Learning human-robot collaboration from human feedback

Jerry Zhi-Yang He (hzyjerry@berkeley.edu), Akshara Rai (akshararai@meta.com) and Anca D. Dragan (anca@berkeley.edu)

Motivation. Consider a human-robot team collaborating on everyday tasks like unloading groceries, preparing dinner, or cleaning the house. Such an assistive robot should coordinate with its partner to efficiently complete the task, without getting in their way. For example, while tidying the house, if its partner starts cleaning the kitchen, the robot could start cleaning the living room to maximize efficiency. If the robot notices its partner loading the dishwasher, it should prioritize bringing dirty dishes from the living room to the kitchen, instead of rearranging cushions. This requires the robot to reason about not only its own embodiment (to avoid getting in the way of the human), but also about its partner's actions and intentions to efficiently assist them. A commercially useful robot should be able to achieve some level of commonsense reasoning of human intentions through pre-training. Then, from interactions and additional feedback, the robot should be able to further accommodate its partner's specific habits and preferences. In this project, we aim to study approaches that can enable a natural way for humans and robots to collaborate, while adapting to each other's needs, and incorporating and seeking human feedback.

Related Work. Embodied AI has seen great advancements in simulation platforms [1, 2, 3] and new task specifications [4, 5]. Object rearrangement is a task of importance for home robotics [6], and a variety of simulators support it [7, 8]. We will utilize the Home Assistant Benchmark (HAB) in AI Habitat [8] for human-robot collaboration. Multi-agent RL (MARL) studies multiple agents acting to complete a task like moving furniture [9]. Unlike these works, we focus on learning embodied agents that can adapt to *new* partner preferences at evaluation time, which one can formulate in two different ways: as *zero-shot coordination* (ZSC) [10, 11] or as assistance POMDPs [12]. Overcooked [13] and Hanabi [14] are common benchmarks for studying such problems [10, 15, 16] in discrete state and action spaces. In contrast to these, we will study ZSC in a complex, visually realistic 3D environment using continuous observations and actions. Learning from human feedback aims to align the objective of the agent with that of the human [17]. While the underlying human reward is often subtle and expensive to collect, researchers have found that people reveal their preferences in various ways through language or reactions [18] and proposed methods [19] for studying them. Recent works [20] have extended preference learning to deep learning with high dimensional features, leading to breakthroughs in LLM [21].

Novelty and Innovation. Our novelty lies in the problem we address - adapting robotic agents to human partners in human-robot collaboration settings. While previous works study learning from human preferences where the robot acts in isolation, we focus on the problem where the robot needs to personalize while collaborating with the human. Our innovation lies in adapting recent progress in learning from human feedback to the task of human-robot collaboration. We believe that this is an understudied, yet incredibly important direction that can guide the personalization of assistive agents such as ChatGPT to collaborate with their users and tune in to their goals and preferences though interactions with them. Our realistic, long-horizon, and embodied test-bed based in Habitat also makes the study more convincing and applicable to embodied applications in robotics.

Technical Objective. Specifially, we aim to achieve the following goals as part of this collaboration:

1. Adapt the AI Habitat simulation [8] to study human-robot collaboration, with a focus on everyday, long-horizon tasks, dealing with realistic sensing and actuation, partial observability in collaborative tasks and unknown partner states and intentions.

2. Develop zero-shot coordination approaches which perform well at long-horizon, everyday tasks. Evaluate learned policies in a human-in-the-loop setting.

3. Develop algorithms that enable few-shot learning and learning from human feedback for adapting the learned policies for personalization.

Potential for Collaboration. We will use the AI Habitat simulator from Meta AI, including recently developed features like human simulation, human-in-the-loop evaluation and Spot robot stack. UC Berkeley collaborators will provide expertise in human-robot collaboration, especially, learning from human preferences and feedback.

References

[1] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor. *arXiv preprint arXiv:1712.05474*, 2019.

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Learning from rgb-d data in indoor environments. *3DV*, 2017.

[3] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. *CVPR*, 2018.

[4] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat. *ICCV*, 2019.

[5] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited. *arXiv preprint arXiv:2006.13171*, 2020.

[6] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. Rearrangement: A challenge for embodied ai. 2020.

[7] Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Manipulathor: A framework for visual object manipulation. In *CVPR*, 2021.

[8] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, et al. Habitat 2.0. *arXiv preprint arXiv:2106.14405*.

[9] Unnat Jain, Luca Weihs, Eric Kolve, Mohammad Rastegari, Svetlana Lazebnik, Ali Farhadi, Alexander G. Schwing, and Aniruddha Kembhavi. Collaborative visual task completion. In *CVPR*, 2019.

[10] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020.

[11] Jerry Zhi-Yang He, Zackory Erickson, Daniel S Brown, Aditi Raghunathan, and Anca Dragan. Learning representations that enable generalization in assistive tasks. In *CoRL*, pages 2105–2114. PMLR, 2023.

[12] Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan. The assistive multi-armed bandit. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 354–363. IEEE, 2019.

[13] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *NeurIPS*, 32, 2019.

[14] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020.

[15] Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-belief learning. In *International Conference on Machine Learning*, pages 4369–4379. PMLR, 2021.

[16] Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *International Conference on Machine Learning*, pages 7204–7213. PMLR, 2021.

[17] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. *Active preference-based learning of reward functions.* 2017.

[18] Hong Jun Jeon, Smitha Milli, and Anca Dragan. Reward-rational (implicit) choice: A unifying formalism for reward learning. *Advances in Neural Information Processing Systems*, 33:4415–4426, 2020.

[19] Erdem Biyik and Dorsa Sadigh. Batch active preference-based learning of reward functions. In *Conference on robot learning*, pages 519–528. PMLR, 2018.

[20] Daniel S Brown and Scott Niekum. Deep bayesian reward learning from preferences. *arXiv preprint arXiv:1912.04472*, 2019.

[21] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.