Individualized Mutual Adaptation in Human-Agent Teams

Huao Li[1], *Student Member*, Tianwei Ni[2], *Student Member*, Siddharth Agrawal[2], *Student Member*,
Fan Jia[2], *Student Member*, Suhas Raja[2], *Student Member*, Yikang Gui[2], *Student Member*,
Dana Hughes[2], *Member*, Michael Lewis[1], *Lifetime Member*, and Katia Sycara[2], *Fellow*

***Abstract*—The ability to collaborate with previously unseen human teammates is crucial for artificial agents to be effective in human-agent teams (HATs). Due to individual differences and complex team dynamics, it is hard to develop a single agent policy to match all potential teammates. In this paper, we study both human-human and human-agent teams in a dyadic cooperative task, Team Space Fortress (TSF). Results show that the team performance is influenced by both players' individual skill level and their ability to collaborate with different teammates by adopting complementary policies. Based on human-human team results, we propose an adaptive agent that identifies different human policies and assigns a complementary partner policy to optimize team performance. The adaptation method relies on a novel similarity metric to infer human policy and then selects the most complementary policy from a pre-trained library of exemplar policies. We conducted human-agent experiments to evaluate the adaptive agent and examine mutual adaptation in human-agent teams. Results show that both human adaptation and agent adaptation contribute to team performance.**

***Index Terms*—Human-robot teams, Adaptive Systems, Team coordination, Team performance.**

I. Introduction

In **human-agent teaming** (HAT), humans and agents perform interdependent actions in order to achieve common team goals [1]. In such a setting, intelligent agents are more than tools for assisting humans, but rather independent actors that closely collaborate with human team members [2]. This imposes additional challenges for agents to perform effective team behaviors including communicating essential information [3], and accounting for human individual differences [2]. Human teammates in HAT might have varying skill levels, knowledge backgrounds, and individual preferences [3]. Individual differences lead to complex behavioral patterns and intentions [4] which may make it impossible for an agent to develop a globally optimal policy to fit observed human behavior. Additional challenges to developing agents to team with humans include inconsistencies in human behavior which pose an obstacle to learning, the large number of examples commonly needed to learn in a large and/or continuous state and action space, and the nonstationarity of human behavior.

The basic problem of an agent attempting to team with an unknown human can be divided into two parts: 1) predicting the human's behavior and 2) choosing actions to move the team toward its common goal. One approach to predicting human behavior is to hypothesize a single model, then use it in choosing agent actions [1]. In this case human departures from the model are treated as noise. An alternative is to consider differences in human behavior to arise from differences among human types providing multiple models. The agent's problem then becomes identifying the human's type, in order to use that type for choosing agent actions.

This paper follows the second approach. Our Team Space Fortress task has two interdependent roles of bait and shooter. Agents are trained under a variety of regimens for each of the roles, creating a diverse set of types for each role. The agents are then paired in self-play to determine each type's best partner. So, for example, for a bait of type i there will be some shooter of type j providing the highest scores in self-play. The problem of teaming with an unknown human then becomes one of identifying the most similar agent-type to the human and choosing its predetermined complement to provide an agent policy which is optimal within the set of available agent-types.

In this paper, we investigate real-time adaptive agents that coordinate with human teammates in a nontrivial strategic game, Team Space Fortress (see Sec. III). We initially studied *human-human* teams to explore characteristics of performance and teamwork at the TSF task in Sec. IV. The complementary policies and complicated team dynamics found in human teams served as guidance for developing agents compatible with individual dif-

[1]School of Information Science, University of Pittsburgh, PA, USA. Email: hul52@pitt.edu, mlewis@sis.pitt.edu
[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. Email: tianwein, katia@cs.cmu.edu, siddhara, danahugh@andrew.cmu.edu

ferences among human policies in *human-agent* teams. In Sec. V, we propose an adaptive agent based on an exemplar policy library and similarity measurement. We evaluated our approach in Sec. VI by having human players play with both agents with exemplar static policy and adaptive policy. Experimental results show that our method is able to identify human policies and predict team performance accurately, with the similarity between the agent-type nearest the human shooter's trajectory and the optimal agent-type accounting for 70% of the variance in team performance. Our proposed online adaptive agent outperforms a random adaptation baseline by 42.3% and achieves stable team state faster than a non-adaptive baseline.

Contributions:

- This paper introduces a novel method treating human behavior as arising from a factored MDP in which task-related components are captured by a library of diverse policies circumventing both problems of task irrelevant actions and nonstationarity of intentions.
- Highly accurate prediction of Human-Agent performance from Agent-Agent self-play validates our optimization by proxy approach.
- Although widely used in Opponent Modeling and Game Theory, application of type-based reasoning to teamwork is largely novel

II. Related Work

Early work in multiagent and human-agent teamwork such as STEAM [5], or Playbook [6] relied on shared protocols and choreographed joint actions to achieve coordination. While efficient, pre-coordinated systems have difficulty incorporating new members who may not conform fully to the system's protocols. This is particularly true for humans [5]. This paper focuses on the complementary problem of enabling agents to adapt their behavior in real-time to observed behavior of a (previously unseen) human team member. Developing agents capable of real-time adaptation in response to observed human behaviors is an active area of interest in several important domains, including human-robot interaction [7], [8], shared autonomy [9], and autonomous driving [10].

The problem of adapting to a previously unknown human is an instance of the more general problem of Ad Hoc Teamwork, cooperating with any previously unknown teammate introduced by [1]. The difficulties of the problem arise if: 1) the Ad Hoc agent does not initially know how its teammates behave 2) yet must choose its own actions so as to influence these teammates to optimally guide the team to its goal. In their example [1] simplifies the problem by postulating the unknown teammate as a naïve but greedy learner reducing the effort to finding optimal actions for the Ad Hoc teammate. Later work follows a similar course considering uncertainty in identifying teammate behavior (choosing between best response vs. another Ad Hoc agent) [11] or when to communicate [12]. While some work such as the naïve but greedy learner in [1] hypothesize a human model directly, a majority of work within robotics has modeled human intentions based on assumptions about human policies (e.g. stationarity and rationality) and then trains robot policies in accordance with these human models [8].

These normative approaches to adaptation in human-agent and human-robot teams develop explicit models of human decision-making in the context of a shared-reward Markov game with partial information, inferring the underlying reward function (i.e., goals and intents) of the human based on observed behavior; the robot or agent policy is then derived from a partially observable Markov decision process, POMDP, which maintains beliefs over the human's reward function [13]. Many of these models assume human policies are stationary, Boltzmann rational, and are generated with ideal understanding of the environmental dynamics. Recent research has been focused on relaxing these assumptions to better capture real-world human behaviors, such as considering non-stationary human rewards [14], mutual adaptation [15], risk-sensitivity [13], imperfect understanding of environment dynamics [9], or more representative models of rationality [16]. Normative approaches to human behavior suffer from several drawbacks that are relevant to our work.

- Individual differences among human behaviors can be very large making them difficult to capture within a single normative model.
- Long agent training times make human-in-the-loop training impractical and real-time adaptation infeasible.
- Explicit models do not scale well to large and/or continuous state and action spaces.
- Inconsistency and/or nonstationarity in human behavior may violate assumptions made in inferring a reward function leading to either failure to converge or a model unrepresentative of human behavior

An alternative borrowed from Bayesian Game Theory posits, not one, but multiple possible types to capture the large range of individual differences among humans by focusing on a relatively small set of policies in the infinite space of possible behaviors. Ample historical data suggests there may be very large variations in human policies, even among those leading to equivalent performance. While there is a long history of type-based reasoning in Game Theory and opponent modeling, applications involving teamwork are rarer. Barrett and Stone [17] implement an algorithm that learns policies to cooperate with past teammates and then reuses these policies to quickly adapt to new teammates in pursuit-evasion [18] and robot soccer [17]. Initial policies are obtained through a variant of Q-learning, while a distribution of team-types is hypothesized and updated with each observation to choose a learned policy to execute. [4] further addresses the problem of non-stationarity among types by executing a type check at each timestep assessing the posterior distribution over types using a convolutional neural network for change point detection to identify behavior switching. [7], by contrast, adopts an a priori definition of human types as things such as expertise and stamina and demonstrate performance improvements for a simulation of humans working with robots. [19], by contrast, follows a factored state approach to types using unsupervised learning to cluster human trajectories by types. Clusters are then used as the basis for learning policies uncontaminated by latent human states such as goals, trust or attention.

[20] developed an algorithm combining stochastic and Bayesian games with the Bellman optimality equation to provide a definition for type-based reasoning. Their original paper included human experiments in which the algorithm beat human opponents and outperformed alternative reasoners at 'paper, scissors, and rock' and Prisoner's Dilemma. In a later paper [21] the authors use their algorithm to define different belief formulations and analyze their convergence properties. In their construction, posterior beliefs about the relative likelihood of types are formed by comparing predictions of types with observed actions. The beliefs and types are then used to find an action which maximizes expected payoffs. Our method differs in that rather than updating with Bayesian improvement in estimates it uses a sliding window and an unchanging uniform prior in the form of the library (section V). This difference appears to make our approach immune to effects of prior beliefs but cannot guarantee convergence to correct beliefs.

III. Team Space Fortress

We have adapted Space Fortress, a game which has been used extensively for psychological research, for teams. Team Space Fortress (TSF) is a cooperative computer game where two players control their spaceships to destroy a fortress [22]. The player can be either human or (artificial) agent, thus there are three possible combinations in teams: human-human, human-agent, and agent-agent.

A sample screen from the game is shown in Fig. 1. At the center of the screen lies a rotating fortress. The fortress and two spaceships can all fire missiles towards each other when they are within a certain range of each other. The first spaceship entering the activation region (the hexagon area) will be shot at by the fortress. The fortress is protected by a shield (hexagonal border around it). The back of the shield opens as the fortress is firing (see Fig. 1) thus becoming vulnerable to being shot through the opening by the spaceships. Spaceships die immediately whenever they hit any obstacles (e.g. boundaries, missiles, the fortress). The game resets every time either fortress or both players are killed. Once the fortress has been killed, both players must leave the activation region before the game respawns. The team performance is measured by the total number of fortress kills that the team achieves in each 1-minute game.

In order to test heterogeneity and tight teamwork, players were assigned roles of either *bait* or *shooter*. The *bait* enters the activation region first and tries to attract the fortress's attention, i.e. becoming vulnerable to being shot by the fortress. When the fortress attempts to shoot at the bait, its shield lifts making it vulnerable. The other player in the role of *shooter* can now shoot at the fortress and destroy it. This sequential dependency of the game requires two players to collaborate closely to achieve the common goal. Spaceships in the initial human-human experiment used 2nd order dynamics without deceleration making them very difficult to control. Friction was added in later experiments to facilitate training agents as well as stabilize human performance. Results from the first experiment, therefore, can provide insight into the complementarity of roles or adaptation but cannot be compared directly to later experiments.

IV. Human-Human Teams

To better understand the team dynamics and overall difficulty in TSF, we first collected data from human-human teams. This experiment shed light on the general strategy of adaptation when humans paired with unknown teammates. Analyzing human teamwork also gives us a baseline when designing agents in terms of reward function and adaptive behaviors. The results of this human teamwork experiment are reported in more detail in [23], [24]. Here we briefly discuss the insights it brings to our work in human agent teaming.

1) Performance analysis: In TSF, baits and shooters have their own role-specific tasks but also must coordinate their positions and sequences of actions. Therefore players' individual capabilities are confounded with

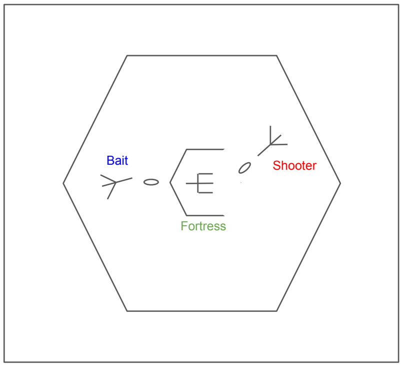


Figure 1: Sample TSF game screen (line drawing version, original screen is in black background). Spaceships are labeled as shooter and bait. Entity at center is the rotating fortress with the border around it as the shield. Activation region is the hexagon area around players' spaceships. One projectile (the ellipsoid) is emitted from the shooter towards the fortress, and another one is from the fortress towards the bait. All the entities are within the rectangle map borders.

their adaptive behaviors when evaluating performance (non-stationarity). To disentangle these effects, a special design was employed which assigns two baits and two shooters to a *squad*. Shooter-bait pairings within a squad were exchanged halfway through the experiment so that both individual and team performances could be observed in a controlled way. We conducted a two-way ANOVA analysis with shooter and bait as two factors to test the contribution of each role to team performance and their interaction effect. Results show that the team performance in TSF is influenced by both players' individual skill level and their ability to collaborate with different teammates. In some teams, pairs of good individual players produced high team performance, which means the team performance was largely determined by how good each individual was. However, interaction effects found in other teams revealed the contribution of team dynamics on performance. In other words, good teams are not simply the addition of two good individual players, but also rely on good coordination such as adaptive actions and complementary strategies. These observations guide our development of agents for HAT that adapt to human individual differences in skill level and playing style.

2) Adaptation analysis: There are multiple possible explanations for the good and bad performance of pairings in addition to players' individual skill level. One hypothesis is that some players are good at adapting their policies to different teammates, therefore they can change their policies rapidly after team reorganisation [23]. Analyzing the similarity between players' trajectories when pairing with different partners could help uncover the relationship between individual adaptation and team performance. In order to quantitatively measure the similarity and distinctness among trajectories, we obtained third-party human judgements on the online crowd-sourcing platform Amazon Mechanical Turk. Results showed that players who changed their policy when switching partners tend to have a better team performance. This finding indicates that individual adaptive actions contribute to team coordination and therefore improve team performance.

V. Agent Adaptation

In this section, we formulate our method for agent adaptation. First we introduce a factored state model of human behavior and its relation to the policy library. Next we introduce the policy library that we developed by training agents in the bait and shooter roles and evaluating their team performance in agent-agent play. This library will be used by the agent adaptation process, for human policy identification and adaptive policy generation. Third, we introduce the policy similarity metric adopted for human policy identification that measures the distance between human and exemplar policies from human-agent trajectories. Finally, we propose an agent adaptation method based on similarity between human and exemplar policies.

A. Human Model

Following [25], [26] we consider human behavior to depend on latent states (such as goal, trust, attention) in addition to the observable states of the task. As the latent factors are unrelated to the task they can contaminate estimates of a type's policy. [25] addresses this problem by clustering trajectories in order to learn policies for clusters rather than from noisy human trajectories. We take the alternative approach of independently learning a diverse set of role-satisfying policies corresponding to potential task-relevant components of human states. If indeed this coverage is wide enough to subsume the role-relevant components of human policies, then our beliefs about human types (policy matches) and shifts between types (nonstationarity) should be correct.

B. Exemplar Policies Library

The exemplar policies library $\mathcal{L} = \mathcal{L}_B \cup \mathcal{L}_S$ consists of two sets of policies in bait ($\mathcal{B}$) and shooter ($\mathcal{S}$) roles, $\mathcal{L}_B$ and $\mathcal{L}_S$ respectively. Both bait policies and shooter policies are trained using a combination of imitation learning (IL), reinforcement learning (RL), and rule-based methods. To make these different policies diverse, we train them using different reward functions, inspired by human-human experiments where there were multiple

Agent Role & IDs	Method	Algorithm	Reward	Obs dim	Act dim	Hyperparameters
B1-B3	RL	A2C [27]	Alive bonus	6	1	Target Speed
B4-B6	RL	A2C, PPO, TRPO [28]	Alive bonus	8	2	Algorithm
B7	RL	PPO [29]	Alive bonus	12	2	N/A
B8-B9	RL	PPO	Distance+Angle+Death	16	2	Reward weights
B10-B13	IL & RL	SQIL [30]	Alive bonus	9	2	Alive bonus+loss weight
S1-S3	RL	DDQN [31]	Distance+Angle	3	1	Target Speed
S4-S7	Rule	Mirror strategy	N/A	N/A	N/A	Distance threshold
S8-S11	IL & RL	SQIL [30]	Fortress kill	8	2	Kill bonus+loss weight

Table I: Brief descriptions of the agents in the library grouped by learning methods used to create different strategy types. For each type of agent, we describe its method, training algorithm, reward design, observation and action dimension, hyperparameters that generate the agent instances in that type, and corresponding agent IDs where the prefix B and S are referred to as Bait and Shooter agents, respectively.

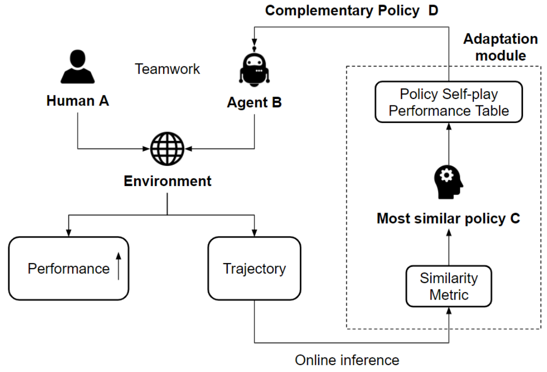


Figure 2: The flowchart of the proposed adaptive agent architecture. The adaptation module (in dotted border) takes the input of the trajectory at current timestamp, and then assigns the adaptive agent with new policy at next timestamp. The adaptation procedure can be deployed in real-time (online).

ways to achieve good performance. There are in total 11 shooter policies/types and 13 bait policies/types in the exemplar policies library. Their brief descriptions are shown in Table I.

1) Reinforcement Learning policies: For RL-based policies, we design the reward functions attempting to encode the desirable behavior of a player in either role. Agent policies are then trained in an agent-only task environment to achieve an optimal behavior with respect to the given reward function. RL-based baits are trained with an *alive* bonus reward, which encourages the bait agents to learn to survive within the activation region while making the fortress vulnerable to the shooter. The RL shooters' reward functions are based on the prior knowledge of TSF game that a good shooter should have a position *opposite* the Bait, which was observed in many successful human-human teams. Other individual differences we found in human game play are implemented as hyperparameters to create diverse policies. For example, the observation that some players adapt to their teammate's action is implemented as whether the agents are able to perceive the teammate's position and velocity in their observation space. Those combinations of RL methods, reward functions and hyperparameters provide in total 12 policies in both shooter and bait roles.

2) Rule-based policies: Rule-based methods are used in developing *shooter* agents with a mirror strategy. The mirror shooters adopt the team strategy of *opposite positions*. They try to keep opposite the current position of the bait, on the back side of the Space Fortress. When an opportunity to destroy the fortress arises, it stops mirroring and fires towards the fortress. By controlling the threshold of distance to the target position, we derived agents S4-7.

3) Imitation Learning policies: Lastly, to include more human-like policies into the library, we trained policies based on soft Q-imitation learning (SQIL) [30] using the demonstrations collected from pilot human-agent experiments. Details about this human data collection are introduced in Sec. VI. SQIL is a pure imitation learning method, so it does not consider the task reward. However, when the human demonstrations are far from optimal, they may hinder learning [32]. We introduced a simple improvement upon SQIL to mitigate this issue: we incorporated a pre-defined task reward function into the objective as inductive bias, resulting in what could be viewed as a mixture of IL and RL policies. We trained both bait and shooter SQIL agents with hyperparameter tuning on the weight between IL and RL reward, and selected the best four agents, namely B10-B13 and S8-S11.

4) Self-play performance: We evaluated the performance of each shooter-bait pair in the exemplar policy library by self-play in the TSF environment, and recorded the results in the self-play performance table $\mathcal{P}$. The table $\mathcal{P}$ has rows with the number of bait policies in $\mathcal{L}_B$ and columns with the number of shooter policies in $\mathcal{L}_S$, with each entry the average performance of the bait-shooter pair. When applied to our policy library, a subgroup of table $\mathcal{P}$ is showed in Table II.

Similar to human-human teams, agent-agent teams also show complementary policy pairs that work extremely well with each other. An example would be B6 (RL bait policy) which yielded a dominant performance when pairing with most of the shooters except for S9 and S11. While for specific shooter policies such as S9 and S11 (IL shooter policies), the best teammate would be B11 (IL bait policies) instead of the more "optimal" B6. Inspection of the self-play table shows that the space of reasonable policies in the TSF game is indeed diverse, and there are multiple ways to achieve good team dynamics and team performance. This confirms again the potential gains from introducing real-time adaptive agents into human-agent teams.

	S1	S2	S3	S4	S7	S9	S11	Avg
B3	5.8	6.6	5.7	**8.1**	**8.1**	2.3	1.9	5.5
B6	6.0	**6.9**	**6.0**	**7.9**	7.6	3.1	3.5	**5.9**
B7	**6.1**	6.6	5.7	7.7	**7.7**	3.0	3.1	5.7
B8	4.6	**5.1**	**5.1**	3.2	3.0	2.1	2.6	3.7
B9	4.6	**5.5**	4.7	2.8	2.7	1.8	2.3	3.5
B11	5.4	**6.3**	5.3	5.7	6.0	**3.5**	**3.6**	5.1
B13	5.3	6.0	4.9	6.1	**6.5**	3.0	3.0	5.0
Avg	5.4	**6.1**	5.3	5.9	5.9	2.7	2.8	4.9

Table II: Self-play agent performance table. Each row is for one bait agent policy named B_i in $\mathcal{L}_B$ ($i \in [1, 13]$), and each column is for one shooter agent policy named S_j in $\mathcal{L}_S$ ($j \in [1, 11]$). Each entry is computed by per-minute team performance (number of fortress kills) of the corresponding pair. We mark the "optimal" shooters in bold teaming with a given bait, and the "optimal" baits teaming with a given shooter.

C. Similarity Metric

Now we introduce the **cross-entropy metric** (CEM) as the policy similarity metric used for our agent adaptation process. Cross-entropy, well-known in information theory, can measure the (negative) distance between two policies π_1, π_2:

$$\text{CEM}(\pi_1, \pi_2) = E_{s,a\sim\pi_1}[\log \pi_2(a|s)] \tag{1}$$

where $\pi_1(\cdot|s), \pi_2(\cdot|s)$ are action distributions given state s.

If the policy π_2 and state-action samples from π_1 are obtained, we can then estimate the cross-entropy between two policies $\text{CEM}(\pi_1, \pi_2)$ by Monte Carlo sampling, even without access to the target policy π_1. In human-agent teaming, human policy π_H, though unknown, can be compared using CEM as the similarity metric with each agent policy π_A in the library because the state-action pairs generated by the human policy can be easily obtained.

Given a sliding window of frames that record the observed behavior of the human policy π_H, we can estimate the CEM between a human policy π_H and any known agent policy π_A by the following formula:

$$\frac{1}{T}\sum_{t=1}^{T} \log \pi_A(a_t|s_t), \quad \text{where } (s_t, a_t)_{t=1}^{T} \sim \pi_H \tag{2}$$

where $(s_t, a_t)_{t=1}^{T}$ are the sequential state-action pairs from human policy play, T is the window size, a hyperparameter to be tuned. We believe this approach of referencing task-relevant aspects of human behavior to an agent model is novel and potentially applicable to human-human comparisons as well, particularly for capturing behavioral alternatives that may have crucial implications for safety or reliability but are missed by normative measures.

D. Adaptive Agent Method

The prerequisite for the architecture is the exemplar policies library $\mathcal{L}$ introduced in the Sec. V-B and the self-play table $\mathcal{P}$ of the library to translate human-agent performance in the adaptation process.

Figure 2 shows the overall flowchart of our adaptive agent framework. When the game starts and a new human player A starts to play as one pre-specified role $R_1 \in \{B, S\}$ in TSF, the adaptive agent framework will first randomly assign a policy B from the library $\mathcal{L}_{R_2}$ in teammate role R_2 such that $\{R_1, R_2\} = \{B, S\}$, and keep track of the joint trajectories (state-action sequences) and record them into memory.

The adaptation process is as follows. As we maintain the latest human trajectories of a pre-specified window size, we use the data to compute the similarity by cross-entropy metric between the human trajectory and any of the exemplar policies in the library $\mathcal{L}_{R_1}$ with same role. Then we find the most similar policy $C \in \mathcal{L}_{R_1}$ to the human trajectory, and look in the performance table $\mathcal{P}$ to find the optimal complementary policy $D \in \mathcal{L}_{R_2}$ to the predicted human policy type C. Finally, we assign the agent D as the complementary policy at next timestamp with the human player.

The adaptation process on the exemplar policies selection is based on the following assumption: if the human policy A with role R_1 is similar to one exemplar policy $C \in \mathcal{L}_{R_1}$ within some threshold, then the human policy A will have similar team performance with teammates as C, i.e., if C performs better with $D \in \mathcal{L}_{R_2}$ than $E \in \mathcal{L}_{R_2}$, so does A. This enables us to adapt the agent policy in real-time from recent data without modeling the human policy directly.

VI. Human-Agent Teaming Experiments

In this section, we first introduce our experiment design for human-agent teaming, then evaluate the human-agent performance when paired with static policy agents (introduced in Sec. V-B) and proposed adaptive agents (introduced in Sec. V-D).

By analyzing the collected human-agent data, we aim to answer the following motivated questions:

1) How well do human players' policies correspond to agent policies in our library?
2) Is our adaptive agent architecture capable of identifying human policies and predicting team performance for human-agent teams?
3) Do our adaptive agents perform better than static policy agents in human-agent teams?

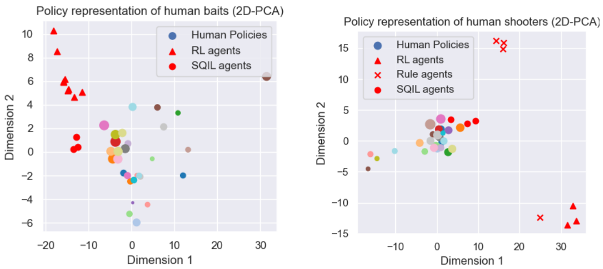
A. *Experiment 1: Static policy agents*

1) Experimental Design: We recruited 104 participants from Amazon Mechanical Turk for our first human-agent experiment. They were paid USD 2 for participating in the 15-min online study. Participants were randomly assigned a role of either shooter or bait and then teamed in randomized order with five artificial agents in the opposite role. Participants completed three 1-min game trials with each agent. The five variants were selected from our static agent library $\mathcal{L}$ balanced for performance in the self-play table and diversity by considering different training methods and reward functions. Specifically, we selected {B3, B6, B7, B8, B9} to test as static baits and {S1, S2, S3, S4, S7} to test as shooters. In the dataset of human and static agent teams, we collected 25 valid sessions from human shooters and 29 valid sessions for human baits.

2) Policy space representation: To quantify the relationship between real human policies in the experiments and agent policies in the library, we compared the distance between the collected human trajectories and agent policies using CEM measurement (see Sec. V-C). This provides us with a high-dimensional policy space w.r.t. exemplar policies in our library. Then, we applied principal component analysis (PCA) based on the log-probability dataset to project the high-dimensional policy space into a 2D plane for a better visualization. The two primary components left explain more than 99% of the variance.

Fig. 3 illustrates where observed human policies fell in relation to the static agent dataset. Policies in the library cover a wide space and subsume the distribution of human policies with particular overlap between human and SQIL policies. Additionally, the distribution of human policies correlates with their team performance (circle size) in that human baits (see Fig 3a) perform better at the left of the figure while human shooters performed better when located at the center (see Fig 3b).

3) Human policy identification: In the proposed adaptive agent architecture, our model infers human policy by associating it with the most similar policy in the library based on CEM measurement, then assigns the agent with the corresponding complementary policy in the self-play table. One way of verifying this method is to see if human-agent teams performed better when the predicted human policy was closer to the complementary match (i.e. best partner) in the self-play table $\mathcal{P}$. Assuming each human maintains a consistent policy over each 1-min interaction when paired with a specific teammate with static policy, we could then calculate, for each human-agent pair, the similarity between human policy and the best partner policy for the agent that the human was playing with.



(a) Human baits

(b) Human shooters

Figure 3: Policy representations of each human baits (left) and shooters (right) in the static agent dataset (after PCA dimension reduction). Each colored node in the figures represents the average policy of a human player, while the size indicates her average team performance. Red nodes are reference points of exemplar policies.

This "similarity to best partner" quantifies the degree to which a human player is similar to the optimal policy for an agent teammate in our architecture. Correlation analysis shows that "similarity to best partner" is positively correlated with team performance in both bait ($r = 0.636, p = .0002$) and shooter ($r = 0.834, p < .0001$) groups. These results in which complementary pairings of the human shooter accounted for 70% of the variance among teams shows the high payoff potentially available from our approach to matching. The result indicates that the complementary policy pairs we found in agent-agent self-play can be successfully extended to human-agent teams, and that our proposed architecture is able to accurately identify human policy types and predict team performance.

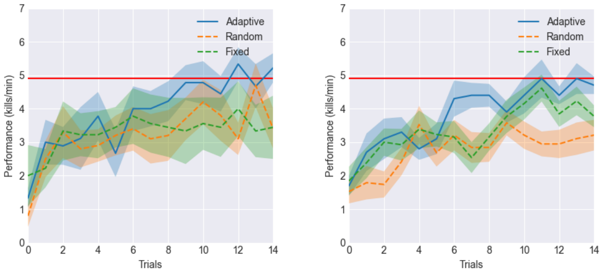
B. *Experiment 2: Adaptive agents*

1) Experimental Design: In the previous experiment and analysis, we validated our proposed architecture on a static agent dataset. In the second experiment, we evaluate the performance of an adaptive agent that dynamically chooses policies complementary to the human. A between-subject design was employed with one experimental group and two control groups.

In the experimental group, participants were paired with the adaptive agent in either shooter or bait role. The **Adaptive** group used the CEM similarity metric (Sec. V-C) to identify the policy most similar to the human behavior over a fixed number of recent preceding game frames. To perform the adaptation procedure; when each trial starts, the agent identifies the most similar agent policy to the policy of the human teammate in the last trial, and then refers to the self-play table to select the policy that would best complement the teammate's estimated policy for the current trial. In the two control groups, two different policy agents were chosen as baselines. The **Fixed** group selected a single policy at random from the library and maintained it through the whole experiment. This baseline controls for human learning from an agent that *does not adapt* allowing the human to adapt to a fixed agent policy. The **Random** condition randomly selects a static policy from the library *for each trial*. This baseline controls for the effects of human adaptation by presenting a new agent policy at each trial preventing human learning and adaptation. All three groups randomly select one static policy for the first trial as initialization.

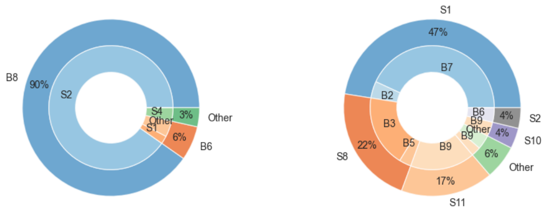
134 participants from Amazon MTurk were randomly divided into the three groups, and played 15 trials of the TSF game with corresponding agent partners. After deleting low quality data (e.g. unfinished trials, long idle times, low performance) we collected sessions from 42 valid human shooters and 29 valid human baits.

2) Results: The average performance of three groups of the human-agent teams are plotted in Fig. 4. We excluded the first 5 trials as a training phase and performed a statistical analyses on the data from last 10 trials. We ran a mixed two-way ANOVA for the number of fortress kills where trial number is the within-subject independent variable and agent type is the between-subject independent variable. This analysis illustrates how team performance changes over time in the three experimental conditions. In human bait - agent shooter teams, average team performance is not significantly different between conditions ($F(2, 25) = 0.67, p = .520$). This conclusion aligns with the pattern of Fig. 4a, in which three groups show strong learning effects with similar slope and finally reach a similar level of performance. More interesting patterns are found in human shooter - agent bait teams (Fig. 4b). There is a significant difference in team performance between three experimental conditions ($F(2, 39) = 3.45, p = .042$). Also, the main effect of trial numbers ($F(9, 351) = 3.53, p < .001$) and interaction effect between conditions and trial numbers ($F(18, 351) = 1.74, p = .031$) are both significant. To analyze the interaction effect, we ran post-hoc t-tests to compare the performance difference between groups in different time periods. Overall, the adaptive group outperforms the random group ($t(25.8) = 2.87, p = .008$) but not the fixed group ($t(20.8) = 1.61, p = .122$). However immediately following the training phase during trials 6-8, the average performance of the adaptive group is significantly higher than the other two groups ($p < .01$).



(a) Human bait - agent shooter (b) Human shooter - agent bait

Figure 4: (Best viewed in color) Average performance of human-agent teams. Solid blue line represents the learning curve of teams in the Adaptive condition, while the other two dashed lines represent Random and Fixed baseline, respectively. Shaded areas indicate one standard error from the mean. Red horizontal lines are the average performance of agent-agent teams in self-play as a reference.



(a) Human bait - agent shooter (b) Human shooter - agent bait

Figure 5: (Best viewed in color) Frequency distributions of adaptive agent's policy selection (inner circle) and human type identification results (outer circle) among all human-agent teams.

3) Discussion: We observe different patterns in human-agent teams when the adaptive agent takes different roles. Particularly for the Adaptive bait, the results align with our expectation that the Adaptive bait outperforms the Random baseline by 42.3% and achieves high team performance faster than the Fixed baseline (agent adapts more rapidly than human). We observe effects of both human adaptation and agent adaptation on team performance as shown in Fig. 4b. The superiority of the Fixed over the Random condition shows that improvement in the Fixed condition is due to human adaptation to the particular agent with which they were paired rather than increasing experience with TSF. For human shooters convergence to a stable team strategy with high performance is being achieved in both Adaptive and Fixed conditions. However, the Adaptive agent is able to *speed up* this process by selecting complementary policies. The unique mutual adaptation in the Adaptive group leads to the best team performance in the early stages (trial 6-8) compared to both baselines. This advantage of fast recovering after team reorganization (i.e. introducing unseen human teammates) brought by mutual adaptation has important implications for highly dynamic environments.

The Adaptive shooter, by contrast, does not show significantly better performance than either of the baselines. The performance gap between the two roles could be partially explained by the distribution of identified human policies and selected, complementary agent policies. As shown in Fig 5a, the adaptive shooter selects a single policy 90% of the time as most human baits are identified to be the same type while there are three to four major policies among human shooters in Fig 5b. Therefore the adaptive bait has a more diverse policy distribution and diverse policies are necessary to benefit from agent adaptation. An examination of table II shows that the dominant human bait policy, B8, is the next to poorest performing policy in the self-play table with 4:7 entries coming in under 4 while the table averages 4.9. Two factors might contribute to these difficulties: 1) The nature of the TSF game creates different policy/type spaces for shooter and bait roles. There might be too little variation in human bait policy for an agent to distinguish into types and select complementary policy accordingly. 2) Our independently constructed bait policy library is not a good representation of the true human type space and thus cannot provide accurate type identification for the agent to adapt to.

VII. Conclusion and Future Work

In this paper, we proposed a novel adaptive agent framework in human agent teaming (HAT) based on the cross-entropy similarity measure and a pre-trained static policy library. The framework was inspired by human teamwork research, which illustrates important characteristics of teamwork such as the existence of complementary policies, influence of adaptive actions on team performance, and the dynamic human policies in cooperation [23], [24]. The CEM measure making it possible to directly measure and compare behavior among humans and agents provides a tool of broad applicability.

We constructed a high-dimensional policy space based on types of policies in a pre-trained library and leveraged it as a reliable way to categorize and pair human policies with appropriate agent policies. The distance between human policy and the optimal complementary policy for his/her teammate is shown to be positively correlated with team performance, which confirms the validity of our proposed framework. An online adaptation method is employed to identify human policy shifting during the course of interaction and adapt the agent policy in real time. HAT evaluation shows that adaptive agents in the shooter role outperform agents using random adaptation strategies and achieve high team performance faster than a non-adaptive strategy. This result confirms the effectiveness of our proposed adaptive agent architecture. Further, the advantage of mutual adaptation in both overall team performance and faster team state convergence may be useful in highly dynamic environments.

This research has limitations that could be improved by future work. First, the effectiveness of the proposed adaptive agent depends on the representativeness of the policy library. A larger or more precise coverage in the policy space of the team task could lead to more accurate estimation of human policy and better selection of complementary policies. In the present study agents were trained in plausible ways we thought likely to encompass actual human policies. In future work, we would like to enrich the static agent library using methods [33] designed to generate a diversity of policies providing assurance of coverage.

Our method is dependent on clearly defined roles to exhaustively compare policy combinations to optimize the library. If boundaries between roles are porous (tasks can be performed by either agent) this process becomes much more difficult. In addition, required comparisons increase exponentially in the number of roles which along with diversity linked increases in number of policies could make optimizing computations expensive for larger problems. At execution, in compensation, comparisons are linear in the number of policies matching an actor's role.

While the adaptive bait agent outperformed randomly assigned static agents in trials 6-8 of the second experiment, human adaptation closed the gap over trials 9-14, suggesting people can learn to compensate for even a poorly matched partner. Finding effects of human adaptation to be on a par with agent adaptation suggests an additional role our agent might play. In keeping with the duty of the Ad Hoc teammate to guide the team to an optimal trajectory, the agent upon detecting an approaching asymptote, could nudge the human toward a higher performance policy pairing. The self-play table could perform a similar function in cases such as the human bait-agent shooter teams where agent policies could support much stronger performance than pairing with the human preferred policy allows. Provided humans could learn to approximate agent performance, nudging could again be used to guide the team toward more advantageous pairings.

Acknowledgment

This research has been supported by ARL award W911NF-19-2-0146, AFOSR/AFRL award FA9550-18-1-0251 and AFOSR award FA9550-18-1-0097.

References

[1] P. Stone and S. Kraus, "To teach or not to teach? decision making under uncertainty in ad hoc teams," in The Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS). International Foundation for Autonomous Agents and Multiagent Systems, May 2010.

[2] A. Decnstanza, A. Marathe, A. Bohannon, A. Evans, E. Palazzolo, J. Metcalfe, and K. McDowell, "Enhancing human-agent teaming with individualized, adaptive technologies: A discussion of critical scientific questions," IEEE Brain, 2018.

[3] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Duran, "Interactive team cognition," Cognitive science, vol. 37, no. 2, pp. 255–285, 2013.

[4] S. A. Manish Ravula and P. Stone, "Ad hoc teamwork with behavior switching agents," in International Joint Conference on Artificial Intelligence (IJCAI), August 2019.

[5] M. Tambe, "Electric elves: What went wrong and why," AI Magazine, vol. 29, no. 2, p. 23, Jul. 2008.

[6] C. Miller and R. Parasuraman, "Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control," Human factors, vol. 49, pp. 57–75, 03 2007.

[7] O. C. Görür, B. Rosman, and S. Albayrak, "Anticipatory bayesian policy selection for online adaptation of collaborative robots to unknown human types," in Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, 2019, pp. 77–85.

[8] A. Thomaz, G. Hoffman, and M. Cakmak, "Computational human-robot interaction," Foundations and Trends in Robotics, vol. 4, no. 2-3, pp. 105–223, 2016.

[9] S. Reddy, A. Dragan, and S. Levine, "Where do you think you're going?: Inferring beliefs about dynamics from behavior," in Advances in Neural Information Processing Systems, 2018, pp. 1454–1465.

[10] J. F. Fisac, E. Bronstein, E. Stefansson, D. Sadigh, S. S. Sastry, and A. D. Dragan, "Hierarchical game-theoretic planning for autonomous vehicles," in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 9590–9596.

[11] N. Agmon, S. Barrett, and P. Stone, "Modeling uncertainty in leading ad hoc teams," in Proc. of 13th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS), May 2014.

[12] R. Mirsky, W. Macke, A. Wang, H. Yedidsion, and P. Stone, "A penny for your thoughts: The value of communication in ad hoc teamwork," in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 254–260, main track.

[13] S. Singh, J. Lacotte, A. Majumdar, and M. Pavone, "Risk-sensitive inverse reinforcement learning via semi-and non-parametric methods," The International Journal of Robotics Research, vol. 37, no. 13-14, pp. 1713–1740, 2018.

[14] D. Hughes, A. Agarwal, Y. Guo, and K. Sycara, "Inferring non-stationary human preferences for human-agent teams," in 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE, pp. 1178–1185.

[15] S. Nikolaidis, D. Hsu, and S. Srinivasa, "Human-robot mutual adaptation in collaborative tasks: Models and experiments," The International Journal of Robotics Research, vol. 36, no. 5-7, pp. 618–634, 2017.

[16] R. Shah, N. Gundotra, P. Abbeel, and A. Dragan, "On the feasibility of learning, rather than assuming, human biases for reward inference," in International Conference on Machine Learning. PMLR, 2019, pp. 5670–5679.

[17] S. Barrett and P. Stone, "Cooperating with unknown teammates in complex domains: A robot soccer case study of ad hoc teamwork." in AAAI, vol. 15. Citeseer, 2015, pp. 2010–2016.

[18] S. Barrett, P. Stone, and S. Kraus, "Empirical evaluation of ad hoc teamwork in the pursuit domain." in AAMAS, 2011, pp. 567–574.

[19] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient model learning from joint-action demonstrations for human-robot collaborative tasks," Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, Mar 2015. [Online]. Available: http://dx.doi.org/10.1145/2696454.2696455

[20] S. V. Albrecht and S. Ramamoorthy, "A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems," 2013.

[21] S. V. Albrecht, J. W. Crandall, and S. Ramamoorthy, "Belief and truth in hypothesised behaviours," Artificial Intelligence, vol. 235, pp. 63–94, 2016.

[22] A. Agarwal, R. Hope, and K. Sycara, "Challenges of context and time in reinforcement learning: Introducing space fortress as a benchmark," arXiv preprint arXiv:1809.02206, 2018.

[23] H. Li, D. Hughes, M. Lewis, and K. Sycara, "Individual adaptation in teamwork," in Proceedings of the 42nd Annual Conference of the Cognitive Science Society, 2020, p. In Press.

[24] H. Li, T. Ni, S. Agrawal, D. Hughes, M. Lewis, and K. Sycara, "Team synchronization and individual contributions in coop-space fortress," in Proceedings of the 64th Human Factors and Ergonomics Society Annual Meeting, 2020, p. In Press.

[25] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient model learning from joint-action demonstrations for human-robot collaborative tasks," in 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2015, pp. 189–196.

[26] V. V. Unhelkar, S. Li, and J. A. Shah, "Semi-supervised learning of decision-making models for human-robot collaboration." CoRL, pp. 192–203, 2019.

[27] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in Advances in neural information processing systems, 2000, pp. 1008–1014.

[28] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in International conference on machine learning, 2015, pp. 1889–1897.

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

[30] S. Reddy, A. D. Dragan, and S. Levine, "Sqil: Imitation learning via reinforcement learning with sparse rewards," arXiv preprint arXiv:1905.11108, 2019.

[31] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," arXiv preprint arXiv:1509.06461, 2015.

[32] Y. Gao, H. Xu, J. Lin, F. Yu, S. Levine, and T. Darrell, "Reinforcement learning from imperfect demonstrations," arXiv preprint arXiv:1802.05313, 2018.

[33] J. Parker-Holder, A. Pacchiano, K. Choromanski, and S. Roberts, "Effective diversity in population based reinforcement learning," in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. International Joint Conferences on Artificial Intelligence Organization, 2020, p. 5923—5929.