In *Proceedings of the 19th European Conference on Multi-Agent Systems* (EUMAS 2022), Düsseldorf, Germany, September 2022

A Survey of Ad Hoc Teamwork Research

Reuth Mirsky[1,2], Ignacio Carlucho[3,⋆], Arrasy Rahman[3], Elliot Fosong[3], William Macke[2], Mohan Sridharan[4], Peter Stone[2,5], and Stefano V. Albrecht[3]

[1] Bar Ilan University, Israel mirskyr@cs.biu.ac.il
[2] The University of Texas at Austin, USA {wmacke, pstone}@cs.utexas.edu
[3] The University of Edinburgh, UK
{ignacio.carlucho, arrasy.rahman, e.fosong, s.albrecht}@ed.ac.uk
[4] The University of Birmingham, UK m.sridharan@bham.ac.uk
[5] Sony AI, USA

Abstract. Ad hoc teamwork is the research problem of designing agents that can collaborate with new teammates without prior coordination. This survey makes a two-fold contribution: First, it provides a structured description of the different facets of the ad hoc teamwork problem. Second, it discusses the progress that has been made in the field so far, and identifies the immediate and long-term open problems that need to be addressed in ad hoc teamwork.

Keywords: Ad Hoc Teamwork · Collaboration Without Prior Coordination · Agent Modelling · Reinforcement Learning · Zero-Shot Coordination

1 Introduction

Ad hoc teamwork (AHT) is defined as the problem of developing agents capable of cooperating on the fly with other agents without prior coordination methods, such a shared task and communication protocols or joint training. Designing an AHT agent is a complex problem, but the underlying capabilities are crucial to enabling agents to take on their designated roles in many practical domains. From service robots and care systems to team sports and surveillance, agents need to reason about the best way to collaborate with other agents and people without prior coordination. Research in AHT has been around for at least 15 years [Rovatsos and Wolf, 2002, Bowling and McCracken, 2005], and it was proposed as a formal challenge by Stone et al. [2010]:

> "To create an autonomous agent that is able to efficiently and robustly collaborate with previously unknown teammates on tasks to which they are all individually capable of contributing as team members."

Since then, hundreds of papers that include the phrase "ad hoc teamwork" have been published (464 according to Google Scholar at the time of writing this paper) and many more address closely related problems under names such as "zero-shot coordination" [Bullard et al., 2020, Hu et al., 2020]. Moreover, much of the work on personalizing agents' interactions with humans can be viewed as instances of AHT [Li et al., 2021].

⋆ Corresponding Author

This survey seeks to make a two-fold contribution. First, it defines the AHT problem by describing the underlying assumptions (Section 2.1), key subtasks (Section 2.2), and the scope of the problem as considered in this paper (Section 3). Second, it surveys the existing work in AHT in terms of the solution methods (Section 4) and the evaluation domains that have been developed (Section 5), and discusses the open problems in the field of AHT (Section 6).

Related initiatives. Several initiatives over the last decade have contributed to research progress in AHT. In particular, between 2014 and 2017, the Multi-Agent Interaction without Prior Coordination (MIPC) workshop series[6] held at AAAI and AAMAS conferences facilitated discussions and presentations in AHT and related topics. The MIPC workshop series was followed by a special journal issue [Albrecht et al., 2017] which featured a collection of new research works in AHT. Moreover, the RoboCup Drop-in Challenge was introduced to provide a platform to develop and evaluate AHT capabilities in the context of soccer-playing robots [Genter et al., 2017]. However, to date there is no comprehensive survey on AHT. We seek to address this gap in the literature and help foster further research in AHT.

2 Background

This section provides a basic formulation of the AHT problem. It takes the original challenge proposed in Stone et al. [2010] and describes it in terms of the inputs and outputs, and the underlying assumptions (Section 2.1). It then describes the subtasks of the problem based on issues addressed in relevant papers (Section 2.2).

2.1 Problem Formulation

The AHT problem focuses on training an agent to coordinate with an unfamiliar group of teammates without prior coordination. In this work, we refer to the trained agent as the **learner**. The learner's **teammates** are assumed to be capable of contributing to the common teamwork task, meaning that they have a set of skills that are useful for the task at hand. Here we describe the inputs, outputs, and the underlying assumptions of this problem.

Input. The inputs of the AHT problem are the teamwork task to be executed, domain knowledge comprising a description of the domain/environment in which the task is to be executed, a (possibly incomplete) list of attributes characterizing each agent (e.g., a set of goals, perception, and action capabilities), a description of the learner's abilities, and a list of teammates. The agent attributes' values might differ between each teammate—also see first assumption below—and some teammates might be able to communicate with each other.

[6] https://mipc.inf.ed.ac.uk

Output. The output of the problem is the learner, represented by a policy that determines the action this agent should execute in any given state of the domain. Depending on the agent's sensors, actuators, and the available communication channels, this policy can be deterministic or stochastic, static or adaptable, and might include ontic (physical) actions and epistemic (knowledge-producing) actions, which in turn may contain verbal or non-verbal communication.

Assumptions. Three key assumptions (i.e., claims or postulates) characterize the AHT problem.

1. **No prior coordination.** The learner is expected to cooperate with its teammates when the task begins without any prior opportunities to establish or specify mechanisms for coordination. For example, it is not possible to prespecify the agents' roles or to have a joint training phase for all agents. The learner might know or assume knowledge of a subset of attributes (e.g., current policies, individual goals) of some subset of its teammates. This knowledge might be acquired from an expert who has had prior interactions with the learner's current teammates, and the assumptions might be the result of generic models or rules based on past interactions in the target domain. The learner's current teammates might or might not be familiar with one another before the current interaction. For example, in drop-in soccer (a spontaneous soccer match where some or all of the team are strangers), a teammate might be perceived to be a good striker because they are fast and the team can work around this assumption even if they have not played with that specific player before.

2. **No control over teammates.** The learner cannot change the properties of the environment, and the teammates' policies and communication protocols; it has to reason and act under the given conditions. We distinguish between *changing the properties* of the environment (e.g. modifying observability level) and *acting in* the environment to change its state (e.g. picking up a box). Similarly, the learner might influence its teammates' actions, but this influence will be in accordance with the pre-defined policy of the teammates. Moreover, teammates' policies may support learning or adaptation, but the learner cannot modify these abilities. Continuing with the soccer example, teammates can learn to work better together with practice, but no teammate can impose their knowledge on the team before the game starts.

3. **Collaborative.** All agents are assumed to have a common objective, but some teammates might have additional, individual objectives, or even completely different rewards. However, these additional objectives do not conflict with the common task [Grosz and Kraus, 1999]. In the drop-in soccer example, different teammates may have incentives in their contract that encourage them to focus on different skills, e.g., goal-scoring rewards for forwards or assist rewards for midfielders. The difference in the individual objectives may result in situations in which an individual agent may seem to be acting contrary to the team reward, but each agent in the team is always acting to achieve the common objective. For example, although passing frequently is considered very important to a team's performance in a soccer game, an individual teammate may choose to dribble forward because of a perceived opportunity to score a goal.

2.2 Subtasks in Ad Hoc Teamwork

Based on a survey of the existing literature, we identified four main subtasks that the learner should be able to perform, although much of the existing work only focuses on addressing a subset of these subtasks.

ST1: Knowledge representation. The learner requires a representation of the domain knowledge. This includes knowledge about the environment (e.g., discrete or continuous, static or dynamic, etc.), its capabilities, and knowledge about potential teammates (e.g., similarity to past teammates, their theory of mind, etc.). These choices influence the solution methods for the other substasks. Most of the attributes characterizing the environment are common to all multi-agent problems. They can be presented in the classical PEAS system [Russell and Norvig, 2021] and are not unique to AHT, so we do not elaborate on these here.

ST2: Modeling teammates. The learner can leverage information about its teammates to improve its decision making. Thus, a key subtask for the learner is to model the information pertaining to teammates' behavior (e.g., classifying teammates by type in order to adapt to different teammates).

ST3: Action selection. The third subtask is the design of mechanisms used by the learner to select actions once it has an estimate of its teammates' behavior (observed or based on models of teammates). Example methods for this subtask include planning methods and expert policies that are learned or based on expert knowledge.

ST4: Adapting to changes. During interaction, the learner might receive new information about its teammates, the environment, or task objectives. Based on this information, the learner needs to adapt its behavior to improve coordination. This adaptation also includes merging the models provided by teammates.

3 Boundaries of Ad Hoc Teamwork

Here we further define the scope of the AHT problem by describing factors that can be considered within the basic problem formulation presented above, and by discussing related research problems.

3.1 Variations of the Ad Hoc Teamwork Problem

We first describe additional factors that define the scope of AHT and influence the subtasks described earlier.

Partial observability. Under conditions of full observability, each agent is aware of the state of the environment, including the location of other agents. Partial observability implies a higher level of complexity in knowledge representation as it introduces uncertainty in certain parts of the domain state. Changing the observability level will affect ST1 and thus the other subtasks described above.

Open environment. Closed environments assume a fixed number of teammates [Rahman et al., 2021]. Relaxing this assumption increases the problem complexity, as the learner will also have to adapt to the changing number of teammates in the environment; this will primarily affect ST2 and ST4.

Communication. Since the exploration of how communication can be leveraged to improve team performance is an important area of research in AHT, we make a distinction based on whether there is any communication channel between agents. When communication exists, it is sometimes presented as predetermined and known protocols, such as the hints allowed in the game of Hanabi [Bard et al., 2020], which affects ST1. If these protocols are unknown in the beginning of the interaction and need to be learned during the task execution, it has an effect on ST3 and ST4.

Adaptive teammates. We make a distinction between work where the teammates learn alongside the learner, or use policies that stay fixed throughout the learning phase of the learner. Unlike multi-agent reinforcement learning (see Section 3.2), which supports joint training for all agents in the team, AHT does not assume that the deployed teammates are the same as those the learner might have trained with. Rather, adaptive teammates learn by reacting to the learner's policy using methods that are not known to the learner, thus affecting ST3 and ST4. An example of such a setup is flocking, where the teammates have a fixed policy, but their actions are directly influenced by the learner [Genter and Stone, 2016].

Mixed objectives. While teammates are assumed to be collaborative, they can have mixed objectives. Two types of scenarios arise depending on the objectives of the learner and its teammates. In the first, the learner and the teammates have a perfectly aligned objective (e.g., the reward functions of all agents are identical). In the second, while all team members have a common goal, each agent might also hold individual goals as long as these are not purely adversarial to the shared one. This factor extends the original formulation in [Stone et al., 2010], is related to the third assumption in Section 2.1, and will primarily affect ST2 and ST3.

3.2 Related Problems

In this section, we highlight the main differences between AHT and other related research problems.

Multi-agent reinforcement learning (MARL). It refers to the use of reinforcement learning methods for jointly training multiple agents to maximize their respective cumulative rewards while working with each other [Busoniu et al., 2008, Devlin and Kudenko, 2016, Papoudakis et al., 2019]. AHT, on the other hand, assumes control over a single agent (the learner) while teammates can have their own learning mechanisms, e.g., a robot interacting with different human. Prior work has shown that the good team performance of MARL methods often comes at the expense of poor performance when interacting with previously unseen teammates [Vezhnevets et al., 2020, Rahman et al., 2021, Hu et al., 2020]. MARL methods are thus not particularly well-suited to AHT.

Ad hoc teaming. The objective is to learn coercive measures that may allow self-interested agents with different skills and preferences to collaborate and solve a task. For example, existing work has trained a manager to assign subtasks to agents based on their skills while also incentivizing agents to complete their tasks [Shu and Tian, 2019]. In contrast, the learner in AHT might incentivize its teammates to act in a certain way, but cannot dictate the teammates' behavior due to the lack of prior coordination.

Agent modelling. These methods infer attributes of teammates' behavior such as beliefs, goals, and actions [Albrecht and Stone, 2018]. Since inferring teammates' behavior is important for decision making in AHT (e.g., ST3 in Section 2.2), agent modeling methods are useful for AHT. However, they can be used for a broader class of problems and are not limited to (or necessarily indicative of) AHT.

Human-agent interaction. The task of creating agents that interact with previously unseen agents has also been explored in the human-agent/robot interaction community. In human-agent interaction, agents have to achieve their goals in the presence of human decision makers. As in AHT, it is often impossible to jointly train humans and agents to coordinate their behavior; agents must instead find a way to coordinate with previously unseen humans, e.g., by using implicit communication or acting in a legible manner [Breazeal et al., 2005, Dragan et al., 2013].

Zero-shot coordination (ZSC). A special case of AHT where teammates' behavior are assumed to arise from a reward function that always provides identical rewards for every agent is known as ZSC [Lupu et al., 2021, Hu et al., 2021, Bullard et al., 2020, 2021]. After training different populations of agents under the same fully cooperative setup, a ZSC agent is evaluated by measuring its performance when cooperating with agents from a different population. While ZSC introduced techniques relevant for AHT, there are AHT problems where the controlled agent must interact with teammates whose reward functions are different from its own.

4 Solution Approaches

As stated earlier, while existing methods for AHT often provide a functioning learner, each method's key contribution can often be mapped to one or more of the four subtasks in Section 2.2. Here we elaborate on common solution methods for each subtask and refer to representative literature.

4.1 Knowledge Representation

The representation of domain knowledge strongly influences the solution approach used in the other subtasks. This information can be acquired from human experts (or expert knowledge), prior knowledge of past teammates, or using self-play.

To support adaptation based on limited information, it is common to equip agents with preconceptions of the likely behaviors or intentions of previously unseen teammates. These preconceptions are based on prior experience with the task; this can be the agent's own experience or that of a human familiar with the task. *Agent modeling* techniques can be used to represent the teammates [Albrecht and Stone, 2018].

Type-based methods. The use of type-based methods is common in the AHT literature. These methods represent prior experience with agents (in the target domain) by a set of hypothesized *types*, where each type models an action selection policy. It is assumed that new teammates encountered by the learner have behaviors specified by one of these types.

A range of type representations have been explored. Early work explored a nested representation of agents' beliefs, where agents perform Bayesian updates to maintain beliefs over physical states of the environment and over models of other agents [Gmytrasiewicz and Doshi, 2005]. It was also common to use hand-coded programs to represent types [Barrett et al., 2011, Albrecht and Ramamoorthy, 2013]. For approaches that employ a learned type set, learned decision trees were a common representation [Barrett et al., 2017]. More recently, latent type methods have been used which learn a neural network-based encoder to map observations of teammates to an embedding of the agent's type [Rabinowitz et al., 2018, Xie et al., 2020, Rahman et al., 2021, Zintgraf et al., 2021].

There are three main approaches to specifying a hypothesized type space: (1) specification by a human expert; (2) learning from data; and (3) using reinforcement learning (RL) methods and access to the environment or an environment model. Barrett et al. [2017] collect diverse behaviors by drawing their types from the output of an assignment presented to a large number of student. Many methods attempt to generate diverse behaviors in a population trained via RL, requiring only access to the target task. They do so using methods such as genetic algorithms [Albrecht et al., 2015a,b, Canaan et al., 2020], regularisation techniques [Lupu et al., 2021], and reward-shaping techniques [Leibo et al., 2021].

Experience replay. Rather than encoding experience in explicit behavioural models, experience replay methods store transition data in a buffer. Transitions observed during an interaction are compared against the stored transitions to identify the current teammate [Chen et al., 2020].

Task recognition. In methods based on task recognition, prior experience or information provided by an expert is encoded as a library of tasks referred to as *plays*, *macro actions*, or *options* [Sutton et al., 1999]. Tasks then encode prior experience as applicability conditions, termination conditions, and high-level specifications of a sequence of low-level actions [Wang et al., 2021].

4.2 Identifying Current Teammates

Once a representation is set, estimating the behavior of current teammates allows the learner to determine a suitable behavior.

Type inference. Methods that represent teammates using types infer beliefs over the hypothesized type space using a history of interactions of the learner with each teammate up to the current timestep. The dominant approach is to use a Bayesian belief update [Albrecht et al., 2016, Barrett et al., 2017]. In such methods, prior beliefs about the teammates' types are updated using the history of interactions and a likelihood of the types based on the history. It is also common to assume uniform priors across types and type parameters [Albrecht et al., 2015a].

Experience recognition. Rather than inferring types, some approaches attempt to measure the similarity of the current observations to that from earlier experience in a more direct manner. PLASTIC-Policy [Barrett et al., 2017] compares the most recently observed state transition to previously stored data. For each team they find the stored transition with the closest state to the current state, and consider the next state observed in that historical transition. They then measure the distance between that state and the observed next state, and use this to compute the likelihood of the team. AATEAM [Chen et al., 2020] takes a more sophisticated approach which uses prior experience buffers to train one attention-based neural network per type, to identify agents from a trajectory rather than a single transition.

Task recognition. For methods which represent prior knowledge as tasks, the learner attempts to infer the current task being carried out by the teammate under consideration. Wang et al. [2021] achieved this by assuming that the teammate was attempting to complete hypothesized tasks and computing the extent to which the teammate's observed behavior is sub-optimal for that task. Melo and Sardinha [2016] consider a setting in which agents both identify the current task and identify the teammate's strategy, with the teammate's behavior subject to a bounded rationality assumption.

4.3 Action Selection

Given current knowledge about task and teammates, agents must decide which action to take to maximize team return.

Planning. Many AHT approaches use planning methods to select actions. Some, such as Bowling and McCracken [2005] and Ravula et al. [2019], use bespoke planning methods suited to the specific task, and chosen by a human expert. Many approaches use the more general Monte Carlo tree search (MCTS) planning procedure [Wu et al., 2011, Barrett et al., 2014, Alford et al., 2015, Sarratt, 2015, Albrecht and Stone, 2017, Malik et al., 2018, Yourdshahi et al., 2018, Eck et al., 2020]. The upper confidence tree (UCT) algorithm [Kocsis and Szepesvári, 2006] for MCTS is often used due to its ability to perform well when the branching factor is large, as is the case when multiple agents are present. These MCTS-based methods require that types are represented by explicit behavioral models to sample teammate actions during rollouts.

Expert policy methods. Selecting actions by choosing a policy from a set of expert policies, and then acting according to the chosen policy. There are many ways in which these expert policies can be obtained prior to the ad hoc interaction: they can be provided by an expert, learned offline, using experience data [Chen et al., 2020, Santos et al., 2021], or by online RL training given the task [Albrecht et al., 2015b]. One of the advantages of expert policy methods over type-based planning methods is that they can handle large or continuous state and action spaces, where MCTS approaches may struggle [Barrett et al., 2017]. However, type-based planning methods are more appropriate when the ad hoc team is likely to have a previously unseen composition, as type-based methods can reason at the level of the types of individual agents. Also, creating expert policies may be impossible when a large variation of situations are encountered. The

E-HBA method attempts to achieve the advantages of both type-based reasoning methods and expert policy methods by combining the two [Albrecht et al., 2015b]. The GPL method [Rahman et al., 2021], suitable in open AHT problems, uses an action-selection mechanism based on E-HBA .

Leading. Some works explicitly consider adaptive teammates, where a learner's choice of action affects its teammates' behaviors. Works such as Agmon et al. [2014] assume teammates employ a known best response strategy, and that the goal is to lead these teammates to a specific joint coordination strategy. These approaches were addressed in simple games using dynamic programming. Xie et al. [2020] consider cases where the learner does not know the teammate's current behavior, nor how this behavior changes across interactions. Thus, deep learning is used to learn an embedding of the teammate's strategy, and model the teammate's behavioral dynamics and teammates' adaptation process.

Metalearning. Metalearning approaches use action selection policies which are trained to facilitate the entire AHT process. The MeLIBA approach [Zintgraf et al., 2021] trains the policy to carry out interactive Bayesian RL, intentionally taking actions which seek to reveal information about the teammate's type. The action selection policies of metalearning approaches is typically conditioned on the learner's prediction of the teammate's type. In this sense, such methods can be compared to expert policy methods.

4.4 Adapting to Current Teammates

During interaction, the learner receives new information, which can be used to adapt its behavior.

Belief revision. Most methods employ belief revision protocols to maintain their belief about the identity of other agents across time. For type-based methods, it is typical to assume each teammate's type does not change over time, and that a good representation of the teammate exists in the hypothesized type space [Albrecht et al., 2016]. However, if it is assumed that teammates' types change over time, the learner must also adapt. The ConvCPD method [Ravula et al., 2019] considers settings in which the type space is known, but agents can switch types. For these settings, they employ a convolutional neural network (CNN)-based changepoint detection approach, which uses image-like representations of type likelihoods across time to detect changes. An alternative approach is to modify the Bayesian belief revision process to allow beliefs to decay towards the priors over time. This approach is useful when a teammate changes to a type which the learner has assigned low (or zero) probability to. In this case, the learner might struggle (or be unable) to quickly update its belief to reflect the new true teammate [Santos et al., 2021]. Sum-based posterior definitions were also proposed to deal with changing types [Albrecht et al., 2016].

Hypothesis space revision. Approaches exist for adapting to agents whose behavior may not be adequately represented in the hypothesized space. TwoStageTransfer is a transfer learning method employed by PLASTIC-Model [Barrett et al., 2017] which

uses observations of new teammates and prior models to finetune a model for the new teammate.

Metalearning. During the metalearning process, the action selection policy learns its own adaptation procedures, avoiding the need to specify particular adaptation schemes [Xie et al., 2020, Zintgraf et al., 2021].

Zero-Shot coordination techniques. The ZSC problem does not allow the learner any behavioral adaptation during ad hoc interactions. For this reason, the focus of these methods is on training agents which robustly coordinate with other agents trained using the same algorithm. One approach is to avoid strategies which are not invariant under symmetries within the underlying tasks [Hu et al., 2020, 2021]. Another approach is based on the hypothesis that there are few strategies which perform well with a diverse set of teammates, so ad hoc agents independently trained against diverse teammates (and themselves) are likely arrive at similar pre-coordinated policies [Lupu et al., 2021].

Communication. The learner can quickly adapt to changes is by communicating with its teammates. This communication can either be a query [Mirsky et al., 2020, Macke et al., 2021], transfer knowledge or preferences [Mead and Weinberg, 2007, Barrett et al., 2014], or providing an advice [Shvo and McIlraith, 2020, Canaan et al., 2020].

5 Evaluation Domains

Many different approaches have been used for evaluating AHT methods. In this section, we categorize them using the identified variations from Subsection 3.1. Some domains might fit more than one category, but we place them according to the first ad hoc teamwork paper they appeared in. In Table 1, we summarize each of the domains and associated papers.

No variations. Some evaluation domains do not have any of the variations outlined in Section 3.1. Among these AHT domains, some of the simplest are matrix games [Albrecht et al., 2015b, Melo and Sardinha, 2016]. These games consist of a payoff matrix for two agents who independently choose actions and then receive a payoff based on the actions each agent chose. The game is then repeated with the goal to maximize long term return over repeated trials. Another common domain is predator prey [Barrett et al., 2011, Ravula et al., 2019, Papoudakis et al., 2021]. This domain consists of several agents (the predators) attempting to surround and capture other agents (the prey). The predator prey domain requires both recognising a teammate's goal (namely which prey they are pursuing), and also collaborating with other agents to surround the prey. In level-based foraging [Albrecht and Ramamoorthy, 2013], the goal of the agent team is to collect food items which are spatially distributed in a grid world. Agents and items have different skill levels which represent different capabilities in agents, requiring that agents decide when and with whom to collaborate in order to collect the items.

Open environments. There are several instances of open domains presented in AHT. First, open variations of the domains mentioned above exist in Rahman et al. [2021]. Another open AHT domain is wildfires, where agents entering and leaving the environment need to work together to contain the spread of wildfires [Chandrasekaran et al., 2016]. Finally, ad hoc flocking and swarming domains enable agents to enter and leave the environment freely [Genter and Stone, 2016].

Partial or noisy observability. Partially observable variants of the domains with no extensions exist in Ribeiro et al. [2022]. One domain that has been prevalent in AHT literature is robot soccer. Drop-in soccer where a group of players need to form a team without playing with each other is common among humans in real life, so it has been a frequented challenge by AI as well [Barrett et al., 2017, Genter et al., 2017]. The problem typically consists of substituting one member of a team with a learner. The performance is then measured on how robust the learner's performance is regardless of which team it is placed in. This domain presents an additional challenge, as each agent can only observe its local environment. Another partially observable domains are military simulation, which simulate various combat and search tasks using unmmaned autonomous vehicles [Alford et al., 2015], and the collaborative card game Hanabi [Bard et al., 2020]. Similar to the RoboCup domain, these domains also present the challenge that agents only have access to their local observations.

Communication. Multiple domains allow communication in some form. The RoboCup domain mentioned above allows limited communication between agents using wireless connections. Others use communication as a more critical part of the domain. The tool fetching domain provides an AHT domain that allows one agent to query another about its goals [Macke et al., 2021]. Unlike other domains mentioned so far, the tool fetching domain is specifically focused on evaluating an agent's ability to communicate effectively. The Hanabi domain also presents a structured communication channel. While in the tool fetching domain the learner can query its teammates, in Hanabi the communication channel allows the learner to provide its teammates with information unknown to them [Bard et al., 2020, Canaan et al., 2020]. Another domain that focuses on communication is the cops and robbers domain [Sarratt, 2015]. In this domain, teammates (cops) must work together to capture another, adversarial agent (the robber). Each agent can query the other to gain information about their current plans [Sarratt, 2015].

Adaptive teammates. So far all domains mentioned are focused on evaluating whether a learner can successfully adapt their behavior to collaborate with diverse teammates. Some domains, however, instead try to evaluate how well learner(s) can influence other agents to achieve better performance. While the above domains can be adapted to have learning teammates, several domains exist with this explicit purpose in mind. Some examples of these are domains focused on incentivising the teammate to take a specific course of action [Wang et al., 2021], or on swarming [Genter and Stone, 2014], where the learner attempts to move in such a way as to influence the overall behavior of the agents around it.

Mixed objectives. Works that make the assumption of coupled objectives, such as ZSC [Hu et al., 2020], utilize an environment in which the reward received by all agents

Table 1: Different environments used for evaluating ad hoc teamwork.

Domain	Paper	Method Description
Matrix Games	Albrecht et al. [2012]	Empirically evaluates various multi-agent learning algorithms in ad hoc mixed teams.
	Chakraborty et al. [2013]	Introduces an optimal algorithm to cooperate with a Markovian teammate.
	Albrecht et al. [2015b]	Combines type-based reasoning for prediction with expert algorithms for decision making.
	Albrecht et al. [2016, 2015a]	Evaluates impact of prior beliefs in type-based reasoning in a range of matrix games.
	Melo and Sardinha [2016]	Extends ad hoc teamwork to scenarios where the current task is unknown in addition to the teammates.
Predator Prey	Barrett et al. [2011]	MCTS (UCT) with type-based reasoning using handcrafted types in the predator prey domain.
	Ravula et al. [2019]	Extends ad hoc teamwork methods to work with teammates which can switch behaviors.
	Papoudakis et al. [2021]	Assumes only local observations of ad hoc teamwork agent are available to model other agents.
LBF	Albrecht et al. [2013]	Develops type-based reasoning based on game theory model to solve ad hoc teamwork problems.
	Albrecht and Stone [2017]	Type-based reasoning with continuous parameterized types and MCTS (UCT).
	Liemhetcharat et al. [2017]	Defines the problem of ad hoc team assignment.
	Yourdshahi et al. [2018]	Introduces new history-based MCTS.
	Rahman et al. [2021]	Uses graph-based learning to handle a dynamic number of agents in the environment.
Wildfires	Eck et al. [2020]	Introduces ad hoc teamwork in open environments with large numbers of agents.
Flocking Swarming	Genter and Stone [2014]	Introduces AHT approaches for influencing a flock's behavior.
	Genter et al. [2015]	Determines where to place agents in a flock.
	Genter and Stone [2016]	Solves how to force agents to join flock in motion.
Robot Soccer	Bowling et al. [2005]	Introduces two new approaches for working with ad hoc teams in robot soccer.
	Barrett and Stone [2014]	Introduces new method for reusing policies learned from previous teammates to accomplish AHT.
	Barrett et al. [2017]	Introduces algorithms for AHT based on previously met teammates, using either policies or models.
Military Simulation	Alford et al. [2015]	Introduces an algorithm for classifying agent behaviors in air combat simulator.
Hanabi	Bard et al. [2020]	Proposes the Hanabi game as a new challenge for AI research, including ad hoc teamwork.
	Canaan et al. [2020]	Creates a meta-strategy for solving ad hoc teamwork in Hanabi using a diverse set of possible teammates.
	Hu et al. [2020]	An effective algorithm for learning from self-play by attempting to seek out new behaviors.
	Hu et al. [2021]	Introduces improved method off-belief learning for learning from self-play in DecPOMDPs.
	Lupu et al. [2021]	Creates a new optimisable metric for determining policy diversity in Hanabi self-play.
Tool Fetching Domain	Mirsky et al. [2020]	Introduces SOMALI CAT problem and proposes solution for determining when queries might be useful.
	Macke et al. [2021]	Proposes a solution for what to query when multiple possible queries are available.
	Suriadinata et al. [2021]	Investigates human behavior in the Tool Fetch Domain.

is the same. Such environments include the lever environment [Hu et al., 2020] and Hanabi [Bard et al., 2020]. Works which do not assume coupled objectives utilize general-sum domains such as level-based foraging [Albrecht and Ramamoorthy, 2013], in which the reward changes depending of the contribution of the agent; or the tool fetching domain where each agent has a distinct role in the team [Mirsky et al., 2020].

6 Conclusion and Open Problems

In this survey, we presented a review of the AHT literature that has been published over the past decade. This long period of time, along with the abundance of published work, enabled us to draw a big picture view of this topic: setting the boundary on what is, and what is not, AHT; identifying the subtasks that an agent needs to tackle as part of an AHT task; and the various levels of complexity in AHT. Many open problems still need to be addressed to achieve a robust agent that is able to interact with teammates without prior coordination and solve real-world problems. Furthermore, AHT research is currently suffering from a lack of standardised comparison between existing AHT approaches, which increases the difficulty of identifying state-of-the-art methods for solving a certain AHT problem.

Future work could address further extensions of the variations of the ad hoc teamwork problem discussed in Section 3.1, or combinations of these variations. For example, considering the presence of teammates with complex adaptive processes, such as teammates which learn via RL while interacting with the learner; or teammates which themselves apply AHT techniques. Current approaches to AHT are not designed to work with adaptive teammates (one notable exception being HBA [Albrecht et al., 2016]), whose presence would mean that the learner needs not only to adapt to teammates' behaviors, but also consider how the teammates adapt to its own behavior. Another extension is the combination of partial observability and open teams, which provides a difficult challenge for the learner, due to this complex dual uncertainty.

In terms of potential solution methods, one of the crucial open problems is improving the generalization to new teammates that have not yet been seen during training. Recent continual learning [Khetarpal et al., 2020] advances showed that training on diverse tasks can result in agents with robust performance in previously unseen tasks [Open-Ended Learning Team et al., 2021]. In the same way, training with a diverse set of teammates can improve the learner's ability to collaborate with new teammates. Lupu et al. [2021] proposed a method to generate diverse teammates for ZSC, but it was not evaluated with collaborative teammates with objectives that might not be fully aligned with the learner's. Recently, Rahman et al. [2022] proposed a method for generating a diverse set teammates specifically for ad hoc teamwork applications. However, results were only obtained in a $5x5$ grid world environment, more work is needed to evaluate how this method performs in more complex environments. These works are a good starting point when designing learners that are robust to different teams, however, they do not specifically address the collaborative aspect of AHT. Additional work is required to properly define the scope of the diverse set of agents a learner should be able to work with. And while generating teammates that display different behaviours and skill levels

can improve generalisation during execution time, this is not an easy task, especially in more complex domains.

AHT research could also benefit from the use of more complex or realistic domains in evaluation. Previous works tended to use simple domains (Section 5), but these solutions might not perform well in realistic domains. We suggest that future AHT research should consider more realistic testbeds, which can rely on robotics simulators extended to handle multi-agent scenarios [Collins et al., 2021], or on existing scenarios such as the DARPA "Spectrum Collaboration Challenge"[7], which will allow for the evaluation of more complex tasks and algorithms. Social navigation, the problem of a robot navigating through a crowd of people and robots, is another relevant robotics challenge [Mirsky et al., 2021]. In this problem, the learner needs to coordinate with previously unmet passerby humans and robots in order to avoid collisions, while allowing each other to get to their destinations. Thus, this challenge poses a series of challenging AHT problems where the learner need to adapt to new incoming teammates based on a highly limited amount of interaction experience.

Another important issue that can be addressed by future work is benchmarking current AHT approaches by providing systematic comparison between them. Existing works in AHT often forgo comparison against other approaches designed to solve the same variation of AHT problems, which makes it hard to identify state-of-the-art approaches in the field. A systematic benchmark between AHT approaches across different environments could therefore be a crucial stepping stone towards further identifying the strengths and weaknesses of different AHT methods.

To conclude, the AHT problem comprises a unique mixture of subtasks that the learner is required to perform, which requires solutions ranging from different fields. In this survey, we identified the existing and open problems in AHT which we hope will contribute to the development of the field, and in turn will advance the multi-agent research community as a whole.

[7] www.darpa.mil/program/spectrum-collaboration-challenge

Bibliography

N. Agmon, S. Barrett, and P. Stone. Modeling uncertainty in leading ad hoc teams. In *AAMAS '14*, pages 397–404, 2014.

S. V. Albrecht and S. Ramamoorthy. Comparative evaluation of MAL algorithms in a diverse set of ad hoc team problems. In *AAMAS '12*, 2012.

S. V. Albrecht and S. Ramamoorthy. A game-theoretic model and best-response learning method for ad hoc coordination in multiagent systems. In *Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '13, page 1155–1156, Richland, SC, 2013. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450319935.

S. V. Albrecht and P. Stone. Reasoning about hypothetical agent behaviours and their parameters. In *AAMAS '17*, pages 547–555, 2017.

S. V. Albrecht and P. Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

S. V. Albrecht, J. W. Crandall, and S. Ramamoorthy. An empirical study on the practical impact of prior beliefs over policy types. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 1988–1994, 2015a.

S. V. Albrecht, J. W. Crandall, and S. Ramamoorthy. E-HBA: Using action policies for expert advice and agent typification. In *AAAI Workshop on Multiagent Interaction without Prior Coordination*, page 7, 2015b.

S. V. Albrecht, J. W. Crandall, and S. Ramamoorthy. Belief and truth in hypothesised behaviours. *Artificial Intelligence*, 235:63–94, 2016.

S. V. Albrecht, S. Liemhetcharat, and P. Stone. Special issue on multiagent interaction without prior coordination: Guest editorial. *Autonomous Agents and Multi-Agent Systems*, 31(4):765–766, 2017. https://doi.org/10.1007/s10458-016-9358-0.

R. Alford, H. Borck, and J. Karneeb. Active behavior recognition in beyond visual range air combat. In *Proceedings of the 3rd Annual Conference on Advances in Cognitive Systems*. Cognitive Systems Foundation, 2015.

N. Bard, J. N. Foerster, S. Chandar, N. Burch, M. Lanctot, H. F. Song, E. Parisotto, V. Dumoulin, S. Moitra, E. Hughes, I. Dunning, S. Mourad, H. Larochelle, M. G. Bellemare, and M. Bowling. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, 280:103216, 2020.

S. Barrett and P. Stone. Cooperating with unknown teammates in robot soccer. In *AAAI Workshop on Multiagent Interaction without Prior Coordination*, page 6, 2014.

S. Barrett, P. Stone, and S. Kraus. Empirical evaluation of ad hoc teamwork in the pursuit domain. In *AAMAS '11*, volume 2, pages 567–574, 2011.

S. Barrett, N. Agmon, N. Hazon, S. Kraus, and P. Stone. Communicating with unknown teammates. In *ECAI 2014*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 45–50. IOS Press, 2014. https://doi.org/10.3233/978-1-61499-419-0-45.

S. Barrett, A. Rosenfeld, S. Kraus, and P. Stone. Making friends on the fly: Cooperating with new teammates. *Artificial Intelligence*, 242:132–171, 2017. https://doi.org/10.1016/j.artint.2016.10.005.

M. Bowling and P. McCracken. Coordination and adaptation in impromptu teams. In *National Conference on Artificial Intelligence*, volume 1 of *AAAI '05*, pages 53–58, 2005.

C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *IEEE/RSJ international conference on intelligent robots and systems*, pages 708–713. IEEE, 2005.

K. Bullard, F. Meier, D. Kiela, J. Pineau, and J. Foerster. Exploring zero-shot emergent communication in embodied multi-agent populations. *arXiv:2010.15896*, 2020.

K. Bullard, D. Kiela, F. Meier, J. Pineau, and J. Foerster. Quasi-equivalence discovery for zero-shot emergent communication. *arXiv:2103.08067*, 2021.

L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008. https://doi.org/10.1109/TSMCC.2007.913919.

R. Canaan, X. Gao, J. Togelius, A. Nealen, and S. Menzel. Generating and adapting to diverse ad-hoc cooperation agents in Hanabi. *arXiv:2004.13710*, 2020.

D. Chakraborty and P. Stone. Cooperating with a markovian ad hoc teammate. In *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems*, volume 1 of *AAMAS '13*, pages 1085–1092. International Foundation for Autonomous Agents and Multiagent Systems, 2013.

M. Chandrasekaran, A. Eck, P. Doshi, and L. Soh. Individual planning in open and typed agent systems. In *Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 82–91, 2016.

S. Chen, E. Andrejczuk, Z. Cao, and J. Zhang. AATEAM: Achieving the ad hoc teamwork by employing the attention mechanism. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 7095–7102, 2020. https://doi.org/10.1609/aaai.v34i05.6196.

J. Collins, S. Chand, A. Vanderkop, and D. Howard. A review of physics simulators for robotic applications. *IEEE Access*, 2021.

S. Devlin and D. Kudenko. Plan-based reward shaping for multi-agent reinforcement learning. *The Knowledge Engineering Review*, (1):44–58, 2016.

A. D. Dragan, K. C. Lee, and S. S. Srinivasa. Legibility and predictability of robot motion. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 301–308. IEEE, 2013.

A. Eck, M. Shah, P. Doshi, and L.-K. Soh. Scalable decision-theoretic planning in open and typed multiagent systems. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 7127–7134. AAAI Press, 2020. https://doi.org/10.1609/aaai.v34i05.6200.

K. Genter and P. Stone. Influencing a Flock via Ad Hoc Teamwork. In *Swarm Intelligence*, volume 8667, pages 110–121. Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-09952-1_10.

K. Genter and P. Stone. Adding influencing agents to a flock. In *AAMAS '17*, pages 615–623, 2016.

K. Genter, S. Zhang, and P. Stone. Determining placements of influencing agents in a flock. In *Proceedings of the 14th International Confer-ence on Autonomous Agents and Multiagent Systems*, pages 247–255. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

K. Genter, T. Laue, and P. Stone. Three years of the RoboCup standard platform league drop-in player competition: Creating and maintaining a large scale ad hoc teamwork robotics competition. *Autonomous Agents and Multi-Agent Systems*, 31(4):790–820, 2017. https://doi.org/10.1007/s10458-016-9353-5.

P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research*, 24:49–79, 2005. https://doi.org/10.1613/jair.1579.

B. J. Grosz and S. Kraus. The evolution of Sharedplans. In *Foundations of Rational Agency*, volume 14 of *Applied Logic Series*, pages 227–262. Springer Netherlands, 1999. https://doi.org/10.1007/978-94-015-9204-8_10.

H. Hu, A. Lerer, A. Peysakhovich, and J. Foerster. "Other-play" for zero-shot coordination. In *International Conference on Machine Learning*, volume 119, pages 4399–4410, 2020.

H. Hu, A. Lerer, B. Cui, L. Pineda, N. Brown, and J. Foerster. Off-belief learning. In *International Conference on Machine Learning*, volume 139, pages 4369–4379, 2021.

K. Khetarpal, M. Riemer, I. Rish, and D. Precup. Towards continual reinforcement learning: A review and perspectives. *arXiv:2012.13490*, 2020.

L. Kocsis and C. Szepesvári. Bandit based monte-carlo planning. In J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, editors, *Machine Learning: ECML 2006*, pages 282–293, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-46056-5.

J. Z. Leibo, E. A. Dueñez-Guzman, A. Vezhnevets, J. P. Agapiou, P. Sunehag, R. Koster, J. Matyas, C. Beattie, I. Mordatch, and T. Graepel. Scalable evaluation of multi-agent reinforcement learning with Melting Pot. In *International Conference on Machine Learning*, pages 6187–6199, 2021.

H. Li, T. Ni, S. Agrawal, F. Jia, S. Raja, Y. Gui, D. Hughes, M. Lewis, and K. Sycara. Individualized mutual adaptation in human-agent teams. *IEEE Trans. on Human Machine Systems*, 2021.

S. Liemhetcharat and M. Veloso. Allocating training instances to learning agents for team formation. *Autonomous Agents and Multi-Agent Systems*, 31(4):905–940, 2017. https://doi.org/10.1007/s10458-016-9355-3.

A. Lupu, B. Cui, H. Hu, and J. Foerster. Trajectory diversity for zero-shot coordination. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7204–7213, 2021.

W. Macke, R. Mirsky, and P. Stone. Expected value of communication for planning in ad hoc teamwork. In *AAAI*, volume 35, page 10, 2021.

D. Malik, M. Palaniappan, J. F. Fisac, D. Hadfield-Menell, S. Russell, and A. D. Dragan. An efficient, generalized Bellman update for cooperative inverse reinforcement learning. *arXiv:1806.03820*, 2018.

R. Mead and J. B. Weinberg. Impromptu teams of heterogeneous mobile robots. In *AAAI*, 2007.

F. S. Melo and A. Sardinha. Ad hoc teamwork by learning teammates' task. *Autonomous Agents and Multi-Agent Systems*, 30(2):175–219, 2016. https://doi.org/10.1007/s10458-015-9280-x.

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 *IJCAI*, 2020. https://doi.org/10.24963/ijcai.2020/36.

R. Mirsky, X. Xiao, J. Hart, and P. Stone. Prevention and resolution of conflicts in social navigation–a survey. *arXiv preprint arXiv:2106.12113*, 2021.

Open-Ended Learning Team, A. Stooke, A. Mahajan, C. Barros, C. Deck, J. Bauer, J. Sygnowski, M. Trebacz, M. Jaderberg, M. Mathieu, N. McAleese, N. Bradley-Schmieg, N. Wong, N. Porcel, R. Raileanu, S. Hughes-Fitt, V. Dalibard, and W. M. Czarnecki. Open-ended learning leads to generally capable agents. *arXiv:2107.12808*, 2021.

G. Papoudakis, F. Christianos, A. Rahman, and S. V. Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *ArXiv*, abs/1906.04737, 2019.

G. Papoudakis, F. Christianos, and S. V. Albrecht. Local information agent modelling in partially-observable environments. *arXiv:2006.09447*, 2021.

N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick. Machine theory of mind. In *International Conference on Machine Learning*, pages 4218–4227. PMLR, 2018.

A. Rahman, N. Höpner, F. Christianos, and S. V. Albrecht. Towards open ad hoc teamwork using graph-based policy learning. In *International Conference on Machine Learning*, volume 139. PMLR, 2021.

A. Rahman, E. Fosong, I. Carlucho, and S. V. Albrecht. Towards robust ad hoc teamwork agents by creating diverse training teammates. In *IJCAI Workshop on Ad Hoc Teamwork*, 2022.

M. Ravula, S. Alkoby, and P. Stone. Ad hoc teamwork with behavior switching agents. In *International Joint Conference on Artificial Intelligence*, pages 550–556, 2019. https://doi.org/10.24963/ijcai.2019/78.

J. G. Ribeiro, C. Martinho, A. Sardinha, and F. S. Melo. Assisting Unknown Teammates in Unknown Tasks: Ad Hoc Teamwork under Partial Observability. *arXiv:2201.03538*, 2022.

M. Rovatsos and M. Wolf. Towards social complexity reduction in multiagent learning: The adhoc approach. Technical Report SS-02-02, AAAI Press, 2002.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Series in Artificial Intelligence. Pearson, 4th edition edition, 2021.

P. M. Santos, J. G. Ribeiro, A. Sardinha, and F. S. Melo. Ad hoc teamwork in the presence of non-stationary teammates. In *Progress in Artificial Intelligence*, 2021.

T. Sarratt. Tuning belief revision for coordination with inconsistent teammates. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pages 177–183, 2015.

T. Shu and Y. Tian. M3rl: Mind-aware multi-agent management reinforcement learning. In *International Conference on Learning Representations*, 2019.

M. Shvo and S. A. McIlraith. Active goal recognition. In *AAAI*, volume 34, pages 9957–9966, 2020.

P. Stone, G. A. Kaminka, S. Kraus, and J. S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *AAAI Conference on Artificial Intelligence*, pages 1504–1509, 2010. https://doi.org/10.5555/2898607.2898847.

J. Suriadinata, W. Macke, R. Mirsky, and P. Stone. Reasoning about human behavior in ad hoc teamwork. page 6, 2021.

R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1): 181–211, 1999.

A. Vezhnevets, Y. Wu, M. Eckstein, R. Leblond, and J. Z. Leibo. OPtions as REsponses: Grounding behavioural hierarchies in multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 9733–9742, 2020.

R. E. Wang, S. A. Wu, J. A. Evans, J. B. Tenenbaum, D. C. Parkes, and M. Kleiman-Weiner. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021. https://doi.org/10.1111/tops.12525.

F. Wu, S. Zilberstein, and X. Chen. Online planning for ad hoc autonomous agent teams. In *International Joint Conference on Artificial Intelligence*, pages 439–445, 2011. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-081.

A. Xie, D. P. Losey, R. Tolsma, C. Finn, and D. Sadigh. Learning latent representations to influence multi-agent interaction. In *Proceedings of the Conference on Robot Learning*. PMLR, 2020.

E. S. Yourdshahi, T. Pinder, G. Dhawan, L. S. Marcolino, and P. Angelov. Towards large scale ad-hoc teamwork. In *2018 IEEE International Conference on Agents*, pages 44–49. IEEE, 2018. https://doi.org/10.1109/AGENTS.2018.8460136.

L. Zintgraf, S. Devlin, K. Ciosek, S. Whiteson, and K. Hofmann. Deep interactive Bayesian reinforcement learning via meta-learning. *arXiv:2101.03864*, 2021.