Collaborative Decision-Making in Ad Hoc Teams

Doctoral Consortium

Rupal Nigam
University of Illinois Urbana-Champaign
Champaign, IL, United States
rupaln2@illinois.edu

ABSTRACT

Ad hoc teaming (AHT) is an open challenge for multi-agent systems, in which an autonomous agent must successfully coordinate with other unknown agents. Consider a search-and-rescue mission where robots are deployed from different organizations and expected to cooperate with each other on the fly—these robots may have different biases in how they achieve a given objective (e.g., risky vs. risk-averse search) or have different capabilities (e.g., sensing vs. manipulation). Adapting to such differences would enable agents to effectively and autonomously complete tasks where the team is unknown prior to deployment. In this work, we leverage reinforcement learning and develop a novel algorithm (GPAT) that enables zero-shot coordination in AHTs. We then propose extending the GPAT algorithm to online adaptation settings. Finally, we propose learning from existing complex coordinating systems, such as air traffic, using inverse reinforcement learning. We hypothesize that these learned reward functions can help gain insights into how these systems coordinate and lead to better-informed AHT policies.

KEYWORDS

zero-shot coordination; ad hoc teams; reinforcement learning

ACM Reference Format:
Rupal Nigam. 2026. Collaborative Decision-Making in Ad Hoc Teams: Doctoral Consortium. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/HDMX4554

1 INTRODUCTION

Ad hoc teaming (AHT) has been proposed as a challenge for multi-agent autonomous systems, in which an agent must successfully coordinate with other unknown agents [15]. The ability of intelligent agents to collaborate effectively is crucial for many real-world applications, from autonomous robots performing coordinated tasks to distributed sensor networks optimizing data collection. Multi-agent reinforcement learning (MARL) is a powerful framework for training such agents. However, traditional MARL approaches often focus on pre-defined teams with established communication protocols and interaction histories. Classic MARL algorithms jointly train all agents in the team together, whereas, in AHT, only a single agent (the learner) is controlled [16], [13]. Generalizing to teammates the learner hasn't seen before is a crucial open problem for AHT [8].

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/HDMX4554

This work explores autonomous decision-making algorithms for AHT agents through the following research questions:

(1) Zero-shot Coordination: How can AHT algorithms enable effective coordination when teamed with unseen and unknown teammates?
(2) Effective Adaptation: How can AHT algorithms quickly and effectively adapt to varying teammate behaviors?
(3) Learning from Existing Coordinating Systems: How can we leverage insights from existing coordinating systems to better design AHT algorithms for the real world?

Addressing these research questions is expected to result in the following contributions:

(1) AHT algorithms that account for diverse teammate behaviors, enable zero-shot coordination, and can effectively infer and switch between complementary policies online.
(2) Demonstrations supporting the real-world applicability of AHT algorithms developed through real multi-robot demos and utilization of real-world aircraft data.
(3) Open-source, Gym-compatible infrastructure and software for evaluation of AHT agents in popular domains.

2 AD HOC TEAMS

We modify the general formulation of MMDPs for AHT as follows. Let $\mathrm{a} \in \mathcal{N}$ be the *learner* (i.e., the agent whose policy we aim to optimize) and $\mathcal{N}_u = \mathcal{N} \setminus \{\mathrm{a}\}$ be the complementary set of all teammates (i.e., uncontrolled agents). We assume that each teammate follows a fixed policy, which is unknown to the learner. Teammate policies may be suboptimal with respect to the team reward r and the ad hoc team considered due to, e.g., the teammates being trained for a different task or with different teammates, or being humans and having inherent biases towards different goals. We formally define this model as an ad hoc MMDP.

Definition 1 (Ad Hoc MMDP)**.** An ad hoc MMDP is defined by a tuple $M \coloneqq \langle \mathcal{S}, \mathcal{N}, \mathrm{a}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, p, r, \{\pi^i\}_{i\in\mathcal{N}_u}, \gamma\rangle$, where $\mathrm{a} \in \mathcal{N}$ is the learner, $\mathcal{N}_u = \mathcal{N} \setminus \{\mathrm{a}\}$ is the complementary set of teammates, and $\pi^i : \mathcal{S} \times \mathcal{A}^i \to [0, 1]$ is the fixed policy of teammate i.

We assume that the reward function, r, is non-negative. Note that any bounded reward can be transformed into a non-negative reward because scalar addition renders the reward to be policy invariant. We refer to an ad hoc MMDP M as an ad hoc team. The performance of a learner policy π^{a} in ad hoc team M can be described by its action-value function, $Q^{\pi^{\mathrm{a}},\pi^{-\mathrm{a}}}(s, a^{\mathrm{a}})$, where $\pi^{-\mathrm{a}}$ is the joint policy of all teammates induced by $\{\pi^i\}_{i\in\mathcal{N}_u}$. Our objective is to compute an optimal learner policy, π^{a^*}. Optimizing a learner policy for an ad hoc team M is equivalent to solving a single-agent Markov decision

process with a transition function $\tilde{p}$ that captures the impact of teammate policies $\pi^{-\mathrm{a}}$.

Assume we are given a partially specified ad hoc team $M_{\setminus \mathcal{N}_u} = \langle \mathcal{S}, \cdot, \mathrm{a}, \mathcal{A}^{\mathrm{a}}, p, r, \cdot, \gamma \rangle$ and m possible ad hoc teammates $\{\langle \mathcal{A}^i, \pi^i \rangle\}_{i=1}^m$. Then let $\mathcal{M}$ be the set of possible ad hoc teams induced by those teammates. Given such a set, we formalize our ZSC in AHT problem:

Problem 1 (Zero-shot Coordination for AHT)**.** *Let $\mathcal{M}_0 = \{M_i\}_{i=1}^n \subseteq \mathcal{M}$ be a given set of source AHTs with which the learner can pretrain. Our objective is to synthesize an optimal learner policy $\pi_{n+1}^{a^*}$ for a new AHT $M_{n+1} \in \mathcal{M} \setminus \mathcal{M}_0$ by leveraging information from pretraining on $\mathcal{M}_0$ but with no online learning with the new team M_{n+1}.*

3 AN ALGORITHM FOR ZSC IN AHT

We address the AHT problem defined in Problem 1 through two key ideas. First, we use a GPI policy to *dynamically* leverage a library of pretrained learner policies to coordinate with a new ad hoc team with no online learning. This ability to dynamically leverage policies without inference is important in scenarios where a learner must use multiple pretrained skills to complete a task. We are also motivated by the fact that a GPI policy guarantees improvement over each library policy in single-agent zero-shot transfer settings where dynamics are fixed and rewards are changed [3]. However, in our setting, new ad hoc teammates induce new dynamics due to their (potentially) different policies, while our team rewards are fixed. To ensure policy improvement, one would need to perform policy evaluation for *each pretrained policy* with respect to the new ad hoc team, which requires many online samples.

We address this issue through our second idea—instead of having GPI operate over value functions of pretrained policies evaluated with respect to the team reward, we have it operate over value functions with respect to the learner's difference rewards. We hypothesize that evaluating with respect to the learner's difference rewards emphasizes contributions of the learner towards the team reward, which will then reduce the impact of the distribution shift induced by the new AHT on the actions selected by the GPI policy:

$$\pi^{\mathrm{a}}(s) \in \operatorname*{argmax}_{a^{\mathrm{a}} \in \mathcal{A}^{\mathrm{a}}} \max_{i \in \{1,\ldots,n\}} Q_{i,\Delta r^{\mathrm{a}}}^{\pi_i^{\mathrm{a}^*}}(s, a^{\mathrm{a}}). \tag{1}$$

Based on the ideas presented above, we propose an algorithm, GPI for Ad Hoc Teaming (GPAT), to address ZSC in AHT. We empirically demonstrate GPAT's performance in a multi-agent foraging environment inspired by [4, 7], a multi-agent predator-prey environment as in [7, 18], and Overcooked [6, 17]. We assume linear rewards for all environments, though we also consider a general reward setting for foraging within our ablation study.

We perform experiments in varying difficulties of a cooperative foraging environment, a predator-prey environment, and in Overcooked. The results, presented in [12], suggest that GPAT can effectively achieve ZSC when its library has at least some relevant skills, but can struggle when there are no relevant skills in the library. Additionally, GPAT is able to coordinate well with multiple teammates and handle dynamic environments. In Overcooked, GPAT significantly outperforms our baselines. Compared to other environments, there is also a greater optimality gap between all methods and the oracle policy, demonstrating the difficulty of the task. Our ablation study emphasizes the necessity of difference rewards for aligned value functions and effective policy switching. Finally, we demonstrate our method in a real-world multi-robot setting using Robotis Turtlebot3 Burgers in the foraging environment.

4 ONLINE ADAPTATION IN AHT

To guarantee policy improvement with GPI, we need to perform policy evaluation with the new AHT due to the new dynamics induced by the new teammates. Recall that we cannot perform policy evaluation with GPAT because the learner's pretrained SFs are not valid for the new AHT. Instead, we propose extending Universal Successor Feature Approximators (USFAs) [5], which leverage universal value functions and SFs in single-agent multitask RL transfer settings. We train *universal* learner SFs that are additionally parameterized through the teammate policy: $\psi^{\pi^{\mathrm{a}}, \pi^{-\mathrm{a}}}(s, a^{\mathrm{a}}, e(\pi^{-\mathrm{a}}))$. Here, e is a policy-encoder that maps policies to vectors, $z \in \mathbb{R}^{\kappa}$. The GPI policy for the learner is now defined as

$$\pi^{\mathrm{a}}(s) \in \operatorname*{argmax}_{a^{\mathrm{a}} \in \mathcal{A}^{\mathrm{a}}} \max_{\mathbf{z}} \psi_i^{\pi^{\mathrm{a}*}}(s, a^{\mathrm{a}}, \mathbf{z})^\top \mathbf{w} \tag{2}$$

Note that to use Equation (2) in practice, the learner must know the AHT task, which is induced through teammate behaviours. We propose using online samples from the new AHT to infer z with inverse reinforcement learning. We also derive theoretical results for this method by extending the bound from [2] to AHT settings.

5 LEARNING FROM EXISTING SYSTEMS

We propose leveraging insights from existing, complex coordinating systems, such as air traffic, using inverse reinforcement learning (IRL). The definition of a reward function for a task is crucial and the most succinct description of the task [1]. However, defining effective reward functions for complex scenarios remains a challenge. Instead of explicitly defining rewards, IRL algorithms aim to infer the underlying reward function from expert demonstrations or collected trajectory data [9]. This allows the agent to learn the decision-making strategies employed by, for example, human pilots, potentially leading to the development of more robust and adaptable AI systems for aviation applications. By learning the reward instead through IRL, we can better capture the agent's environment and can improve transfer capabilities between different tasks as the reward function is inherently more transferable than a decision-making policy [14]. Using these recovered reward functions, AHT agents can optimize better-informed policies. We review how IRL has been leveraged for aviation applications in [10] and present our work for learning reward functions from real aircraft trajectories with multi-task IRL in [11].

6 DISCUSSION

This work explores autonomous decision-making algorithms for AHT agents by investigating how to effectively coordinate with unseen and unknown teammates in zero-shot settings. Ongoing efforts and future directions include theoretically analyzing our approach. We also discuss how GPAT can be extended to the online adaptation setting by developing sample-efficient ways to update pretrained policies and using inference for better adaptation to teammates. Finally, we propose leveraging inverse reinforcement learning to gain insights from air traffic, an existing complex coordinating system, to then optimize better-informed AHT policies.

ACKNOWLEDGMENTS

The author would like to thank Dr. Huy Tran for guidance and helpful discussions. This work was funded in part by NASA TTT Award 80NSSC23M0221, ONR N00014-20-1-2249, a NASA grant awarded to the Illinois/NASA Space Grant Consortium, and a GAANN grant from the U.S. Department of Education.

REFERENCES

[1] Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning (ICML '04)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/1015330.1015430

[2] André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado van Hasselt, and David Silver. 2017. Successor features for transfer in reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4058–4068. https://dl.acm.org/doi/10.5555/3294996.3295161

[3] Andre Barreto, Will Dabney, Remi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. 2017. Successor Features for Transfer in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/350db081a661525235354dd3e19b8c05-Paper.pdf

[4] André Barreto, Shaobo Hou, Diana Borsa, David Silver, and Doina Precup. 2020. Fast reinforcement learning with generalized policy updates. *Proceedings of the National Academy of Sciences* 117, 48 (Dec. 2020), 30079–30087. https://doi.org/10.1073/pnas.1907370117 Publisher: Proceedings of the National Academy of Sciences.

[5] Diana Borsa, André Barreto, John Quan, Daniel Mankowitz, Rémi Munos, Hado van Hasselt, David Silver, and Tom Schaul. 2018. Universal Successor Features Approximators. http://arxiv.org/abs/1812.07626 arXiv:1812.07626 [cs, stat].

[6] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. 2019. On the Utility of Learning about Humans for Human-AI Coordination. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vol. 32. arXiv:1910.05789

[7] Pengjie Gu, Mengchen Zhao, Jianye Hao, and Bo An. 2022. Online Ad Hoc Teamwork under Partial Observability. In *International Conference on Learning Representations*. https://openreview.net/forum?id=18Ys0-PzyPI

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

[9] Andrew Y. Ng and Stuart J. Russell. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 663–670. https://dl.acm.org/doi/10.5555/645529.657801

[10] Rupal Nigam, Jimin Choi, Niket Parikh, Max Z. Li, and Huy Tran. 2025. *A Survey of Current Applications of Inverse Reinforcement Learning in Aviation and Future Outlooks*. AIAA SCITECH 2025 Forum. https://doi.org/10.2514/6.2025-1540 arXiv:https://arc.aiaa.org/doi/pdf/10.2514/6.2025-1540

[11] Rupal Nigam, Nadezhda D. Dimitrova, Aastha Acharya, Huy T. Tran, and Husni R. Idris. 2026. Understanding Visual Flight Rule Traffic Behavior Through Inverse Reinforcement Learning. *AIAA Aviation* (2026).

[12] Rupal Nigam, Niket Parikh, Hamid Osooli, Mikihisa Yuasa, Jacob Heglund, and Huy T. Tran. [n.d.]. *Zero-Shot Coordination in Ad Hoc Teams with Generalized Policy Improvement and Difference Rewards*. https://doi.org/10.48550/arXiv.2510.16187 arXiv:2510.16187 [cs]

[13] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *J. Mach. Learn. Res.* 21, 1, Article 178 (Jan. 2020), 51 pages.

[14] Stuart Russell. 1998. Learning agents for uncertain environments (extended abstract). In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory* (Madison, Wisconsin, USA) *(COLT' 98)*. Association for Computing Machinery, New York, NY, USA, 101–103. https://doi.org/10.1145/279943.279964

[15] Peter Stone, Gal Kaminka, Sarit Kraus, and Jeffrey Rosenschein. 2010. Ad Hoc Autonomous Agent Teams: Collaboration without Pre-Coordination. *Proceedings of the AAAI Conference on Artificial Intelligence* 24, 1 (Jul. 2010), 1504–1509. https://doi.org/10.1609/aaai.v24i1.7529

[16] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (Stockholm, Sweden) *(AAMAS '18)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2085–2087.

[17] Rose E. Wang, Sarah A. Wu, James A. Evans, Joshua B. Tenenbaum, David C. Parkes, and Max Kleiman-Weiner. 2020. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. https://doi.org/10.48550/arXiv.2003.11778 arXiv:2003.11778 [cs].

[18] Dong Xing, Pengjie Gu, Qian Zheng, Xinrun Wang, Shanqi Liu, Longtao Zheng, Bo An, and Gang Pan. 2023. Controlling Type Confounding in Ad Hoc Teamwork with Instance-wise Teammate Feedback Rectification. In *Proceedings of the 40th International Conference on Machine Learning*.