COLE}_{SV\text{-}WPG}$ excludes the WPG element (Eq. 3) while computing the Shapley value. $\text{COLE}_{SV\text{-}UCGG}$ eliminates the SUCG component (Eq. 8) while sampling training partners. The results include the mean and standard error over five different random seeds. The gray bars indicate the rewards obtained when playing with themselves; the hashed bars indicate the performance when starting positions are switched.

of each generation. Thus, we can conclude that our proposed COLE_{SV} effectively overcomes the problem of cooperative incompatibility.

7.4 Ablation Study

In this section, we aim to investigate the efficiency of each component within our proposed algorithm. Refer to Fig. 11 where we compare the performance of our algorithm against three ablated models namely, $\text{COLE}_{SV\text{-}WPG}$, $\text{COLE}_{SV\text{-}UCGG}$, thereby study the individual impact of these specific components. Specifically, $\text{COLE}_{SV\text{-}WPG}$ excludes the WPG element (Eq. 3) while computing the Shapley value. Therefore, in the case of $\text{COLE}_{SV\text{-}WPG}$, the computation of the coalition value employs the average of all utilities within the coalition. That is, $v(C) = \frac{1}{n^2} \sum_{i \in C} \sum_{j \in C} \mathbf{w}(i, j)$, where C represents a coalition within the full coalition set $\mathcal{N}$. Additionally, we examine the effect of the SUCG element (Eq. 8) on the performance of our proposed algorithm during the sampling of training partners. This model is referred to as $\text{COLE}_{SV\text{-}UCGG}$.

Fig. 11 illustrates the performance in terms of average episode rewards over trajectories of 400 timesteps, when coupled with the unseen human proxy partner H_{proxy}. This data presents both the mean and the standard error across five unique random seeds. The gray bars reveal the rewards yielded when the models compete against themselves, while the hashed bars represent the performance when the starting positions are alternated. It is evident from the figure that the WPG element (Eq. 3) significantly contributes to the performance when interacting with the human proxy partner, particularly noticeable in the Asymm. Adv. layout. Furthermore, as depicted in Fig. 11, all baseline models demonstrate subpar

Table 3: Performance comparison of COLE_{SV} and COLE_{R} on five layouts. Mean scores and standard errors (denoted in parentheses) are measured over five different seeds. In the Position column, L and R signify respective initial positions on the left and right of the layout. The term AVG stands for the average score obtained from both sides. COLE_{SV} has demonstrated significantly superior performance compared to COLE_{R} across all layouts, with the exception of the Asymm. Adv. layout. Within this specific Asymm. Adv. setting, COLE_{R} outperformed in the left position and exhibited a comparable score in the right position.

Methods	Position	Layouts				
		CRAMPED RM.	ASYMM. ADV.	COORD. RING	FORCED COORD.	COUNTER CIRC.
COLE_{R}	L	131.20 (14.18)	158.40 (31.66)	80.80 (9.60)	21.60 (10.91)	63.20 (9.93)
	R	135.20 (14.40)	156.80 (19.17)	116.00 (8.39)	27.20 (26.70)	52.00 (10.73)
	AVG	133.20 (14.43)	157.60 (26.18)	98.40 (19.77)	24.40 (20.59)	57.60 (11.76)
COLE_{SV}	L	144.00 (9.80)	142.40 (35.29)	116.00 (17.53)	36.00 (11.31)	77.60 (6.50)
	R	148.00 (5.66)	155.20 (12.75)	141.60 (4.80)	56.80 (15.05)	86.40 (10.61)
	AVG	146.00 (8.25)	148.80 (27.29)	128.80 (18.14)	46.40 (16.89)	82.00 (9.84)

performance when coordinating with the human proxy model in position 0 of the Asymm. Adv. layout. The computation of the Shapley value employing the WPG is fundamental to resolving the asymmetry, as depicted in Fig. 9. We think the reason behind it is that WPG can provide a just evaluation of each position's strategic cooperative ability. In relatively straightforward layouts such as Cramped Rm. and Asymm. Adv., the algorithm displays performance comparable to the ablated models $\text{COLE}_{SV\text{-}WPG}$, $\text{COLE}_{SV\text{-}UCGG}$. However, in the remaining three complex layouts, there is a significant enhancement in the COLE_{SV}'s performance.

7.5 Comparison of COLE_{SV} and COLE_{R}

Table 3 presents a comparative analysis of the coordination payoffs achieved by two practical algorithms, COLE_{SV} and COLE_{R}, across five different layouts. The results, acquired via five distinct seed values with the human proxy model H_{proxy}, feature standard errors, which are indicated within parentheses. The column named 'position' designates the diverse initial positions of the two practical algorithms respectively; here, 'L' stands for left, 'R' for right, while 'AVG' signifies the average derived from the two positions.

As demonstrated in Table 3, COLE_{SV}, utilizing the Shapley value as the core, substantially outperforms COLE_{R} in the Cramped Rm., Coord. Ring, Forced Coord., and Counter Circ. layouts. For the latter, more challenging layouts (Coord. Ring, Forced Coord., and Counter Circ.), the average scores of COLE_{SV} with the H_{proxy} model have shown approximately 30%, 91%, and 43% improvements over COLE_{R}. Interestingly, despite the two algorithms demonstrating similar performance on the right-hand starting position of the Asymm. Adv. layout, COLE_{R} exhibits superior performance when initiated from the left-hand starting position.

7.6 Evaluation with Human Players

We recruited 130 students as participants from universities in different majors, and each provided written informed consent. Each participant received a reward of 50 CNY for their participation. To motivate them to be more engaged and attentive, we established a goal (a number of rewards) for each layout. Participants who successfully reached the goal were awarded an additional bonus of 5 CNY for each layout. Experiments were conducted online where each participant completed their own experiment on a certain web page using a computer, and each experiment costs about 15 to 20 minutes. We do not collect any personally identifiable information, and no significant risk to participants was expected.

7.6.1 Experimental Setting

Participants were first acquainted with the experiment and the rules of Overcooked, with details provided in Section 6. Each participant played a sequence of 5 pairs of games, totaling 10 rounds, with 2 agents (one being COLE, while the other was randomly selected from the four agents in section 7.2.1). The experiment employed the in-group setting for each baseline and COLE_{SV} pair. We also randomized the order of agents within each pair to account for skill differences arising from the inner order variance.

The sequence of five different layouts remained consistent, and we did not compare objective scores across layouts. The five pairs corresponded to five distinct layouts, arranged in the same order for all experiments (1. Cramped Rm., 2. Asymm. Adv., 3. Coord. Ring, 4. Forced Coord., 5. Counter Circ.). Each pair featured one round with the COLE_{SV} agent and one round with the other agent. Under these conditions, we can assume that each participant possesses the same skill level and prior knowledge when encountering the same layouts.

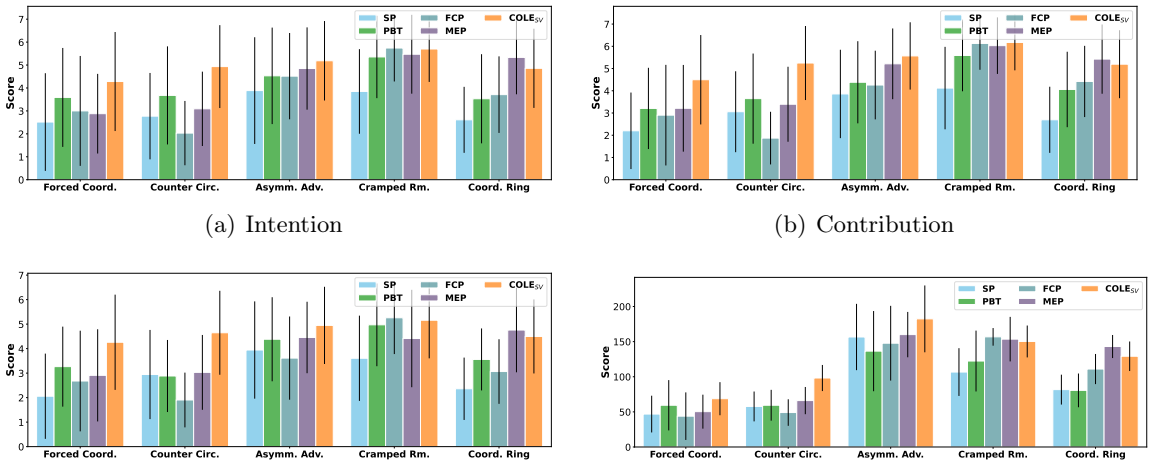
During the experiment, participants were unaware of the specific algorithm names. To ensure fairness, only the color of the chef's hat in the game was used to differentiate between the two algorithms. Each game lasted for 1 minute (approximately 400 steps).

7.6.2 Results of Human Evaluation

In this section, we discuss the findings from our Human-AI experiment. We initially recruited 148 participants from Shanghai Jiaotong University to participate in the study. However, we had to exclude 18 questionnaires from our analysis due to issues such as incomplete responses or negative scoring (e.g., awarding 0 points to all aspects). Thus, our final sample consisted of 130 valid data entries.

The demographic breakdown of the 130 valid participants is as follows: 91 males and 39 females. In terms of prior experience with the Overcooked video game, 70 participants had never played it, 28 had played it but did not consider themselves skilled, 25 regarded themselves as intermediate players, and 7 viewed themselves as experts.

As described in Section 7.2.1, all participants are required to fill in after-game and final questionnaires. The after-game questionnaires are about the subjective evaluation of the playing between participants and one AI agent. The subjective evaluation needs human players to score three questions from 1 to 7, "the agent and I have good teamwork" (short as teamwork), "the agent is contributing to the success of the team" (short as contribution), and "I understand the agent's intentions" (short as intention). As illustrated in Fig. 7.6.2, our proposed COLE_{SV} surpasses all baseline algorithms in three layouts that require cooperation



(a) Intention

(b) Contribution

(c) Teamwork

(d) Rewards

Figure 12: Comparison of subjective evaluation scores and rewards between our COLE$_{SV}$ method and baseline algorithms across five different layouts. "Intention", "Contribution" and "Teamwork" are shorts for "I understand the agent's intentions", "The agent is contributing to the success of the team", and "the agent and I have good teamwork", respectively. Fig. (d) depicts the average rewards comparison. Our COLE$_{SV}$ method obviously outperforms the baselines in Forced Coordination, Counter Circulation, and Asymmetric Advantage. The first three layouts necessitate cooperation for higher rewards. In the simplest layout, Cramped Room, PBT, FCP, MEP, and COLE$_{SV}$ all achieve comparably high scores when playing with humans. In the Coordinated Ring layout, COLE$_{SV}$'s performance is slightly below MEP but surpasses the other algorithms.

(Forced Coord., Counter Circ., and Asymm Adv) based on subjective evaluations of intention, contribution, teamwork, and the objective metric of game rewards. Additionally, human players perceive that all methods, except SP, exhibit similar performance in the Crampped Rm layout. In Crampped Rm., both players are situated within a small rectangular area, enabling them to achieve high scores even with minimal cooperation.

In the Coord. Ring layout, COLE$_{SV}$ marginally underperforms MEP yet surpasses other baseline algorithms. Analyzing human and MEP co-play trajectories revealed MEP's enhanced performance in human subjective experiments arises from its consistent counter-clockwise strategy on the layout. This allows humans to effortlessly adapt to the MEP agent during human-AI experiments by selecting the opposite direction. This predictability leads to misconceptions in questionnaire responses regarding AI's contribution, intention, and teamwork. Participants may perceive a better grasp of MEP's intentions and a more substantial contribution from the agent, but this is primarily due to its easily adaptable and fixed strategy. Nonetheless, when MEP collaborates with diverse partners lacking human-level intelligence, its performance declines, as shown in Fig.9. In contrast, COLE's strategies exhibit greater diversity and are less predictable than MEP's. COLE adapts its routes based

Table 4: The average scores obtained by COLE_{SV} when participants were asked to answer three evaluation questions comparing COLE_{SV} to an assigned baseline. The questions were: "Which agent cooperates more fluently?" (abbreviated as fluency), "Which agent did you prefer playing with?" (abbreviated as preference), and "Which agent did you understand better?" (abbreviated as understanding). A value closer to 1 indicates a stronger preference for COLE_{SV} among participants.

Metrics	**COLE_{SV} v.s. Baselines**			
	SP	**PBT**	**FCP**	**MEP**
Fluency	0.87	0.71	0.78	0.60
Preference	0.88	0.70	0.85	0.68
Understanding	0.80	0.63	0.78	0.61

on varying situations, which can lead to more conflicts. Consequently, in human subjective evaluations, COLE slightly underperforms compared to MEP. Trajectory visualizations of humans with MEP and COLE are available on our demo page.

In addition to evaluating individual games, human players are also asked to compare the two AI models they played with. They are required to rate the AI performances across three dimensions: fluency through "Which agent cooperates more fluently?", preference by "Which agent did you prefer playing with?", and understanding with "Which agent did you understand better?". The mean rating scores are presented in Table 1.

When the value is greater than 0.5, it indicates that human players prefer COLE_{SV} over the baseline. Consequently, a value closer to 1 signifies a more pronounced preference for COLE_{SV} compared to the baseline model. As demonstrated in the table, all values are larger than 0.6, suggesting that human players concur that COLE_{SV} performs better.

When compared to the SP method, COLE_{SV} receives the highest scores (above 0.8), implying that participants believe that COLE_{SV} will outperform SP with a probability of over 80%. In comparison to the state-of-the-art MEP, COLE_{SV}'s scores also exceed 0.6. These results indicate that human players agree that COLE_{SV} performs better in terms of fluency, preference, and understanding compared to the baseline.

In summary, human players concur that COLE_{SV} is more understandable and contributes significantly to their teamwork, particularly in the three layouts that necessitate cooperation: Forced Coord., Counter Circ., and Asymm Adv.

8. Discussion

In this section, we discuss the limitations of our work, with a particular focus on the theoretical aspects, and outline directions for future work.

In this work, we theoretically prove that the algorithm will converge to the local best-preferred strategy, with a convergence rate that is Q-sublinear when using in-degree preference centrality. However, this claim is based on the strict assumption that the RL oracle can train an approximate best response that ranks among the top-k most preferred strategies. The

oracle is proposed as a solver that utilizes a series of techniques such as approximation methods, reinforcement learning algorithms, and optimization techniques to efficiently learn and approximate the best-response policy, which is commonly used in multi-agent reinforcement learning (McMahan et al., 2003; Lanctot et al., 2017b; McAleer et al., 2020; Balduzzi et al., 2019). However, in practice, the approximate best-response policy may deviate significantly from the true best-response policy, which can influence the convergence results. Our implementation includes an additional judge module. If the preference centrality η does not rank among the top-k most preferred strategies, the training process is repeated. However, this approach cannot completely resolve the issue. If k is too large, resulting in a lower ranking, the framework will converge slowly (may be $n \to \infty$) to the best-preferred strategy. Conversely, if k is too small, resulting in a higher ranking, the assumption for each generation will not be guaranteed, which can easily lead to learning failure.

Besides, in our implemented algorithm COLE_{SV}, we introduce the Shapley Value as the tool and develop the Graphic Shapley Value to analyze the cooperative ability. Although we have utilized Monte Carlo permutation sampling to reduce the computational complexity, the computational complexity is still high. Therefore, we only maintain a population of 50 for the limitation of computational resources.

Future Work. Future research will focus on exploring more relaxed assumptions for the theoretical analysis and developing an adaptive mechanism that automatically selects a suitable value for the hyperparameter k to enhance the convergence rate without requiring additional iterations for each update. Additionally, improving the efficiency of the graphical Shapley Value solver and exploring other solvers for evaluating cooperative abilities are crucial for further refining the framework. Beyond this, future work will also involve creating practical algorithms for more complex games beyond Overcooked, extending the applicability and robustness of our approach to a wider range of scenarios and challenges.

9. Conclusion

In this study, we introduce graphic-form games and preference graphic-form games as intuitive reformulations of cooperative games. These reformulations can effectively evaluate and pinpoint cooperative incompatibility during the learning process of Zero-Shot Coordination (ZSC) algorithms, thereby further addressing the issue of cooperative incompatibility in zero-shot human-AI coordination. Additionally, we propose the COLE framework, designed to iteratively approximate the best response preferred by teammates within the latest population. Theoretically, we provide proof that COLE framework converges towards the locally optimal strategy preferred by the rest of the population. If the in-degree centrality is selected as the preference centrality function, the convergence rate would achieve Q-sublinear status.

We also implemented an online pipeline for the Overcooked Human-AI experiment, which allows for easy modifications of questionnaires, model weights, and other elements. To the best of our knowledge, it is the first comprehensive human-AI experimentation pipeline for zero-shot human-AI coordination evaluation including turnkey experimental procedures and scale design. Through the pipeline, we engaged 130 participants in human experiments, and the results highlighted a general preference for our approach over SOTA methods across various subjective metrics. Furthermore, our objective experiments in the Overcooked environment demonstrated that our algorithm, referred to as COLE_{SV}, exceeded

the performance of SOTA algorithms when coordinating with new AI agents and the human proxy model. It also demonstrated the efficient resolution of cooperative incompatibility.

Acknowledgments

Yang Li and Shao Zhang contributed equally in this work. Yang Li is supported by the China Scholarship Council (CSC) Scholarship. The Shanghai Jiao Tong University team is supported by National Natural Science Foundation of China (No.62106141). The authors thank Xihuai Wang for his kind assistance and advice, and both Jia Guo and Tao Shi for their support for our Human-AI experiment platform development. The authors also extend heartfelt thanks to the participants from Shanghai Jiao Tong University.

Appendix A. Proofs of Theorem 4.4

Theorem 4.4. *Let $s_0 \in \mathcal{S}$ be the initial strategy and $s_i = \text{oracle}(s_{i-1}, \mathcal{J}(s_{i-1}, \phi_{i-1}))$ for $i \in \mathbb{N}$. Under the effective functioning of the approximated oracle as characterized by Eq. 5, we can say that the sequence $\{s_i\}$ for $i \in \mathbb{N}$ could converge 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 approach 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 first remark on the RL oracle defined in Eq. 5: $s_{n+1} = \text{oracle}(s_n, \mathcal{J}(s_n, \phi_n))$, *with* $\mathcal{R}(\eta(s_{n+1})) > k$. Under the assumption of the approximated oracle functioning effectively, it follows that the preference centrality η_t associated with generated strategy s_t at generation t resides among the first k positions when arranged in ascending order of centrality values. For simplicity, We define a group g_t at generation t as the set of strategies with the lowest k preference centrality values.

Lemma A.1. *Provided that the approximated oracle is functioning effectively, the maximal preference centrality value, denoted as η_{g_t}, within group g_t is expected to exhibit a non-increasing trend with each successive generation. Furthermore, a consistent non-change trend will last for no more than k generations, followed by a decrease.*

Proof. In light of the approximated oracle's definition, the preference centrality value of the strategy s_t, generated at generation t, will occupy one of the first k positions when sorted in ascending order based on preference centrality values. Consequently, the initial k strategies of group g_t will undergo an update, in which the strategy ranking k-th with the highest preference centrality value is substituted by either the $(k-1)$-th strategy or the newly generation strategy, s_t.

- (Case 1.) If the k-th strategy is replaced by the newly generated strategy s_t, the maximum preference value within the group will experience a reduction.
- (Case 2.) If the k-th strategy is replaced by the $(k-1)$-th strategy, the maximum preference value within the group may remain unchanged or experience a reduction.

– (Case 2.1.) If the preference values of s_{k-1} and s_k are not identical, meaning $\eta(s_{k-1}) < \eta(s_k)$, the replacement will cause a reduction in the maximum preference value.
– (Case 2.2.) If the preference values of s_{k-1} and s_k are identical, meaning $\eta(s_{k-1}) = \eta(s_k)$, the replacement will result in no change in the maximum preference value. The number of generations without changes is limited and may not always occur. In the extreme case where the preference values of all top-k strategies are identical, the k replacements will cause a decrease in the maximum preference value within the group. Thus, we can conclude that in this case, the maximum preference value in the group will decrease after at most k generations.

Until now, we have proven that the group g_t will exhibit a non-increasing trend if the approximated oracle is functioning effectively. Furthermore, a consistent non-change trend will last for no more than k generations before a decrease occurs. □

Let η_{g_t} denote the largest preference centrality in the group g. Thus, we can derive the subsequent equation based on Lemma A.1.

$$\eta_{g_t} = \eta_{g_{t-1}} - \epsilon_{t-1}, \tag{9}$$

where ϵ_{t-1} is a non-negative value and $0 \leq \epsilon \leq \eta_{g_{t-1}}$. By further simplifying the equation, we have

$$\begin{aligned} \eta_{g_t} &= \eta_{g_{t-1}} - \epsilon_{t-1}, \\ &= \eta_{g_{t-1}} - \alpha_{t-1}\eta_{g_{t-1}}, \\ &= \beta_{t-1}\eta_{g_{t-1}}, \end{aligned} \tag{10}$$

where the second line employs $\eta_{g_{t-1}}$ to substitute the residual term, with the adjustment parameters $0 \leq \alpha_{t-1} \leq 1$ 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:

$$\begin{aligned} \eta_{g_t} &= \beta_{t-1}\eta_{g_{t-1}}, \\ &= \beta_{t-1}\beta_{t-2}\eta_{g_{t-2}}, \\ &= \cdots, \\ &= \prod_{i=0}^{t-1} \beta_i \times \eta_{g_0}. \end{aligned} \tag{11}$$

For any $\beta \in \{\beta_0, \cdots, \beta_{t-1}\}$, we have $1 \geq \beta \geq 0$. In addition, we set β_t as a very small positive number if $\eta_t = 0$. Furthermore, we ascertain that the coefficient β is not consistently zero. This is due to the fact that $\beta = 0$ would imply a preference centrality of zero for the strategy, which is not universally attainable within the context of the RL oracle. This very limitation underpins our rationale for introducing the approximated RL oracle. Besides, according to Lemma A.1, we conclude that $\beta = 1$ do not always hold in every generation. Thus, we can conclude that η_t will approach zero within the population as outlined in equation 11.

Through this proof, we have substantiated that under the effective functioning of the RL oracle as characterized by Eq. 5, the sequence s_i is progressing towards the zero of preference centrality within the population. That is, the sequence is converging to a strategy denoted by s^*, which represents a locally best-preferred solution. □

Appendix B. 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

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

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. □

Appendix C. Experimental Details of COLE$_{\mathbf{SV}}$

This paper utilizes Proximal Policy Optimization (PPO) (Schulman et al., 2017) as the oracle algorithm for our set of strategies $\mathcal{N}$, 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 ten removal strategies.

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, COLE$_{\mathrm{SV}}$ takes 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.

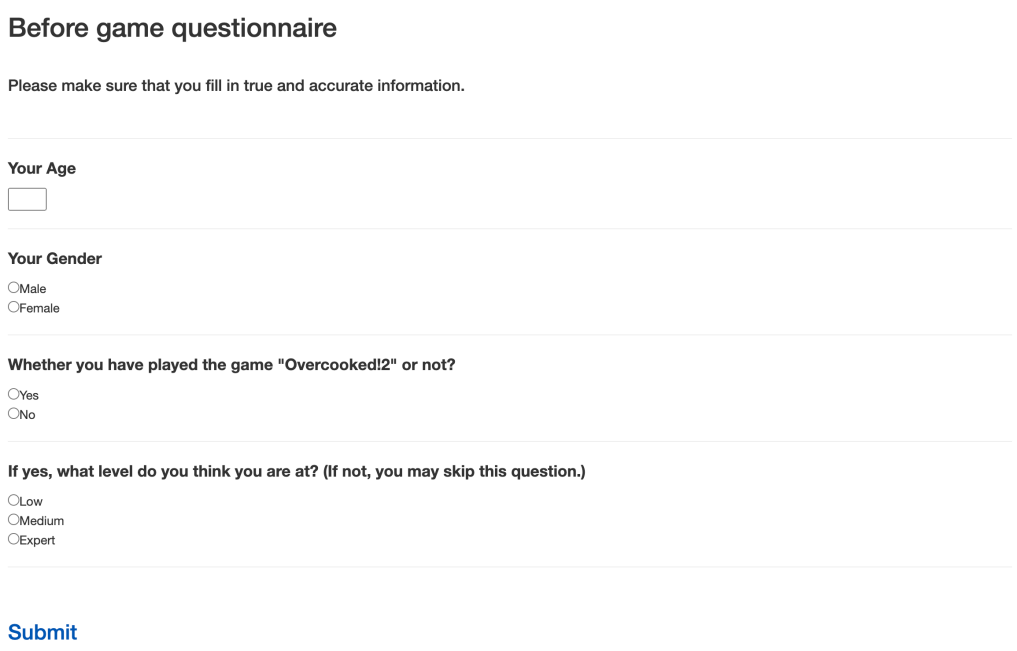


Figure 13: Screenshots of the Human-AI Experiment Platform - participant information questionnaire.

- 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.

Appendix D. Implementations of Baselines

In this part, we will introduce the detailed implementations of baselines. We train and evaluate self-play and PBT based on the Human-Aware Reinforcement Learning repository* (Carroll et al., 2020) and used Proximal Policy Optimization (PPO) (Schulman et al., 2017) as the

*. https://github.com/HumanCompatibleAI/human_aware_rl/tree/neurips2019.

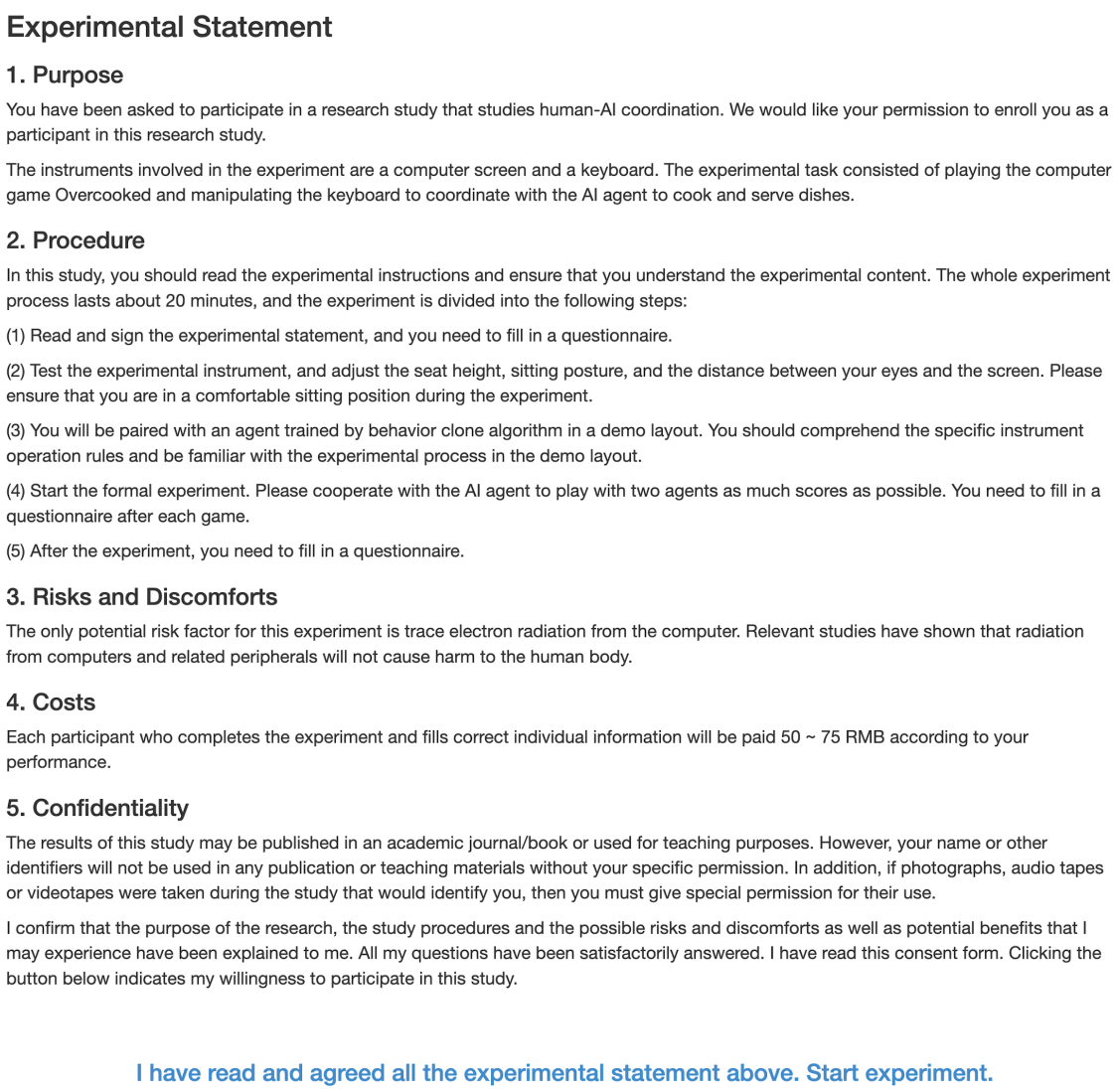


Figure 14: Screenshots of the Human-AI Experiment Platform - experiment statement.

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†.

†. The code of MEP original implementation: `https://github.com/ruizhaogit/maximum_entropy_population_based_training`.

Figure 15: Screenshots of the Human-AI Experiment Platform - instruction providing an overview of the environment interface, operation instructions, and game objectives.

Appendix E. Visual Overview of the Human-AI Experiment Platform

In this section, we provide a visual representation of the Human-AI Experiment Platform to offer readers a better understanding of the system's user interface, layout, and functionality.

Figure 16: Screenshots of the Human-AI Experiment Platform - trial playing: players engage with the human proxy model to familiarize themselves with the system.

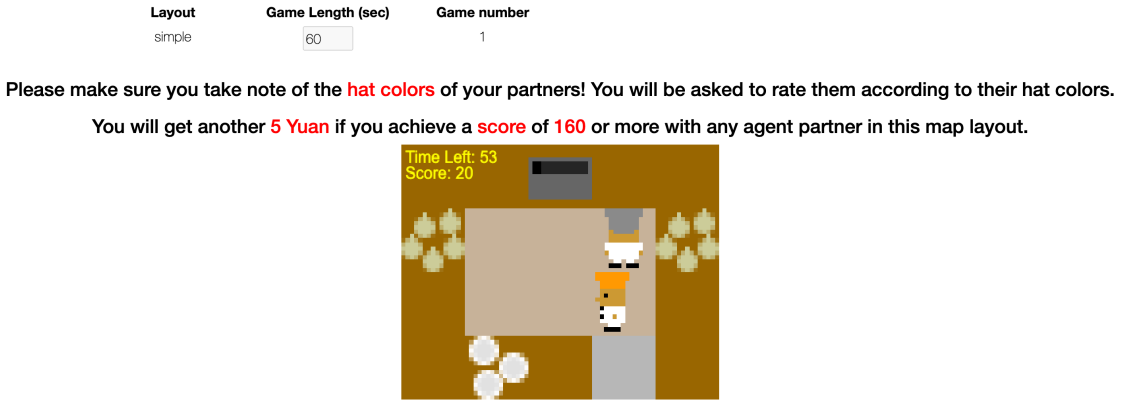


Figure 17: Screenshots of the Human-AI Experiment Platform - Game Playing: displaying the layout name, game length (in seconds), and game number at the top, followed by reminders about hat colors and the bonus awarded for the current game. The middle area is the game interface, where human players interact with the AI agent.

The code of the platform can be find at `https://github.com/liyang619/COLE-Platform`. By including screenshots of various stages of the experiment process in Fig. 14, Fig. 13, Fig. 15, Fig. 16, Fig. 17, Fig. 18, and Fig. 19, we aim to provide a comprehensive overview of the platform, enabling readers to better grasp its design and implementation.

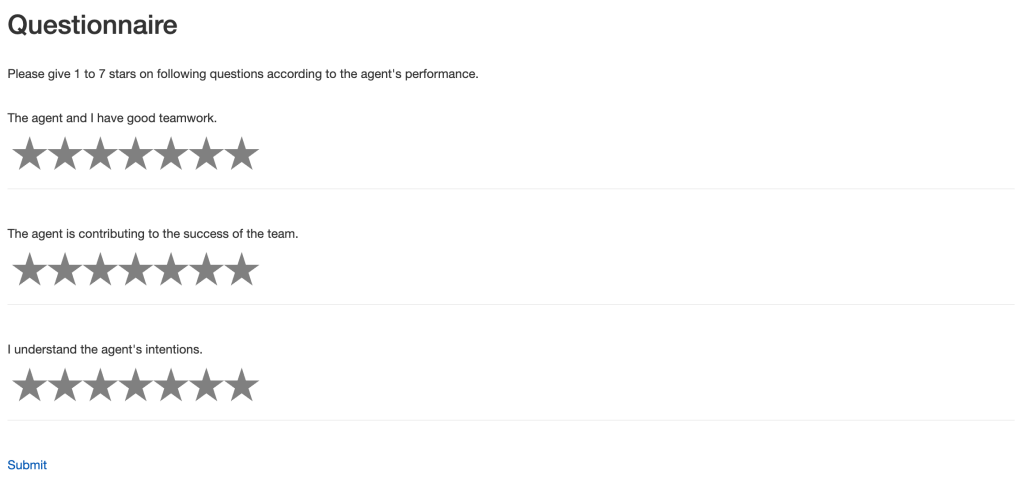


Figure 18: Screenshots of the Human-AI Experiment Platform - Individual Questionnaire: Participants score the performance of the AI teammate in the finished game.

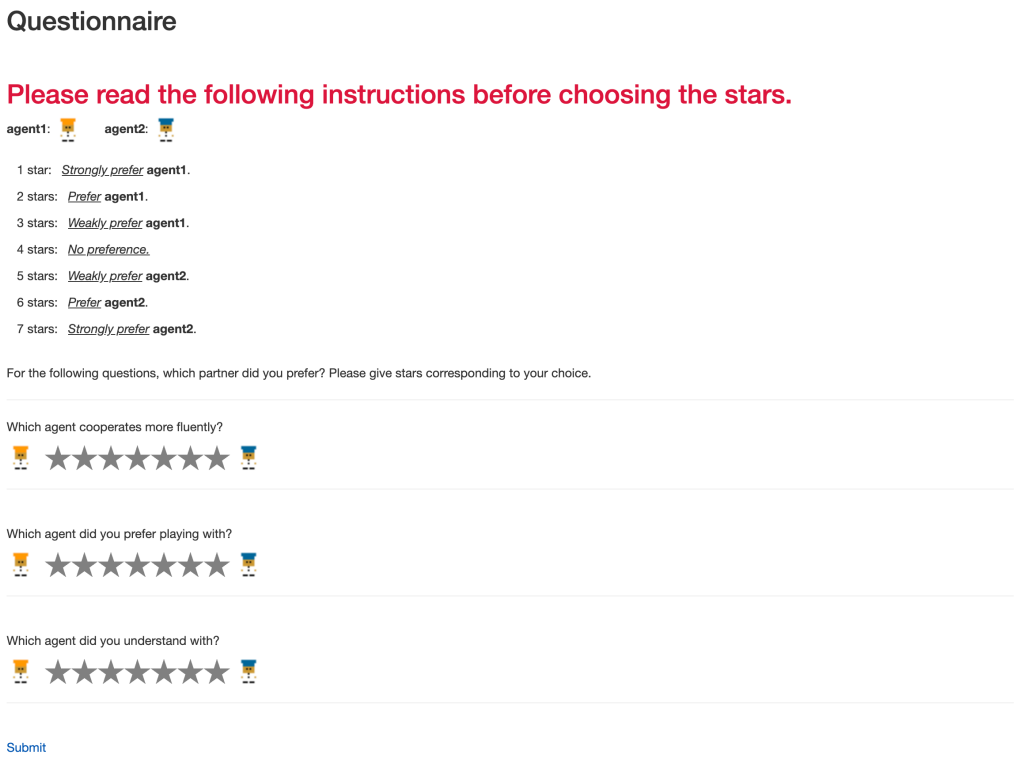


Figure 19: Screenshots of the Human-AI Experiment Platform - Final Questionnaire: Participants compare and rank the performance of AI partners.

References

Aoki, S., Lin, C.-W., & Rajkumar, R. (2021). Human-robot cooperation for autonomous vehicles and human drivers: Challenges and solutions. *IEEE communications magazine*, *59*(8), 35–41.

Balduzzi, D., Garnelo, M., Bachrach, Y., Czarnecki, W., Pérolat, J., Jaderberg, M., & Graepel, T. (2019). Open-ended learning in symmetric zero-sum games. In Chaudhuri, K., & Salakhutdinov, R. (Eds.), *International Conference on Machine Learning (ICML)*, Vol. 97 of *Proceedings of Machine Learning Research*, pp. 434–443. PMLR.

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

Böhmer, W., Kurin, V., & Whiteson, S. (2020). Deep coordination graphs. In *International Conference on Machine Learning (ICML)*, pp. 980–991. PMLR.

Canaan, R., Gao, X., Togelius, J., Nealen, A., & Menzel, S. (2022). Generating and adapting to diverse ad-hoc partners in hanabi..

Carroll, M., Shah, R., Ho, M. K., Griffiths, T. L., Seshia, S. A., Abbeel, P., & Dragan, A. (2020). On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, *32*.

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

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

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

Charakorn, R., Manoonpong, P., & Dilokthanakul, N. (2023). Generating diverse cooperative agents by learning incompatible policies. In *The Eleventh International Conference on Learning Representations*.

Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., & Graepel, T. (2021). Cooperative ai: machines must learn to find common ground. In *Nature Publishing Group*.

de Berardinis, J., Pizzuto, G., Lanza, F., Chella, A., Meira, J., & Cangelosi, A. (2020). At your service: Coffee beans recommendation from a robot assistant.. HAI '20, p. 257–259, New York, NY, USA. Association for Computing Machinery.

De Peuter, S., & Kaski, S. (2022). Zero-shot assistance in novel decision problems..

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

Fudenberg, D., & Levine, D. K. (1998). *The Theory of Learning in Games*. MIT Press Books. The MIT Press.

Gao, Y., Liu, F., Wang, L., Lian, Z., Wang, W., Li, S., Wang, X., Zeng, X., Wang, R., Wang, J., Fu, Q., Yang, W., Huang, L., & Liu, W. (2023). Towards effective and interpretable human-agent collaboration in MOBA games: A communication perspective. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Guestrin, C., Lagoudakis, M., & Parr, R. (2002). Coordinated reinforcement learning. In *International Conference on Machine Learning (ICML)*, Vol. 2, pp. 227–234.

Hoffman, G. (2019). Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems*, *49*(3), 209–218.

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

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., Fernando, C., & Kavukcuoglu, K. (2017). Population based training of neural networks..

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment anything..

Knott, P., Carroll, M., Devlin, S., Ciosek, K., Hofmann, K., Dragan, A., & Shah, R. (2021). Evaluating the robustness of collaborative agents.. AAMAS '21, p. 1560–1562, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

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

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., & Graepel, T. (2017b). A unified game-theoretic approach to multiagent reinforcement learning..

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

Lerer, A., & Peysakhovich, A. (2018). Learning social conventions in Markov games..

Li, S., Zheng, P., Liu, S., Wang, Z., Wang, X. V., Zheng, L., & Wang, L. (2023a). Proactive human–robot collaboration: Mutual-cognitive, predictable, and self-organising perspectives. *Robotics and Computer-Integrated Manufacturing*, *81*, 102510.

Li, Y., Xiong, K., Zhang, Y., Zhu, J., McAleer, S. M., Pan, W., Wang, J., Dai, Z., & Yang, Y. (2023b). Jiangjun: Mastering xiangqi by tackling non-transitivity in two-player zero-sum games..

Li, Y., Zhang, S., Sun, J., Du, Y., Wen, Y., Wang, X., & Pan, W. (2023c). Cooperative open-ended learning framework for zero-shot coordination. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., & Scarlett, J. (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 20470–20484. PMLR.

Liu, X., Jia, H., Wen, Y., Yang, Y., Hu, Y., Chen, Y., Fan, C., & Hu, Z. (2021). Towards unifying behavioral and response diversity for open-ended learning in zero-sum games. In Beygelzimer, A., Dauphin, Y., Liang, P., & Vaughan, J. W. (Eds.), *Advances in Neural Information Processing Systems*.

Lou, X., Guo, J., Zhang, J., Wang, J., Huang, K., & Du, Y. (2023). PECAN: leveraging policy ensemble for context-aware zero-shot human-ai coordination. In Agmon, N., An, B., Ricci, A., & Yeoh, W. (Eds.), *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, pp. 679–688. ACM.

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

Mahajan, A., Samvelyan, M., Gupta, T., Ellis, B., Sun, M., Rocktäschel, T., & Whiteson, S. (2022). Generalization in cooperative multi-agent systems..

McAleer, S., Lanier, J., Wang, K., Baldi, P., Fox, R., & Sandholm, T. (2022). Self-play psro: Toward optimal populations in two-player zero-sum games..

McAleer, S., Lanier, J., Fox, R., & Baldi, P. (2020). Pipeline psro: A scalable approach for finding approximate nash equilibria in large games. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc.

McMahan, H. B., Gordon, G. J., & Blum, A. (2003). Planning in the presence of cost functions controlled by an adversary. In *International Conference on Machine Learning (ICML)*, ICML'03, p. 536–543. AAAI Press.

Meier, R., & Mujika, A. (2022). Open-ended reinforcement learning with neural reward functions. In *ICLR Workshop on Agent Learning in Open-Endedness*.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.. Tech. rep., Stanford InfoLab.

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

Peng, P., Wen, Y., Yang, Y., Yuan, Q., Tang, Z., Long, H., & Wang, J. (2017). Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play starcraft combat games. In *arXiv preprint arXiv:1703.10069*.

Rahman, A., Carlucho, I., Höpner, N., & Albrecht, S. V. (2022). A general learning framework for open ad hoc teamwork using graph-based policy learning..

Rahman, M. A., Hopner, N., Christianos, F., & Albrecht, S. V. (2021). Towards open ad hoc teamwork using graph-based policy learning. In *International Conference on Machine Learning*, pp. 8776–8786. PMLR.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms..

Shapley, L. S. (1971). Cores of convex games. *Int. J. Game Theory*, *1*(1), 11–26.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144.

Srivastava, R., Steunebrink, B., Stollenga, M., & Schmidhuber, J. (2012). Continually adding self-invented problems to the repertoire: First experiments with powerplay..

Strouse, D., McKee, K., Botvinick, M., Hughes, E., & Everett, R. (2021). Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, *34*, 14502–14515.

Team, O.-E. L., Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., McAleese, N., Bradley-Schmieg, N., Wong, N., Porcel, N., Raileanu, R., Hughes-Fitt, S., Dalibard, V., & Czarnecki, W. M. (2021). Open-ended learning leads to generally capable agents. *ArXiv*, *abs/2107.12808*.

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

Tuyls, K., Pérolat, J., Lanctot, M., Leibo, J. Z., & Graepel, T. (2018). A generalised method for empirical game theoretic analysis. *CoRR*, *abs/1803.06376*.

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

Wang, X., Tian, Z., Wan, Z., Wen, Y., Wang, J., & Zhang, W. (2023). Order matters: Agent-by-agent policy optimization. In *The Eleventh International Conference on Learning Representations*.

Wang, X., Zhang, S., Zhang, W., Dong, W., Chen, J., Wen, Y., & Zhang, W. (2024). Zsc-eval: An evaluation toolkit and benchmark for multi-agent zero-shot coordination. In *arXiv preprint arXiv:2310.05208*.

Wei, H., Chen, J., Ji, X., Qin, H., Deng, M., Li, S., Wang, L., Zhang, W., Yu, Y., Lin, L., Huang, L., Ye, D., Fu, Q., & Yang, W. (2022). Honor of kings arena: an environment for generalization in competitive reinforcement learning. In *NeurIPS*.

Wen, M., Wan, Z., Zhang, W., Wang, J., & Wen, Y. (2024). Reinforcing language agents via policy optimization with action decomposition. In *arXiv preprint arXiv:2405.15821*.

Xing, W., & Ghorbani, A. (2004). Weighted pagerank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research, 2004.*, pp. 305–314.

Xue, K., Wang, Y., Yuan, L., Guan, C., Qian, C., & Yu, Y. (2022). Heterogeneous multi-agent zero-shot coordination by coevolution..

Yang, Y., Luo, J., Wen, Y., Slumbers, O., Graves, D., Bou Ammar, H., Wang, J., & Taylor, M. E. (2021). Diverse auto-curriculum is critical for successful real-world multiagent learning systems..

Ye, D., Chen, G., Zhang, W., Chen, S., Yuan, B., Liu, B., Chen, J., Liu, Z., Qiu, F., Yu, H., Yin, Y., Shi, B., Wang, L., Shi, T., Fu, Q., Yang, W., Huang, L., & Liu, W. (2020).

Towards playing full MOBA games with deep reinforcement learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., & Lin, H. (Eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yu, C., Gao, J., Liu, W., Xu, B., Tang, H., Yang, J., Wang, Y., & Wu, Y. (2023). Learning zero-shot cooperation with humans, assuming humans are biased. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhang, S., Xu, H., Jia, Y., Wen, Y., Wang, D., Fu, L., Wang, X., & Zhou, C. (2023). Geodeepshovel: A platform for building scientific database from geoscience literature with ai assistance. *Geoscience Data Journal*, *10*(4), 519–537.

Zhang, S., Yu, J., Xu, X., Yin, C., Lu, Y., Yao, B., Tory, M., Padilla, L. M., Caterino, J., Zhang, P., & Wang, D. (2024). Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Zhao, R., Song, J., Haifeng, H., Gao, Y., Wu, Y., Sun, Z., & Wei, Y. (2021). Maximum entropy population based training for zero-shot Human-AI coordination..