Nested Training for Mutual Adaptation in Human-AI Teaming*

Extended Abstract

Upasana Biswas
Arizona State University
Tempe, Arizona, USA

Durgesh Kalwar
Arizona State University
Tempe, Arizona, USA

Subbarao Kambhampati
Arizona State University
Tempe, Arizona, USA

Sarath Sreedharan
Colorado State University
Fort Collins, Colorado, USA

ABSTRACT

Mutual adaptation is essential in human–robot teaming, as humans adjust their behavior in response to the robot. Prior work trains against diverse but static partners, missing adaptive human responses, while simultaneous multi-agent learning often yields brittle coordination conventions that fail to generalize. We model human–robot teaming as a finite-Level Interactive Partially Observable Markov Decision Process (I-POMDP), explicitly representing human adaptation within the state. To approximately solve this formulation, we introduce a nested training regime in which agents at a level are trained against adaptive agents at a level below. This exposes agents to adaptation while preventing emergence of opaque coordination strategies. In the Overcooked domain with required-cooperation, our method outperforms standard baselines with unseen adaptive partners and demonstrates stronger adaptability during interaction.

Keywords

I-POMDP, Human-AI Teaming, Cooperative AI, Multi-agent Reinforcement Learning

ACM Reference Format:
Upasana Biswas, Durgesh Kalwar, Subbarao Kambhampati, and Sarath Sreedharan. 2026. Nested Training for Mutual Adaptation in Human-AI Teaming: Extended Abstract. 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/OFJL6947

1 INTRODUCTION

Designing reinforcement learning (RL) agents that can collaborate with diverse human partners remains a central challenge in multi-agent reinforcement learning (MARL) due to the inherent diversity in human partners the agent can be paired with.

Existing approaches tackle this by training against populations of simulated strategies [1, 3, 4, 7, 8, 17, 22], improving robustness to unseen partners but typically treating them as static. The resulting policies are often conformant rather than truly adaptive [2, 12, 16]. More fundamentally, human–AI collaboration is inherently interactive. Standard training paradigms largely treat adaptation as one-sided, focusing on the agent responding to a fixed or implicitly stationary partner [9, 19, 21], not considering that humans adapt their strategies in response to the agent's behavior [11, 13].

We address the challenge of mutual adaptation by modeling human-robot teaming as a finitely nested Interactive Partially Observable Markov Decision Process (I-POMDP) [6] and introducing a nested reinforcement learning regime that approximates Level-2 reasoning while preventing convergence to partner-specific conventions. Our method consistently outperforms state-of-the-art baselines in a required-cooperation variant of Overcooked [4, 5], achieving higher success rates and more stable coordination with previously unseen adaptive partners.

2 METHODOLOGY

We ground our study in a setiing where (i) the presence of multiple equilibria corresponding to distinct solutions to the task, and (ii) mutual adaptation, where the human adjusts their behavior in response to the agent. We model human–AI collaboration as a finitely nested I-POMDP [6]. We learn Level-2 reasoning via a nested training regime in a two-player human–robot game, consistent with empirical findings that human belief modeling rarely exceeds depth two [15, 18].

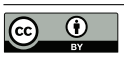
At **Level-1**, human policies $\{\pi_H^j\}_{j=1}^J$ are trained against a finite set of fixed (Level-0) robot policies $\{\pi_R^k\}_{k=1}^K$. Each resulting π_H^j at Level-1 is adaptive to Level-0 policies, corresponding to Level-1 reasoning where the human maintains beliefs over possible robot behaviors. At **Level-2**, the robot policy π_R is trained against the set of adaptive human policies $\{\pi_H^j\}_{j=1}^J$, which represent Level-1 human models. This corresponds to Level-2 reasoning, such the robot maintains beliefs over Level-1 policies i.e. human adaptive policies. Because training occurs against fixed lower-Level policies rather than simultaneous co-adaptation, the learned policy avoids collapse to a single coordination convention (proof in Appendix A.4).

To approximate belief updates tractably, we learn a latent embedding $z_t = f_\theta(h_t)$, where h_t summarizes interaction history, and condition the policy as $a_t \sim \pi_\theta(a \mid o_t, z_t)$. This amortizes uncertainty over partner types while enabling end-to-end optimization. Overall, the nested regime operationalizes Level-2 interactive reasoning within a practical RL framework, promoting generalization across unseen adaptive partners.

3 RESULTS

We evaluate agents using average success rate over N episodes when paired with eight unseen adaptive partners across 10 rounds

*An extended version of this paper - https://arxiv.org/abs/2602.17737. Code is available at https://github.com/upasana27/adaptive-RL.



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/OFJL6947

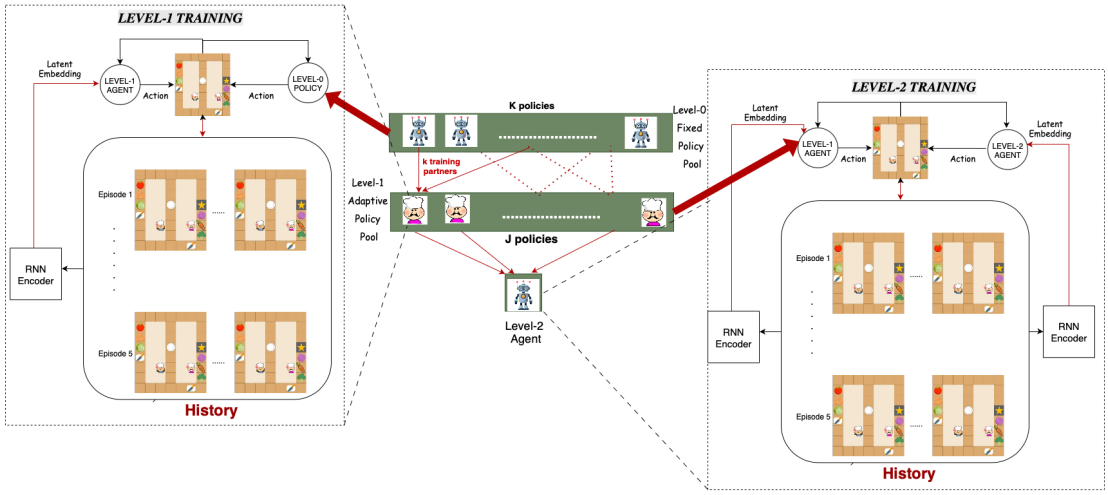


Figure 1: Overview of the nested training regime. Level-1 human policies are trained against fixed robot policies, producing a set of adaptive behaviors. The Level-2 robot then trains against these adaptive human policies, using a latent embedding to summarize interaction history and approximate nested I-POMDP beliefs, enabling reasoning over multiple adaptive partner strategies.

(5 or 25 episodes per round; environment details in Appendix A.1). We compare against LIAM [14], LILI [20], PACE [10], and the Generalist policy. Our method consistently achieves the highest success rates. In the short evaluation, it attains 0.90 average success, substantially outperforming the next-best baseline (GENERALIST, 0.575), while PACE and LILI perform inconsistently and LIAM fails to complete tasks. Under extended evaluation, performance remains high (0.935), whereas baselines show limited improvement despite longer interaction horizons. Moreover, our agent performs reliably across all partner seeds (Table 1), while baselines succeed only with specific partners, indicating overfitting to particular behaviors. To understand the performance gap, we analyze the interaction dynamics between teammates. Agents trained with the proposed regime exhibit structured mutual adaptation consistent with the Level-2 I-POMDP formulation. Level-1 agents display waiting behavior, delaying commitment until their Level-0 partner reveals their type through their actions, reflecting Level-1 reasoning over L0 strategy types. Level-2 agents adapt further by proactively committing first, anticipating the waiting behavior of Level-1 partners and reasoning over their partner's adaptation. These behavioral patterns are statistically validated. In contrast, baseline agents exhibit persistent oscillations between recipe choices when paired with adaptive partners, failing to establish a stable coordination convention (trajectory analysis in Appendix A.5). Together, these results demonstrate that explicitly modeling partner adaptation enables stable coordination under mutual adaptation, whereas standard training methods do not.

Agent	P1	P2	P3	P4	P5	P6	P7	P8
Short Evaluation (10 rounds × 5 episodes)								
Proposed Method	1.0	1.0	0.8	0.8	1.0	1.0	1.0	0.6
PACE	0.4	1.0	0.4	0.4	0.4	1.0	0.4	0.4
Generalist	0.4	0.4	1.0	0.6	0.0	0.8	0.8	0.6
LILI	0.8	0.2	0.6	0.4	0.2	0.4	0.6	0.4
LIAM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Extended Evaluation (10 rounds × 25 episodes)								
Proposed Method	1.0	1.0	1.0	0.88	1.0	1.0	0.96	1.0
PACE	0.76	0.76	0.2	0.6	0.76	0.76	0.28	0.76
Generalist	0.6	0.68	0.76	0.68	0.68	0.72	0.4	0.64
LILI	0.36	0.52	0.72	0.6	0.44	0.4	0.52	0.48
LIAM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 1: Per-partner success rates under short and extended evaluations. Baselines exhibit high variance and instability across adaptive partners, while our method maintains consistently strong performance.

4 CONCLUSION

We introduced a nested reinforcement learning framework for enabling mutual adaptation in human–AI collaboration. Grounded in a finitely nested I-POMDP formulation, our approach structurally prevents convergence to arbitrary conventions while promoting generalization across adaptive partners. Empirical results demonstrate that the proposed method outperforms existing baselines and achieves stable coordination with previously unseen adaptive agents. Future work will include conducting user studies, performing qualitative analysis of the adaptive behaviors that emerge, and extending the framework to mixed-motive settings.

ACKNOWLEDGEMENTS

This research is supported in part by ONR grant N0001423-1-2409 and DARPA grant HR00112520016, and gifts from Qualcomm and Amazon. The research is supported in part by NSF grant 2303019 and a JP Morgan Faculty Award.

References

[1] Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. 2020. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence* 280 (March 2020), 103216. https://doi.org/10.1016/j.artint.2019.103216

[2] Upasana Biswas, Vardhan Palod, Siddhant Bhambri, and Subbarao Kambhampati. 2025. Who is Helping Whom? Analyzing Inter-dependencies to Evaluate Cooperation in Human-AI Teaming. arXiv:2502.06976 [cs.MA] https://arxiv.org/abs/2502.06976

[3] Rodrigo Canaan, Xianbo Gao, Julian Togelius, Andy Nealen, and Stefan Menzel. 2023. Generating and Adapting to Diverse Ad Hoc Partners in Hanabi. *IEEE Transactions on Games* 15, 2 (2023), 228–241. https://doi.org/10.1109/TG.2022.3169168

[4] Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca Dragan. 2020. On the Utility of Learning about Humans for Human-AI Coordination. arXiv:1910.05789 [cs.LG] https://arxiv.org/abs/1910.05789

[5] Rujikorn Charakorn, Poramate Manoonpong, and Nat Dilokthanakul. 2023. Generating diverse cooperative agents by learning incompatible policies. In *The Eleventh International Conference on Learning Representations*.

[6] Piotr J Gmytrasiewicz and Prashant Doshi. 2005. A framework for sequential planning in multi-agent settings. *Journal of Artificial Intelligence Research* 24 (2005), 49–79.

[7] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. 2017. Population Based Training of Neural Networks. arXiv:1711.09846 [cs.LG] https://arxiv.org/abs/1711.09846

[8] Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. 2021. Trajectory Diversity for Zero-Shot Coordination. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 7204–7213. https://proceedings.mlr.press/v139/lupu21a.html

[9] Long Ma, Yuanfei Wang, Fangwei Zhong, Song-Chun Zhu, and Yizhou Wang. 2024. Fast Peer Adaptation with Context-aware Exploration. arXiv:2402.02468 [cs.AI] https://arxiv.org/abs/2402.02468

[10] Long Ma, Yuanfei Wang, Fangwei Zhong, Song-Chun Zhu, and Yizhou Wang. 2024. Fast peer adaptation with context-aware exploration. *arXiv preprint arXiv:2402.02468* (2024).

[11] 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. arXiv:2202.10450 [cs.MA] https://arxiv.org/abs/2202.10450

[12] Reuth Mirsky, Ignacio Carlucho, Muhammad Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano Albrecht. 2022. *A Survey of Ad Hoc Teamwork Research*. 275–293. https://doi.org/10.1007/978-3-031-20614-6_16

[13] Stefanos Nikolaidis, Anton Kuznetsov, David Hsu, and Siddhartha Srinivasa. 2016. Formalizing human-robot mutual adaptation: A bounded memory model. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 75–82. https://doi.org/10.1109/HRI.2016.7451736

[14] Georgios Papoudakis, Filippos Christianos, and Stefano Albrecht. 2021. Agent modelling under partial observability for deep reinforcement learning. *Advances in Neural Information Processing Systems* 34 (2021), 19210–19222.

[15] Josef Perner and Heinz Wimmer. 1985. "John thinks that Mary thinks that..." attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology* 39, 3 (1 Jan. 1985), 437–471. https://doi.org/10.1016/0022-0965(85)90051-7

[16] Peter Stone, Gal A. Kaminka, Sarit Kraus, and Jeffrey S. Rosenschein. 2010. Ad hoc autonomous agent teams: collaboration without pre-coordination. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence* (Atlanta, Georgia) *(AAAI'10)*. AAAI Press, 1504–1509.

[17] Gerald Tesauro. 1994. TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play. *Neural Computation* 6, 2 (1994), 215–219. https://doi.org/10.1162/neco.1994.6.2.215

[18] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. 2005. Understanding and Sharing Intentions: The Origins of Cultural Cognition. *Behavioral and Brain Sciences* 28 (11 2005), 675–735. https://doi.org/10.1017/S0140525X05000129

[19] Yuanfei Wang, Fangwei Zhong, Jing Xu, and Yizhou Wang. 2022. ToM2C: Target-oriented Multi-agent Communication and Cooperation with Theory of Mind. arXiv:2111.09189 [cs.MA] https://arxiv.org/abs/2111.09189

[20] Annie Xie, Dylan Losey, Ryan Tolsma, Chelsea Finn, and Dorsa Sadigh. 2021. Learning latent representations to influence multi-agent interaction. In *Conference on robot learning*. PMLR, 575–588.

[21] Xiaopeng Yu, Jiechuan Jiang, Wanpeng Zhang, Haobin Jiang, and Zongqing Lu. 2022. Model-Based Opponent Modeling. arXiv:2108.01843 [cs.LG] https://arxiv.org/abs/2108.01843

[22] Rui Zhao, Jinming Song, Yufeng Yuan, Hu Haifeng, Yang Gao, Yi Wu, Zhongqian Sun, and Yang Wei. 2022. Maximum Entropy Population-Based Training for Zero-Shot Human-AI Coordination. arXiv:2112.11701 [cs.AI] https://arxiv.org/abs/2112.11701