White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/223377/

Version: Published Version

Proceedings Paper:
Loftin, R., Bandyopadhyay, S. and Çelikok, M.M. (2025) On the complexity of learning to cooperate in populations of socially rational agents. In: Vorobeychik, Y., Das, S. and Nowé, A., (eds.) Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems. 24th International Conference on Autonomous Agents and Multiagent Systems, 19-23 May 2025, Detroit, Michigan, USA. International Foundation for Autonomous Agents and Multiagent Systems, pp. 233-241. ISBN: 9798400714269. ISSN: 1558-2914. EISSN: 1558-2914.

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

On the Complexity of Learning to Cooperate in Populations of Socially Rational Agents

Saptarashmi Bandyopadhyay*
Department of Computer Science
University of Maryland
College Park, USA
saptab1@umd.edu

Mustafa Mert Çelikok*
Department of Intelligent Systems
Delft University of Technology
Delft, The Netherland
m.m.celikok@tudelft.nl

Robert Loftin*
Department of Computer Science
University of Sheffield
Sheffield, UK
r.loftin@sheffield.ac.uk

ABSTRACT

Artificially intelligent agents deployed in the real world must be able to reliably *cooperate* with humans (as well as other, heterogeneous AI agents). To provide formal guarantees of successful cooperation, we must make some assumptions about how these partner agents could plausibly behave. Realistic assumptions must account for the fact that other agents may be just as adaptable as our agent is. In this work, we consider the setting where an AI agent must cooperate with members of some *target population* of agents in a finitely repeated two-player general-sum game, where individual utilities are private. Two natural assumptions in this setting are 1) all agents in the target population are individually rational learners, and 2) when paired with another member of the population, with high-probability the agents will achieve the same expected utility as they would under some Pareto-efficient equilibrium strategy of the underlying stage game. Our theoretical results show that these assumptions alone are insufficient to select an AI strategy that achieves *zero-shot* cooperation with members of the target population. We therefore consider the problem of learning such a cooperation strategy using observations of members of the target population interacting with one another, and provide upper bounds on the sample complexity of learning such a cooperation strategy. Our main result shows that, under the above assumptions, these bounds can be much stronger than those arising from a "naive" reduction of the problem to one of imitation learning.

KEYWORDS

Social Cooperation, Rational Agents, Upper Bound, Stackelberg Equilibria, Imitation Learning, Reinforcement Learning, Offline, Replicator Dynamics

ACM Reference Format:
Saptarashmi Bandyopadhyay*, Mustafa Mert Çelikok*, and Robert Loftin*. 2025. On the Complexity of Learning to Cooperate in Populations of Socially Rational Agents. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Imagine a hospital deploying an AI assistant to help their surgeons plan and execute complex surgeries. For instance, the AI assistant can take the role of a second surgeon in complex procedures that benefit from a two-surgeon approach [15]. When first deployed, the AI agent is unlikely to have comprehensive knowledge of the population consisting of its potential partners (i.e. the surgeons working in the hospital). Differences between human surgeons, such as preferences, capabilities, and internal states—including surgical experience, familiarity with specific procedures, or even mental focus under pressure—can critically impact cooperation. A successful AI agent should be able to adapt its strategy to each human surgeon it partners with. The central question of this paper is how to efficiently learn such adaptive and cooperative *meta-strategies* from a dataset of cooperative interactions between the members of the target population.

To illustrate, consider experience level. An experienced surgeon may prefer a fast strategy to reduce surgery duration, improving post-op recovery time. Conversely, an inexperienced surgeon may prefer a slower, cautious strategy. Here a *strategy* refers to the policy an agent follows in a single collaborative surgery. Our goal is to learn an adaptive *meta-strategy* that maps from the history of interactions (e.g. history of collaborative surgeries performed so far) to strategies, which allows for the AI agent to adapt to the needs of its current human partner over time.

It is possible to learn a good AI *strategy* for individual partners using past surgical data through imitation learning (e.g. [16]). However, learning a good *meta-strategy* through imitation becomes impractical as task complexity, partner diversity, and task duration increase. Imitation learning here would mean learning a function mapping from histories of multiple surgeries to new surgical strategies. To do so, the AI agent would require datasets that capture long-term interactions between human surgeons and cover the full range of surgeon and patient profiles. Additionally, in high-stakes environments like surgery, imperfect imitation may lead to unacceptable failure modes, resulting in the AI agent's role being terminated.

The problem setup. To formalize the above intuitions, we model the interaction between the AI agent and the individual members of the population as a repeated, two-player, general-sum matrix game with private types. Each agent's type is their private information, where different types of agents have distinct payoff functions. Types embed behavioural differences amongst the agents through payoffs, inducing general-sum games between partners with different types,

*Author order is alphabetical.

even though they are collaborating on a task that requires teamwork (e.g., no-conflict games [1] or ad-hoc teamwork [27]). Each stage game represents a complete interaction between two agents. For instance, in the case of our surgery example, a single stage game of the repeated game corresponds to a complete surgery.

Contributions. We present a novel definition called a *socially intelligent* population, where the member agents are (1) Consistent, meaning an individual will perform at least as good as its best pure strategy in hindsight and (2) Compatible in pairs, meaning they achieve performance comparable to a Pareto-optimal Nash equilibrium (PONE). The former, also called the no-regret property, is often seen as a minimal requirement for rationality, whereas the latter has been used to describe successful cooperation by Powers and Shoham [21]. Our main contribution is an upper bound on the sample complexity for learning cooperation meta-strategies in socially intelligent populations. While consistent partners (as defined in section 3) do not guarantee success for imitation learning, we demonstrate that compatibility within the population makes imitation feasible. However, the lower bound on sample complexity grows exponentially due to the need to account for histories. We derive an upper bound in Theorem 5.3 showing that for socially intelligent partners—who are both consistent and compatible—it is possible to learn cooperation meta-strategies more efficiently than through imitation alone. A consequence of the lower bound in Theorem 4.5 is that, even when the target population can be assumed to be socially intelligent, without additional information about this population we cannot find a strategy that can reliably cooperate with members of this population. We therefore consider a more realistic interpretation of the zero-shot coordination problem, where the AI agent must cooperate with an entirely new partner (whose private type is unknown), but has observations of other members of the population, so it can learn the strategies (perhaps better thought of as "meta-strategies") that these agents use to coordinate with new partners.

Outline. In section 2, we discuss the intuition of our approach and the motivations behind it. Then in section 3, we define our multi-agent setting and provide background on consistency (in the form of Hannan-consistency) and external regret. We introduce a novel definition of compatibility in definition 3.2, inspired by Powers and Shoham [21]. Section 3.3 introduces our definition of social intelligence and presents a realistic class of agents that meet this criterion. In section 4, we frame the learning problem as minimizing *altruistic regret* and derive lower bounds on its sample complexity. Finally, in section 5, we present our upper-bound result, showing that a strategy we call *imitate-then-commit* can leverage the social intelligence of the population to learn cooperative strategies more efficiently.

2 MOTIVATION

Socially intelligent populations. Our focus in this paper is on populations where members have established conventions that enable effective cooperation. For instance, two surgeons can plan and execute a complex surgery together efficiently, even if they have not worked together before, because they share a common set of conventions learned through similar education. This can be seen as the members of the population being *compatible* with each other. In addition, each member should individually satisfy a base level of rationality. Our definition of social intelligence formalizes these intuitions.

	Fast	**Balanced**	**Cautious**
Fast	(4, 2)	(5, 4)	(3, 3)
Balanced	(5, 4)	(6, 6)	(4, 5)
Cautious	(3, 3)	(4, 5)	(7, 7)

Table 1: Payoff matrix for a repeated two-player game with an experienced surgeon (row) and inexperienced surgeon (column). The best cooperative outcome is achieved when experienced surgeon slows down to match the inexperienced.

General-sum games. In our setting, agents with different types will have distinct payoff functions due to different behavioural propensities. Consider the example given in Table 1 for a pair of experienced and inexperienced surgeons. Even though the surgery is a cooperative task, the agents have non-identical payoffs due to differences between their types (i.e. experience level). Here, the general-sum aspect models the potential failure of coordination between the agents due to their private types. If the row player is experienced and mistakenly thinks its partner is also experienced, it will choose the fast approach, leading to the sub-optimal cooperation outcomes. However, if for instance the agents learned each other's types through repeated interactions, they can both choose the cautious approach towards the optimal cooperation outcome (Cautious, Cautious).

Our approach proposes that the AI agent initially mimics the behaviour of a team member over a short horizon, gathering enough information to infer its partners' types, while behaving as expected from a member of the team. For instance, the AI agent can start by imitating the average behavior of a human surgeon from the dataset, gradually inferring the human partner's type. This preliminary imitation might not be immediately efficient for the specific partner, but as long as it remains human-like, it is more likely to be tolerated. Once the partner's type is inferred, the AI agent can transition to a type-conditioned strategy that is well-aligned with its partner. This approach would ensure that the partner is more likely to engage with the AI agent as a trusted collaborator, avoiding early-stage friction that might otherwise lead to the termination of the AI agent's involvement. Our formalization of the repeated two-player general sum matrix games setting is motivated from the notion of replicator dynamics [3, 25] in evolutionary game theory. The replicator equation represents the proportion of each type in a population as the difference of the fitness of a population for that type to the average fitness across all types. The replicator dynamics construct helps to understand the type of the two agents sampled for our repeated two-player general sum matrix games.

Our theoretical results apply to various real-world scenarios where the goal of an AI agent is to learn how to cooperate with self-interested agents with private types such as humans. Most importantly, our AI agent itself is not necessarily self-interested, since its goal is to assist or cooperate with partners coming from a population. However, the partners it is trying to cooperate with

are self-interested. Our framework offers efficient bounds for learning viable cooperation meta-strategies based solely on observed interactions between the members of the population. We provide further motivating examples of populations and use cases in the supplementary materials [1] like customer support chatbots, legal assistants, software development assistants, AI health coach and human-robot cooperation in a factory.

3 PRELIMINARIES

Repeated two-player matrix games with private types. First, we define a *class* of repeated two-player matrix games with private types with the tuple $\mathcal{G} = (\mathcal{I}, \mathcal{A}, \Theta, G, T)$ where $\mathcal{I} = \{1, 2\}$ is the set of agents, $\mathcal{A}$ is the set of N pure strategies available to both agents (called *actions* henceforth), Θ is a space of types, G is a function that maps an agent type $\theta \in \Theta$ to a payoff matrix $G(\theta) \in \mathbb{R}^{N\times N}$, and $0 < T < \infty$ is a fixed number of stages. Let $\boldsymbol{\theta} = (\theta_1, \theta_2)$ denote a joint type for both agents. Then, a specific instance of a game from this class is given by $\mathcal{G}(\boldsymbol{\theta}) = (\mathcal{I}, \mathcal{A}, G(\boldsymbol{\theta}), \boldsymbol{\theta}, T)$ such that $G(\boldsymbol{\theta}) = [G(\theta_1), G(\theta_2)^\top]$ is its payoff matrix.

Throughout the paper, we will assume that a joint type $\boldsymbol{\theta}$ directly induces the game $\mathcal{G}(\boldsymbol{\theta})$, and the class $\mathcal{G}$ is fixed. Then in a single *episode*, the agents play $\mathcal{G}(\boldsymbol{\theta})$ for T stages. We let a_t^1 and a_t^2 denote the actions chosen by agents 1 and 2 in stage $0 < t \leq T$. For mixed strategies $\sigma, \sigma' \in \Delta(\mathcal{A})$, we let $G(\sigma, \sigma'; \theta_i) = \sigma^\top G(\theta_i)\sigma'$. We overload a_t^1 and a_t^2 to also denote the mixed strategies that assign all probability mass to actions a_t^1 and a_t^2, such that $G(a_t^1, a_t^2; \theta_1)$ and $G(a_t^1, a_t^2; \theta_2)$ are agent 1 and 2's realized payoffs at stage t. We also assume that without the loss of generality, for all $\theta \in \Theta$, $G(a_t^1 = a, a_t^2 = a', \theta) \in [0, 1], \forall a, a' \in \mathcal{A}$. In other words, payoffs are always bounded in $[0, 1]$.

Let $\mathcal{H}_t = (\mathcal{A} \times \mathcal{A})^t$ be the set of histories of length t (with $\mathcal{H}_0 = \{\emptyset\}$), and let $\mathcal{H}_{\leq t} = \bigcup_{s=0}^{t} \mathcal{H}_s$ be the set of all histories of length at most t. The meta-strategy space Π for an agent is then the space of mappings $\pi : \Theta \times \mathcal{H}_{\leq T-1} \mapsto \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the set of probability distributions over the action set. As a functional of types, a meta-strategy $\pi(\theta, \cdot)$ maps a type θ to a *behavioral strategy* [26, Chapter 5.2.2] that maps histories of play to action distributions, such that $a_t^i \sim \pi_i(\theta_i, h_{t-1})$. We denote agent i's expected total payoff for following meta-strategy π against π' as

$$M_i(\pi, \pi'; \theta, \theta') = \mathrm{E}\left[\sum_{t=1}^{T} G(a_t^i, a_t^{-i}; \theta_i) \,\middle|\, \pi_i = \pi, \pi_{-i} = \pi', \theta_i = \theta, \theta_{-i} = \theta' \right] \quad (1)$$

where the expectation is with respect to the strategies.

3.1 Consistency

A natural criterion for rationality is that an agent should attempt to achieve a payoff nearly as large as the best response to its partner's average strategy, which we refer to as *consistency*. To account for the non-stationary behavior of other agents', we specifically consider

[1]The supplementary materials for our paper can be found in its arXiv version at https://arxiv.org/abs/2407.00419

	A	B
A	2, 2	0, 0
B	0, 0	1, 1

(a) A fully-cooperative 2x2 matrix game.

	C	D
C	2, 2	0, 3
D	3, 0	1, 1

(b) The prisoner's dilemma game.

Table 2

Hannan consistency [13], which in our finite-time setting simply requires that an agent have bounded *external regret* over T stages. The external regret for agent i is defined as

$$R_i^{\mathrm{ext}}(h; \theta_i) = \max_{a^i \in [N]} \sum_{t=1}^{|h|} \left\{ G(a^i, a_t^{-i}(h); \theta_i) - G(a_t^i(h), a_t^{-i}(h); \theta_i) \right\} \quad (2)$$

where $a_t^i(h)$ denotes the action i played at stage t within the history $h \in \mathcal{H}_{\leq T}$.

Definition 3.1 (Consistency)**.** For $\delta, \epsilon, T > 0$, an agent $i \in \{1, 2\}$ is (δ, ϵ, T)-*consistent* if, for all types $\theta \in \Theta$, and *any* partner strategy, we have that $\frac{1}{T} R_i^{\mathrm{ext}}(h_T; \theta) \leq \epsilon$ with probability at least $1 - \delta$.

In essence, consistency requires an agent i to achieve bounded external regret regardless of its type or partner. We also define the *expected* external regret $\bar{R}_i^{\mathrm{ext}}(h; \theta)$ by replacing the $a_t^i(h)$ (the action i played at stage t) with their full meta-strategy $\pi^i(\theta, h_t)$. $R_i^{\mathrm{ext}}(h; \theta)$ and $\bar{R}_i^{\mathrm{ext}}(h; \theta)$ are related by the inequality

$$R_i^{\mathrm{ext}}(h_t; \theta) \leq \bar{R}_i^{\mathrm{ext}}(h_t; \theta) + \sqrt{\frac{T}{2} \ln \frac{1}{\delta}}, \quad (3)$$

which holds with probability (w.p.) at least $1 - \delta$ for all $t \leq T$ simultaneously (this follows directly from [6, Lemma 4.1]). We therefore only need to bound $\bar{R}_i^{\mathrm{ext}}(h_t; \theta)$ to provide high-probability regret bounds.

3.2 Compatibility

Even in a fully cooperative game, the fact that both agents are consistent does not guarantee that they will achieve an optimal outcome. In the 2×2 game in Table 2a for example, both (A, A) and (B, B) are Nash equilibria to which consistent agents could converge, but only (A, A) is optimal. In general-sum games, consistency may preclude Pareto-optimal outcomes, as in the classic prisoner's dilemma game (Table 2b), where the only outcome in which neither player incurs positive regret is (D, D), which is Pareto-dominated by (C, C). Therefore, similar to Powers and Shoham [21], we define successful cooperation in terms of the *Pareto-optimal Nash equilibria* (PONE) [18] of a game $\mathcal{G}(\boldsymbol{\theta})$.

Let $\mathcal{N}(\boldsymbol{\theta}) \subseteq \Delta(\mathcal{A}) \times \Delta(\mathcal{A})$ be the set of Nash equilibria (NE) of $\mathcal{G}(\boldsymbol{\theta})$. For a fully-cooperative game, $\mathcal{N}(\boldsymbol{\theta})$ will contain all globally optimal strategy profiles. It may, however, also contain joint strategies that are highly sub-optimal. Let $\mathcal{P}(\boldsymbol{\theta}) \subseteq \mathcal{N}(\boldsymbol{\theta})$ denote the set of Pareto optimal Nash equilibria. In this work, we say that a strategy profile $\langle \sigma_1, \sigma_2 \rangle \in \mathcal{P}(\boldsymbol{\theta})$ if and only if $\langle \sigma_1, \sigma_2 \rangle \in \mathcal{N}(\boldsymbol{\theta})$,

and there does not exist $\langle \sigma_1', \sigma_2' \rangle \in \mathcal{N}(\boldsymbol{\theta})$ such that $G(\sigma_1', \sigma_2'; \theta_1) \geq G(\sigma_1, \sigma_2; \theta_1)$ and $G(\sigma_2', \sigma_1'; \theta_2) \geq G(\sigma_2, \sigma_1; \theta_2)$, *and* $G(\sigma_i', \sigma_{-i}'; \theta_i) \geq G(\sigma_i, \sigma_{-i}; \theta_i)$ for some $i \in \{1, 2\}$. This means that $\langle \sigma_1, \sigma_2 \rangle$ is a PONE if it is a Nash equilibrium of $\mathcal{G}(\boldsymbol{\theta})$, and it is not Pareto-dominated by any other Nash equilibrium of $\mathcal{G}(\boldsymbol{\theta})$. Intuitively, if two agents are individually consistent, and willing to cooperate with each other, their joint payoff profile should come close to a PONE. We formalize this intuition as follows:

Definition 3.2 (Compatibility)**.** For $\delta, \epsilon, T > 0$, two agents π^1 and π^2 are (δ, ϵ, T)-*compatible* if, when played together, for any joint type $\boldsymbol{\theta}$, w.p. at least $1 - \delta$, $\exists \langle \sigma_1^*, \sigma_2^* \rangle \in \mathcal{P}(\boldsymbol{\theta})$ s.t.

$$\frac{1}{T} \sum_{t=1}^{T} G(\sigma_i^*, \sigma_{-i}^*; \theta_i) - G(a_t^i, a_t^{-i}; \theta_i) \leq \epsilon, \tag{4}$$

for both $i = 1$ and $i = 2$.

A pair of agents is compatible if, when paired together, with high-probability over their path of play h_T there will exist some PONE that does not ϵ-dominate their realized payoffs. Note that this definition is the approximate and finite-horizon version of the one provided in [21].

For the populations we consider, compatibility is a reasonable assumption. In a way, we focus on populations that have evolved over a long time learning to cooperate with each other. In the case of our illustrative surgery example (section 1), the population has evolved dynamically over the course of human medical history, learning and adapting its conventions to enable compatibility. Behaviours that are not compatible cannot survive in this population, considering medical professionals must confer to certain rules, guidelines, and behavioural norms amongst each other.

3.3 Socially Intelligent Agents

We argue that it is natural to model an existing population of cooperating agents as a set of approximately compatible, but otherwise heterogeneous agents. We therefore introduce the more general idea of a socially intelligent *class* of agents that are compatible with any other member of their class:

Definition 3.3 (Social Intelligence)**.** A set C of agents forms a *socially intelligent class* w.r.t. Θ if, for some $\delta, \epsilon, T > 0$, each agent $\pi \in C$ is (δ, ϵ, T)-consistent for all $\theta \in \Theta$, and any two agents $\pi, \pi' \in C$ are (δ, ϵ, T)-compatible over all joint types Θ. An individual agent π is called *socially intelligent* if it forms a socially intelligent class $\{\pi\}$ with itself.

The consistency requirement ensures that any agent in the population always has bounded average regret, whereas the approximate compatibility means if both agents are from C, with high probability there will exist some PONE that does not ϵ-dominate their path of play. Below we describe a socially intelligent class based on a pre-agreed *handshake protocol*. These protocols can be thought of as handshakes that allow the members of a socially intelligent population identify each other's types efficiently.

Handshake protocols. For a type space Θ, we first define a function s that maps from each joint type $\boldsymbol{\theta}$ to a strategy profile in $\mathcal{P}(\boldsymbol{\theta})$ such that $s(\boldsymbol{\theta}) \in \mathcal{P}(\boldsymbol{\theta})$. We can think of this function as a common "convention" the agents in C have settled upon. Since we assume private types, members of C do not know each other's type at the beginning of their interaction. If any type $\theta \in \Theta$ can be communicated to others in a sequence of $k < T$ actions, then agents in C can agree on a handshake protocol. Let the protocol be a map κ from types to a history-dependent policy. Then, at the beginning of each episode, both agents will play their corresponding $\kappa(\theta_i)$ for k-steps in order to communicate their types.

This handshake protocol is quite general. For example, consider the illustrative example of two surgeons with different experience levels from the section 1. When two new agents are paired together, they might both choose the cautious strategy for the first couple of surgeries. Over time, surgeon 1 might shift its strategy to balanced and then to fast, signalling to the surgeon 2 that they are experienced and prefer to be fast. If throughout this period, surgeon 2 sticks to being cautious, this handshake would signal to both that the surgeon 1 is experienced, while 2 is inexperienced.

After identifying each other through their initial behaviour, the agents play $s((\theta_i, \theta_{-i}))$ for the remaining $T - k$ steps. The agents must still ensure (authenticate) their partner does not deviate from $s((\theta_i, \theta_{-i}))$ for safety against adversarial "imposter agents" outside C which can still play $\kappa(\theta_{-i})$, posing as a member of C. Since playing a PONE jointly will lead to low regret for both, if i's regret exceeds a certain threshold, this would indicate $-i$ is deviating from s significantly. The threshold can be chosen by the aid of the following lemma,

Lemma 3.4. *For any $\delta, T > 0$, if both players follow strategy $s(\boldsymbol{\theta})$ at each stage, then with probability at least $1 - \delta$ we have*

$$\bar{R}_i^{ext}(h_t; \theta_i) \leq \sqrt{2T \ln \frac{2}{\delta}} \quad \textit{and} \quad R_i^{ext}(h_t; \theta_i) \leq 2\sqrt{2T \ln \frac{4}{\delta}}, \tag{5}$$

which follows from an application of the Azuma-Hoeffding inequality (shown in supplementary material section 1.1). Then the question is what safe strategy should the i fall back into, if the rule is triggered. We base the fallback strategy on the *multiplicative weights* [12] update rule, defined as:

$$s^i_{\text{mw},k}(h_t; \theta_i) \propto s^i_{\text{mw},k}(h_{t-1}; \theta_i) \exp\left(-\eta G(k, a_{t-1}^{-i}(h); \theta_i)\right) \tag{6}$$

for $k \in N$, where $s^i_{\text{mw}}(h_0; \theta_i)$ is the uniform strategy. Define $\pi^{\text{mw},T}$ as the agent that plays $s^i_{\text{mw}}(h_t; \theta_i)$ with learning rate $\eta = \sqrt{8 \ln(N/T)}$. The expected external regret of $\pi^{\text{mw},T}$ is bounded as

$$\bar{R}_i^{\text{ext}}(h_T; \theta_i) \leq \sqrt{\frac{T}{2} \ln N} \tag{7}$$

surely [6, Theorem 2.2]. We then define the agent's overall meta-strategy $\pi^{T,\epsilon}$ as follows:

(1) In first k steps, play $\kappa(\theta_i)$.
(2) If $-i$'s behaviour in h_k not compatible with $\kappa(\theta)$ for any $\theta \in \Theta$, switch to $\pi^{\text{mw},T}$ for all subsequent stages.
(3) While $\bar{R}_i^{\text{ext}}(h_t; \theta_i) \leq k + \epsilon(T-k) - \sqrt{\frac{T-k}{2} \ln N} - 1$, play $s_i(\boldsymbol{\theta})$.
(4) Otherwise, switch to $\pi^{\text{mw},T}$ for all subsequent stages.

The theorem below shows that agents that follow the meta-strategy above form a socially intelligent class among themselves. All proofs have been deferred to the supplementary material section 1.

Theorem 3.5. *For any $\delta, T > k$, let $\epsilon_0 \geq \sqrt{\frac{2}{(T-k)} \ln \frac{2}{\delta}}$, and let $\epsilon_1 = \epsilon_0 + \sqrt{\frac{1}{2(T-k)} \ln N} + \frac{1}{(T-k)}$. Then for $\epsilon = \epsilon_1 + \sqrt{\frac{(T-k)}{2} \ln \frac{1}{\delta}}$, the π^{T,ϵ_1} is (δ, ϵ, T)-socially intelligent.*

4 LEARNING TO COOPERATE

Going forward, we will assume that our agent (henceforth referred to as the "AI agent") will take the role of agent 1, while the other agent (referred to as the "partner") will be agent 2. Our goal is to choose a meta-strategy for the AI agent that can cooperate with a partner drawn from some *target population* nearly as effectively as agents from this population cooperate with one another. For the class of games $\mathcal{G} = (\mathcal{I}, \mathcal{A}, \Theta, G, T)$ as defined in section 3, we will let the target population be a set C of agents forming a (δ, ϵ, T)-SI class with respect to Θ. Ideally, we would hope to choose an AI meta-strategy π that can cooperate with C *without* any additional information about the strategies in C. Looking at the handshake protocol example in Section 3.3, we can see that in many cases a population is likely to use arbitrary conventions to coordinate their behavior, and intuitively we would imagine cooperation to be impossible without prior knowledge of these conventions. (We make this intuition formal in Theorem 4.5).

We therefore consider the problem of learning a cooperative meta-strategy using prior observations of members of the target population interacting with one another. We define a *social learning problem* by a tuple $\{\mathcal{G}, C, \rho, \mu\}$, where C is the target population (SI w.r.t. Θ), ρ is a distribution over C, while μ is a distribution over the joint type space $\Theta \times \Theta$. We can think of C as the set of possible strategies that any member of the target population might follow, while ρ is the frequency of those strategies within the population. To choose an AI strategy, we leverage a dataset $\mathcal{D} = \{(\theta_1^j, \theta_2^j, h_T^j) | j \in [n]\}$ covering n *episodes* of length T. In each episode j, two agents π_j^1 and π_j^2 are sampled independently from ρ, and played together under the joint type $\boldsymbol{\theta}^j \sim \mu$. The AI agent observes the full history h_T^j, along with the agents' types θ_1^j and θ_2^j. We denote a specific learning algorithm as a data conditioned strategy $\pi(\mathcal{D})$.

4.1 Altruistic Regret

We seek an AI strategy that minimizes the regret relative to some Pareto optimal solution to $G(\boldsymbol{\theta})$. Rather than minimizing regret in terms of the AI's own payoffs, however, we seek to minimize *partner's* relative to their (worst case) PONE in $G(\boldsymbol{\theta})$. We formalize this regret with the following definition:

Definition 4.1 (Altruistic Regret)**.** Let $(\sigma_i^*, \sigma_{-i}^*)$ denote the PONE with the *lowest payoff* for the agent $-i$ where $i \in \{1, 2\}$. The altruistic regret of agent i is defined as

$$R_i^{\text{alt}}(h_T; \theta_{-i}) = \sum_{t=1}^{T} G(\sigma_i^*, \sigma_{-i}^*; \theta_{-i}) - G(a^i(h_t), a^{-i}(h_t); \theta_{-i}). \quad (8)$$

In practical cooperation tasks, we would expect outcomes that have low regret for the partner will have low regret for the AI agent as well.

The cooperation objective for the AI agent can then be formalized as minimising the altruistic regret. Unlike the definition suggests, the AI agent must know its own type as well. This is due to the fact that as seen in the handshake protocols example, if the AI agent fails to imitate a human of its type or fail to communicate its type correctly, the partner might switch to a safe strategy.

The goal for the AI agent is to minimize its *expected* altruistic regret over partners sampled from ρ and types sampled from μ. The following lemma shows that we can treat the problem of minimizing regret with respect to a heterogeneous population C as that of minimizing regret w.r.t. a single stochastic strategy.

Lemma 4.2. *Let C be a finite set of agents that are (δ, ϵ, T)-socially intelligent w.r.t. type space Θ, and let ρ be a distribution over C. There exists a mixed strategy $\bar{\rho}$ that forms an (δ, ϵ, T)-socially intelligent class, and which is equivalent to playing against partners sampled from ρ in expectation.*

Proof. In a perfect recall game, every behavioural strategy has an equivalent mixed strategy and vice-versa [2]. Thus ρ can equivalently be defined as a distribution over mixed strategies so that $\rho \in \Delta(\Delta(N))$. Then defining $\bar{\rho}(a) = \int_{\Delta(N)} \sigma(a)\, d\rho(\sigma)$ where $a \in [N]$ denotes a pure strategy (i.e. action) completes the proof.

In order to show the joint impact of consistency and compatibility on the learning problem, we discuss the cases where the population is either consistent or compatible, but not both.

4.2 Consistency without Compatibility

Assume that C consists of agents that are consistent but not necessarily compatible. The most general class in this case is the class of all no-external-regret learners (no-regret henceforth). It is a well-established result that the long-run average of no-regret learning converges to the set of coarse correlated equilibria. The question is whether the AI agent can learn to do better than a coarse correlated equilibrium when paired with a member of C, using only a dataset $\mathcal{D}$ that consists of histories of play for different Coarse Correlated Equilibria (CCE).

Theorem 4.3. *There exists a consistent yet incompatible class of agents C such that even with an infinite amount of data, in the worst-case, the AI agent suffers constant altruistic regret.*

Proof. The proof follows from the theorem 5.1 of Monnot and Piliouras [20] which shows that given any coarse correlated equilibrium of a two-player normal-form game, there exists a pair of no-regret learners that would converge to it. Since C can be any subset of no-regret learners, we cannot exclude those who converge to inefficient CCE. If the class C contains only the agents that converge to Pareto-inefficient CCE, we cannot hope to learn optimal strategies from any dataset. For example, consider the payoff matrices given in table 3 for two pairs of types. Here, the payoff of each agent depends only on its partner. In both games, there exists only one PONE with payoffs $(3, 4)$. However, every pure and mixed strategy profile is a CCE in both matrices. In the worst case, we may have a class of agents C that only converge to the CCE (B, B). More importantly, the behavior of each agent does not need to carry any information about their type, since each agent's payoff depends solely on its partner. When the AI's strategy is deployed, it will face a partner drawn from C whose type is unknown, regardless of the imitation demonstrations dataset. Since we cannot infer type from behavior any more, there is no way for the AI agent to know

	A	B
A	3, 2	1, 2
B	3, 4	1, 4

(a) The game matrix for types (θ_1, θ_2)

	A	B
A	3, 4	1, 4
B	3, 2	1, 2

(b) The game matrix for types (θ_3, θ_4)

Table 3: A class of games where an agent's payoff depends only on its partner.

which of the two game matrices it is playing. At best, the AI (row player) can choose one of the two PONE with uniform probability and commit to it. There is a 0.5 probability that the AI will play the wrong PONE, incurring constant altruistic regret.

4.3 Compatibility without Consistency

Assume that the members of C are compatible, but not consistent. We can construct such a class as in Section 3.3, with agents using a handshake protocol to exchange type information, and then playing the agreed-upon PONE of the current game. However, if at any any time an agent deviates from this chosen solution, there is no restriction on what strategy each agent will follow from that point forward. The members of C may even employ grim-trigger strategies that "punish" any mistake on the part of the other agent by following a highly sub-optimal strategy. Even if at some point in the future they could potentially switch back to a cooperative strategy (i.e., forgive the other agent), this may not occur within the finite horizon T. A single mistake at any time on the part of the AI agent may yield the maximal altruistic regret for the remainder of the interaction. The AI must therefore learn to imitate at least one member of C perfectly using the dataset $\mathcal{D}$, and the problem of learning to cooperate reduces to imitation learning (specifically the *no-interaction* setting of Rajaraman et al. [24]).

We can derive a lower bound on the altruistic regret in this case by considering a game in which there is only a single type (such that individual payoffs are common knowledge), and each agent's payoffs depend only on their own actions. Specifically, the first $N-1$ actions each yield a payoff of 1, regardless of the other agent's action, while the Nth action yields a payoff of 0. In this game we can construct a compatible class C such that, for the first $k \leq T$ steps, the agents execute some "authentication protocol", which allows them to identify other agents following strategies in C. For the first $k-1$ steps, each agent samples one of the first $N-1$ actions, with the sequence of actions forming a *challenge code* that the other agent must respond to by selecting the correct action at step k. If an agent's partner fails to provide the correct response at step k, the agent will follow the Nth action for the remaining $T-k$ steps, such that it receives no further payoff from that point forward. Using such strategies, and an approach similar to that of [24], we can derive a lower bound on the altruistic regret as a function of the number of samples in the dataset $|\mathcal{D}|$.

Theorem 4.4. *Let $K = |\mathcal{D}|$ be the number of interaction histories in the dataset. For any $k < T$, and any $\delta, \epsilon \geq 0$, there exists a class of games $\mathcal{G}$, and class C of (δ, ϵ, T)-compatible agents such that, for any data-dependent meta-strategy $\hat{\pi}(\mathcal{D})$, the altruistic regret is lower-bounded as*

$$E\left[R_i^{alt}(h_T; \theta_{-i})\right] \geq \frac{T-k}{e}\frac{N-2}{N-1}\min\left\{\frac{1}{2}, \frac{(N-1)^{k-1}-1}{2K+1}\right\}, \quad (9)$$

where the expectation is taken over h_T, $\boldsymbol{\theta}$, and $\mathcal{D}$. Then, for small altruistic regret, the sample complexity grows exponentially in k.

Proof sketch. We choose an "authentication" function $f : [N-1]^{k-1} \mapsto [N-1]$ that maps each possible $(k-1)$-step history of actions to a specific action in $[N-1]$. We then construct a class C consisting of a single meta-strategy that, for the first $k-1$ steps selects its actions so that the initial $k-1$ step history of its actions is distributed according to a specific, nearly uniform distribution μ. At step k, agent i chooses action $f(h_{k-1}^{-i})$, where h_{k-1}^{-i} is the sequence of actions chosen by the other agent $-i$. So long as agent $-i$ response with the correct action $f(h_{k-1}^{i})$ at step k, agent i will continue to choose actions in $[N-1]$. Therefore, in self-play C will be (δ, ϵ, T)-compatible for *any* $\delta, \epsilon > 0$.

The AI, however, is unaware of f, and must estimate this function from $\mathcal{D}$. If the AI's strategy fails to correctly authenticate at step k, its partner will switch to the Nth action, which yields a payoff of zero, such that the AI will suffer an altruistic regret of $T-k$. Because f is deterministic, a meta-strategy found via imitation learning will correctly authenticate for any history h_{k-1}^{-i} found in $\mathcal{D}$, but has a probability of $1-1/(N-1)$ of failing to authenticate for an unseen history. By sampling from a carefully chosen distribution μ, we can ensure that the probability of encountering an unseen history is greater than $\min\left\{\frac{1}{2}, [(N-1)^{k-1}-1]/(2K+1)\right\}$, which leads immediately to the lower-bound on the expected altruistic regret. Note that we can choose any $k < T$ so as to maximize this lower bound for any values of T, N and K.

4.4 Lower Bound for Socially Intelligent Populations

Theorem 4.5. *Let $K = |\mathcal{D}|$. For any $\delta, \epsilon > 0$, there exists a class of games $\mathcal{G}$, and class C of (δ, ϵ, T)-socially intelligent agents such that, for any data-dependent meta-strategy $\hat{\pi}(\mathcal{D})$, the altruistic regret is lower-bounded as*

$$E\left[R_i^{alt}(h_T; \theta_{-i})\right] \geq \Omega\left((T-k)\min\left\{\frac{1}{2}, \frac{(N)^{k-2}-1}{2K+1}\right\}\right), \quad (10)$$

for some $k \geq T\epsilon$.

Proof sketch. Similar to the proof for Theorem 4.3, we can define a class of games in which players must exchange their private types to be compatible, while at the same time they can implement consistent behavior without revealing anything about their types. We can construct a socially intelligent class of agents for this class of games. We can then augment these agents such that they implement a k-step authentication protocol (as in Theorem 4.4) before switching to the socially intelligent meta-strategy if authentication succeeds. If authentication fails, the agents will switch to some alternative consistent meta-strategy. So long as $k-1 \leq T\epsilon$, the resulting class

of agents will be (δ, ϵ, T)-socially intelligent. As discussed in section 4.2, a consistent meta-strategy may never communicate an agent's type. Without knowing its partner's type, the AI agent may suffer arbitrarily large altruistic regret at each step, as it cannot identify the actions that will maximize its partner's utility.

5 UPPER BOUND FOR SOCIALLY INTELLIGENT POPULATIONS

Algorithm 1 The $\tilde{T}$-step *imitate-then-commit* meta-strategy (denoted by $\pi^{IC}_{\tilde{T}}$). It is assumed here that the AI acts as agent 1.

Inputs: Interaction dataset $\mathcal{D}$, imitation time $\tilde{T}$.
Initialize the imitation policy $\hat{\pi}^1_{\tilde{T}}(D)$.
for step $t = 1, \ldots, \tilde{T}$ **do**
  Execute action $a^i_t \sim \hat{\pi}^1_{\tilde{T}}(h_t; D)$
end for
for action $j \in N$ **do**
  $z_j = \sum_{i \in N} \hat{z}(h_{\tilde{T}})_{i,j}$
  for action $i \in N$ **do**
    $x_j(i) = \hat{z}(h_{\tilde{T}})_{i,j} / z_j$
  end for
end for
Sample $x = x_j$ with probably z_j
for step $t = \tilde{T} + 1, \ldots, T$ **do**
  Execute action $a^i_t \sim x$
end for

A key idea behind this work is that against a socially intelligent target population, rather than trying to imitate a member of the population perfectly throughout the entire episode, the AI agent only needs to imitate them long enough to learn about its partner's private type. Once it has this information, the AI agent can leverage the fact that the partner's strategy is consistent against *any* strategy, and try to "coerce" the human partner into playing a strategy that minimizes the altruistic regret. We will refer to such meta-strategies as *imitate-then-commit* (IC) strategies, which use the previous observations $\mathcal{D}$ to learn an imitation strategy that it follows for the first $\tilde{T} < T$ steps of the interaction. In this section we provide an upper bound on the expected altruistic regret of a specific (IC) meta-strategy, as a function of the number of episodes in $\mathcal{D}$, subject to the following assumptions:

Assumption 5.1. For $\delta_0, \delta_1, \epsilon_0, \epsilon_1 > 0$, and $\tilde{T} < T$, we have that

(1) ρ is $(\delta_0, \epsilon_0, T)$-consistent.
(2) ρ is $(\delta_1, \epsilon_1, \tilde{T})$-compatible.

Imitation learning. Under an imitate-then-commit meta-strategy, the sample complexity is defined entirely by the number of episodes the AI agent needs to observe to learn a good $\tilde{T}$-step imitation policy. Fortunately, imitation learning is a well-studied problem, and we can largely leverage existing complexity bounds. The one caveat is that in this setting we need bounds on the total variation distance between the distribution over the partial history $h_{\tilde{T}}$ under the population strategy ρ, and that under the learned strategy. Given the dataset $\mathcal{D}$, we define the imitation strategy $\hat{\pi}^1_{\tilde{T}}(\mathcal{D})$ such that $\hat{\pi}^1_{\tilde{T}}(h; \mathcal{D})$ is the empirical distribution over agent 1's actions for each history h occurring in $\mathcal{D}$, while $\hat{\pi}^1_{\tilde{T}}(h; \mathcal{D})$ is the uniform distribution over N for $h \notin \mathcal{D}$. We also define the *marginal* imitation strategy $\hat{\pi}^1_{\tilde{T}} = \mathrm{E}_{\mathcal{D}}[\hat{\pi}^1_{\tilde{T}}(h; \mathcal{D})]$, where the expectation is taken over the sampling of the dataset $\mathcal{D}$ itself. We then have the following bound on the distribution of $h_{\tilde{T}}$ under the imitation strategy:

Lemma 5.2. *Let $p_{\tilde{T}}$ be the distribution over partial histories $h_{\tilde{T}}$ under the population strategy ρ paired with itself, and let $\hat{p}_{\tilde{T}}$ be their distribution under $\hat{\pi}^1_{\tilde{T}}$ paired with ρ. We have that*

$$\|p_{\tilde{T}} - \hat{p}_{\tilde{T}}\|_{TV} \leq \min\left\{1, \frac{N^{2(\tilde{T}+1)}\tilde{T}\log(K)}{K}\right\}, \tag{11}$$

where $K = |\mathcal{D}|$.

This upper bound follows directly from that of [24] via Lemma 1 of [8] (see supplementary material section 2.1 for full proof). We note that the imitation strategy $\hat{\pi}^1_{\tilde{T}}(h; \mathcal{D})$ marginalizes over agent 1's private type, and so the AI does not need to know its own type.

Imitate-then-commit strategy. For history $h_{\tilde{T}} \in \mathcal{H}_{\tilde{T}}$, we let $\hat{z}(h_{\tilde{T}}) \in \Delta(N \times N)$ denote the empirical *joint* strategy played up to and including step $\tilde{T}$. We show that, given $\hat{z}(h_{\tilde{T}})$, it is possible to construct a *mixture* ν over mixed strategies $x \in \Delta(N)$ such that, in expectation over ν, the partner's payoff under their best response to $x \sim \nu$ will be at least as large as their payoff under $\hat{z}(h_{\tilde{T}})$. The IC strategy described in Algorithm 1 follows $\hat{\pi}^1_{\tilde{T}}(h; \mathcal{D})$ for the first $\tilde{T}$ steps, and then commits to a mixed strategy x for he remainder of the interaction. We then have the following upper bound on the altruistic regret achievable with an imitate-then-commit strategy:

Theorem 5.3. *Given that Assumption 5.1 holds for ρ, if the AI follows $\pi^{IC}(\mathcal{D})$ (Algorithm 1) as agent 1, its altruistic regret satisfies*

$$E\left[\frac{1}{T}R^{alt}_1(h_T, \theta_2)\right] \leq \delta(K) + \epsilon_1 + \delta_1 + \frac{T - \tilde{T}}{T}(\epsilon_0 + \delta_0), \tag{12}$$

where $K = |\mathcal{D}|$ and $\delta(K)$ is defined as

$$\delta(K) = \min\left\{1, \frac{N^{2(\tilde{T}+1)}\tilde{T}\log(K)}{K}\right\} \tag{13}$$

and where the expectation is taken over h_T, $\boldsymbol{\theta}$, and $\mathcal{D}$.

Proof sketch: By Lemma 5.2, we can learn an imitation strategy such that the corresponding distribution over $h_{\tilde{T}}$ and $\hat{z}(h_{\tilde{T}})$ is close to that under ρ in self-play. As ρ is compatible, both agents' payoffs under $\hat{z}(h_{\tilde{T}})$ must be close to those under *some* PONE. Finally, we can construct a mixture ν for agent 1 such that agent 2's payoffs under its (approximate) best-response are almost as large as those under $\hat{z}(h_{\tilde{T}})$ (see supplementary material section 2.2).

6 RELATED WORKS

Our work is closely related to the previous targeted learning model [7, 21, 22],which defines similar compatibility and consistency criteria. The notion of targeted optimality [7] include convergence to learning an approximately best response in a multi-agent model with high probability in a tractable number of steps against a population of memory-bounded adaptive agents. The main difference with our

work is that targeted learning only requires consistency against a specific target class of partners, which generally would not include the agent itself, or other adaptive agents. We require socially intelligent agents to be consistent against all possible partner strategies. We also require that cooperation and consistent learning occur over a fixed time horizon T, rather than asymptotically. These differences mean that a hypothetical "universally cooperative" agent might be able to leverage the consistency of its partner to achieve cooperation without a prearranged convention. "Universal cooperation" in a population is relevant for specific populations as described in our illustrative examples like surgeons in a hospital, workers in a factory etc. In these settings it is reasonable to assume that most agents will be able to cooperate with each other professionally, since if a member fails to do this, they would be not be a member of that population. Furthermore, our model allows agents in the population to possess highly conflicting preferences, and our definition of cooperation only requires that agents identify mutually beneficial joint strategies when these exist. Socially intelligent agents can be modeled as individually rational learners [17] to achieve Pareto-efficient joint behavior. Our research builds on this work by considering a learning setting where the agent when paired with any member of the population will achieve at least the same utility with high probability as the Pareto-efficient approach.

The problem of training agents to be able to cooperate with previously unseen partners is sometimes referred to as *ad hoc teamwork* [19, 27] or *zero-shot coordination* [14], especially in the context of multiagent reinforcement learning. Many approaches in reinforcement learning train cooperative policies that are *robust* to possible strategies that a human or an AI agent can follow [5]. A lot of these methods build a "population" of partner strategies and maximizes the diversity of this population in order to train the AI's policy against it [10, 28]. Other approaches assume that there is no prior coordination between the agents [14] to learn rational joint strategies while estimating the agents' mutual uncertainty about one-another's strategies [30]. Ad-hoc multiagent coordination can be helpful to learn cooperation among AI agents with the "other-play" algorithm [14] that finds such a strategy as a solution to the corresponding *label free coordination* problem [30]. A possible approach to solve these problems can be self-play [31] where the agent can optimize themselves by playing with past iterations of themselves in order to estimate the strategies of unseen partners. However, the "self-play" approach can learn cooperative strategies which can "over-fit" [29] to one another in the population of agents. A key goal of Ad hoc coordination (teamwork) and aligned research in zero-shot coordination work has been to avoid this type of overfitting [9]. Our problem domain is closely related to both *ad hoc teamwork* or *zero-shot coordination*, since we consider training an agent to cooperate with previously unseen partners, and assume no control over the partner. Even though population-based training approaches to ad hoc teamwork are common, they focus on fully cooperative environments such as Dec-POMDPs, where the main issue is creating a diverse enough population to train with [23]. We consider partners that are self-interested, and do not assume identical payoffs.

Finally, in the case of Hannan-consistent partners, our problem setting is closely related to strategizing against and learning to manipulate no-regret learners [4, 11]. This line of work studies whether an optimizer agent can achieve better payoff than CCE against no-regret learners by learning to enforce a Stackelberg equilibria on them. Their emphasis is on online learning and the optimizer's payoff, while we focus on the offline setting and cooperation.

7 CONCLUSION

We provide formal guarantees for successful and reliable cooperation of AI agents with populations of socially intelligent agents. We present a novel definition of social intelligent populations based on the assumptions that 1) members of the population are individually rational, and 2) pairs of members can achieve performance comparable to a Pareto-optimal Nash equilibrium. We formalize the notion of consistency and compatibility of agents in repeated, two-player, general-sum matrix games with private types. Our theoretical guarantees are in the offline cooperation setting where the agent has to cooperate with unseen partners in the population to strategise against and manipulate no-regret policies for which we formalize the idea of altruistic regret. We prove that the assumptions on its own are insufficient to learn *zero-shot* cooperation with partners of the socially intelligent target population. We provide upper bounds on the sample complexity needed to learn a successful cooperation strategy along with lower bounds on when the multi-agent cooperation setting is needed with respect to the populations' trajectories, the state space and the length of the learning episodes. The bounds in these settings of the agent actively querying the MDP without knowing the transition dynamics of the population or the agent observing the populations' transition dynamics are much stronger than the bounds that can be derived by naively reducing the cooperation problem to one of reinforcement learning. These complexity analysis and formally proven bounds can be helpful to sustainably model the alignment problem of AI agents.

ACKNOWLEDGMENTS

Mustafa Mert Çelikok is funded by the Hybrid Intelligence Center, grant number 024.004.022, a 10-year programme granted by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybridintelligence-centre.nl.

REFERENCES

[1] Rapoport Anatol. 1966. A Taxonomy of 2× 2 Games. *Yearbook of the Society for General Systems Research* 11 (1966), 203–214.
[2] Robert J . Aumann. 1964. *28. Mixed and Behavior Strategies in Infinite Extensive Games*. Princeton University Press, Princeton, 627–650. https://doi.org/doi:10.1515/9781400882014-029
[3] Tilman Börgers and Rajiv Sarin. 1997. Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory* 77, 1 (1997), 1–14.
[4] William Brown, Jon Schneider, and Kiran Vodrahalli. 2023. Is Learning in Games Good for the Learners?. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 54228–54249. https://proceedings.neurips.cc/paper_files/paper/2023/file/a9ea92ef18aae17627d133534209e640-Paper-Conference.pdf
[5] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the Utility of Learning about Humans for Human-AI Coordination. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/f5b1b89d98b7286673128a5fb112cb9a-Paper.pdf
[6] Nicolo Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, Learning, and Games*. Cambridge University Press.

[7] Doran Chakraborty and Peter Stone. 2010. Convergence, Targeted Optimality, and Safety in Multiagent Learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* (Haifa, Israel) *(ICML'10)*. Omnipress, Madison, WI, USA, 191–198.

[8] Kamil Ciosek. 2022. Imitation Learning by Reinforcement Learning. In *International Conference on Learning Representations*.

[9] Brandon Cui, Hengyuan Hu, Luis Pineda, and Jakob Foerster. 2021. K-level Reasoning for Zero-Shot Coordination in Hanabi. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 8215–8228. https://proceedings.neurips.cc/paper_files/paper/2021/file/4547dff5fd7604f18c8ee32cf3da41d7-Paper.pdf

[10] Brandon Cui, Andrei Lupu, Samuel Sokota, Hengyuan Hu, David J Wu, and Jakob Nicolaus Foerster. 2023. Adversarial Diversity in Hanabi. In *International Conference on Learning Representations*.

[11] Yuan Deng, Jon Schneider, and Balasubramanian Sivan. 2019. Strategizing against No-regret Learners. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/8b6dd7db9af49e67306feb59a8bdc52c-Paper.pdf

[12] Yoav Freund and Robert E Schapire. 1999. Adaptive Game Playing Using Multiplicative Weights. *Games and Economic Behavior* 29, 1-2 (1999), 79–103.

[13] James Hannan. 1957. Approximation to Bayes Risk in Repeated Play. *Contributions to the Theory of Games* 3, 2 (1957), 97–140.

[14] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. "Other-Play" for Zero-Shot Coordination. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 4399–4410. https://proceedings.mlr.press/v119/hu20a.html

[15] Asad M Lak, Abdullah M Abunimer, Caroline MW Goedmakers, Linda S Aglio, Timothy R Smith, Melvin Makhni, Rania A Mekary, and Hasan A Zaidi. 2021. Single-versus Dual-Attending Surgeon Approach for Spine Deformity: A Systematic Review and Meta-Analysis. *Operative neurosurgery (Hagerstown, Md.)* 20, 3 (2021), 233–241.

[16] Hoang M. Le, Yisong Yue, Peter Carr, and Patrick Lucey. 2017. Coordinated Multi-Agent Imitation Learning. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1995–2003. https://proceedings.mlr.press/v70/le17a.html

[17] Robert Loftin, Mustafa Mert Çelikok, and Frans A. Oliehoek. 2023. Towards a Unifying Model of Rationality in Multiagent Systems. arXiv:2305.18071 [cs.AI]

[18] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic Theory*. Vol. 1. Oxford University Press New York.

[19] Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V. Albrecht. 2022. A Survey of Ad Hoc Teamwork Research. In *Multi-Agent Systems*, Dorothea Baumeister and Jörg Rothe (Eds.). Springer International Publishing, Cham, 275–293.

[20] Barnabé Monnot and Georgios Piliouras. 2017. Limits and Limitations of No-regret Learning in Games. *The Knowledge Engineering Review* 32 (2017), e21.

[21] Rob Powers and Yoav Shoham. 2004. New Criteria and a New Algorithm for Learning in Multi-Agent Systems. In *Advances in Neural Information Processing Systems*, L. Saul, Y. Weiss, and L. Bottou (Eds.), Vol. 17. MIT Press. https://proceedings.neurips.cc/paper_files/paper/2004/file/220a7f49d42406598587a66f02584ac3-Paper.pdf

[22] Rob Powers and Yoav Shoham. 2005. Learning Against Opponents with Bounded Memory. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (Edinburgh, Scotland) *(IJCAI'05)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 817–822.

[23] Muhammad Rahman, Jiaxun Cui, and Peter Stone. 2024. Minimum Coverage Sets for Training Robust Ad Hoc Teamwork Agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17523–17530.

[24] Nived Rajaraman, Lin Yang, Jiantao Jiao, and Kannan Ramchandran. 2020. Toward the Fundamental Limits of Imitation Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 2914–2924. https://proceedings.neurips.cc/paper_files/paper/2020/file/1e7875cf32d306989d80c14308f3a099-Paper.pdf

[25] Peter Schuster and Karl Sigmund. 1983. Replicator Dynamics. *Journal of Theoretical Biology* 100, 3 (1983), 533–538.

[26] Yoav Shoham and Kevin Leyton-Brown. 2008. *Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations*. Cambridge University Press.

[27] Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. 2010. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 24. 1504–1509.

[28] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with Humans without Human Data. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 14502–14515. https://proceedings.neurips.cc/paper_files/paper/2021/file/797134c3e42371bb4979a462eb2f042a-Paper.pdf

[29] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with Humans without Human Data. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 14502–14515. https://proceedings.neurips.cc/paper_files/paper/2021/file/797134c3e42371bb4979a462eb2f042a-Paper.pdf

[30] Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. 2021. A New Formalism, Method and Open Issues for Zero-Shot Coordination. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 10413–10423. https://proceedings.mlr.press/v139/treutlein21a.html

[31] Jaleh Zand, Jack Parker-Holder, and Stephen J. Roberts. 2022. On-the-fly Strategy Adaptation for Ad-hoc Agent Coordination. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1771–1773.