Learning Human-Robot Interactions to improve Human-Human Collaboration

Radu Stoican, Angelo Cangelosi, Christian Goerick, Thomas Weisswange

2022

Preprint:

This is an accepted article published in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022) - Workshop on Human Theory of Machines and Machine Theory of Mind for Human-Agent Teams (TOM4HAT). The final authenticated version is available online at: https://doi.org/[DOI not available]

Learning Human-Robot Interactions to improve Human-Human Collaboration

Radu Stoican[1], Angelo Cangelosi[1], Christian Goerick[2], and Thomas H. Weisswange[3]

***Abstract*— Most research in human-robot interaction focuses on either the single-human case or the multi-human case where there is direct interaction between the robot and each human. The multi-human scenario in which some of the humans depend on the robot, but do not interact with it directly, is currently less studied. In this paper, we introduce a human-human-robot collaboration task, in which the robot interacts directly with only one of the humans. The goal of the robot is to find the optimal way of helping the two humans achieve their objective. We decided to use meta-reinforcement learning to solve the task, giving the robot the ability to quickly adapt to new human behavior. We trained and tested an agent on a version of the proposed environment that uses simulated human behavior. Initial results show that our task is learnable.**

I. INTRODUCTION

Human-robot collaboration (HRC) is concerned with human-robot interaction (HRI) scenarios in which humans and robots work together, aiming to achieve a shared objective [1]. Researchers study HRC to measure and understand the effects robots have on humans in scenarios where robots are integrated into teams of humans.

Direct interaction is the most obvious and most studied type of interaction in HRI. However, in a human-robot team, a robot's behavior may affect the entire team, even if it interacts with only some of the humans. These indirect interactions are less studied but can be just as important for HRC.

It is important to consider trust when humans and robots interact. As a consequence, there is a large amount of work on trust in HRI, most of it focusing on the humans' trust in robots [2]. Some issues make the development of trustworthy social robots difficult. One of these is finding the optimal amount of trust [2], since maximizing trust can lead to over-trusting, which can cause problems [3]. Another issue is caused by the large number of factors that can affect trust. The extensive body of research on trust in HRI has shown that, besides the robot's behavior, trust can also be affected by the human's personality or the HRI task [3].

Humans' wellbeing (WB) is another social construct that can be influenced by a robot. The study of WB in the broader field of human-computer interaction is closely related to positive computing. Positive computing focuses on the design of technologies that have the goal of directly improving users' WB [4], [5]. Software built for positive computing promotes positive emotions, and can have objectives like stress management [6], self-help [7] or socialising [8]. In general, WB can be split into evaluative, hedonic, eudaimonic, and social WB [9], [10]. Hedonic WB is concerned with day-to-day emotions [9] and can be relatively easy to influence and measure in HRI experiments.

In this work, we propose a human-human-robot collaboration environment in which the main focus is indirect interaction. In the environment, a third party, represented by a human, collaborates with a pair consisting of a robot and another human. The assumption we make is that the third party is affected by the performance of the human-robot team, but only interacts directly with the other human, not the robot. The environment we propose can be used to measure and analyze the third party's trust and hedonic WB. Because the third party has to interact with the other human to be affected by the robot's actions, it is possible to study how the robot's behavior influences trust and WB in human-human interaction.

We use reinforcement learning (RL) to train an agent that solves the proposed environment. RL studies both the set of problems concerned with behaving optimally in an environment and the algorithms used to find this behavior [11]. The optimal behavior is learned by interacting with the environment. RL has been used in human-robot interaction scenarios (HRI) [12], where agents learn how to behave by interacting with humans. However, RL is known to be very sample inefficient [13]. This is a problem in robotics, where collecting data is expensive [14]. RL is also limited when it comes to quickly adapting to new environments. These two issues are especially true in HRI [12], [15], due to the complex human behavior.

An approach for avoiding the sample inefficiency issue when adapting to new environments is meta-reinforcement learning (meta-RL). Meta-RL agents are trained on several tasks to find a policy that can be easily adapted to new, but similar, tasks. In other words, solving a single task is no longer the main goal. Meta-RL is more concerned with learning how to learn about new tasks. Because of this, meta-RL is appropriate for our environment, as some of the agent's objectives are to quickly understand the third party's goals and to adapt to dynamic teams, where each human member brings in different expertise and expectations.

However, a limitation of current meta-RL algorithms is that they only generalize to very similar tasks [16]. This provides restrictions when applying meta-RL to real-world robotics tasks, like HRI. Algorithms that adapt better and benchmarks that contain more diverse tasks are required before meta-RL can provide the type of generalization required for interacting with diverse, complex environments.

[1]Department of Computer Science, University of Manchester, UK
[2]Honda Research Institute USA, Inc.
[3]Honda Research Institute Europe GmbH, Offenbach, Germany

Meta-RL holds potential for HRI, with agents being able to quickly adapt to new humans or new tasks defined by the humans. Yet, research on the usage of meta-RL for HRI is still very limited. Meta-RL has been used for studying how adapting fast to new humans influences trust in HRI [17]. It has also been used as an attempt at solving an important issue in RL for HRI, where the reward function depends on the social context [18]. Since there is no unique optimal reward function for all tasks, the agent uses meta-RL to adapt to each reward. Still, these studies are restricted by the meta-RL issues outlined above.

Therefore, besides the focus on multi-human indirect interaction, our proposed environment has a second objective. The environment is defined as a set of similar tasks. These tasks are much more varied than most of the task collections used in the current meta-RL literature. We aim to create tasks that can only be solved by an agent that adapts. We also aim to ensure that a task's optimal exploration policy is specific to that task. Therefore, it is expected that existing meta-RL algorithms will not be adaptable enough to solve our environment.

In this paper, we present the main concepts of our work, while leaving objectives like measuring trust and WB as future work. We make two contributions:

1) Propose an HRI environment that can only be solved by an artificial agent that understands indirect interactions and quickly adapts to new humans.
2) Present preliminary empirical results that show how RL can be used to solve the proposed environment.

II. RELATED WORK

A. Trust and Wellbeing in Human-Robot Interaction

In [19], trust modeling is used to assess humans' skills in an HRC task. The paper shows that knowing when to trust a human and when not to can lead to improved task performance. There are similarities between this work and ours, but we choose to focus more on the multi-human scenario and indirect interactions. The experiments presented in [20] aimed to show that when given the choice between using a human and a robot to complete a task, participants are more likely to choose the agent they trust more. While trust impacted the participants' choice, other factors, like task type, were more important. Trust was also studied in the context of HRC in [21], with results showing that participants' trust decreases when the robot makes a mistake, unless the mistake was due to a limitation already known by the participant.

The literature on WB in HRI is still limited, compared to trust. A study on how factory workers' WB is affected after introducing a collaborative robot is presented in [22]. The effects on medical staff after using robots in a hospital are analyzed in [23]. However, their interviews show that staff are concerned with issues like depersonalization, which relates more to eudaimonic WB than hedonic WB. The hedonic WB of passengers in self-driving cars is measured in [24]. Despite this not being an HRI study, there are similarities to our work, as the focus is on a third party, the passenger, that has no direct control over the artificial agent.

B. Reinforcement Learning in Human-Robot Collaboration

In [25], RL is used to study how the fairness of the artificial agent affects humans' trust in a multi-human HRC scenario. The goal of the agent is to share resources based on each participant's usefulness in the task. The authors report that an unfair agent that favors stronger participants leads to the weaker participants trusting the system less. An RL framework for improving safety in HRC is presented in [26]. The framework takes safety requirements as inputs and uses RL to produce a safe policy.

C. Meta-Reinforcement Learning

Meta-learning focuses on the idea of learning to learn: given a set of tasks and an algorithm, the algorithm will get better at solving tasks from the set as it gains more data for each task and the number of tasks increases [27]. Meta-RL combines meta-learning with RL.

There are three main types of meta-RL algorithms: recurrent, gradient-based, and context-based. Context-based algorithms are the current state-of-the-art in meta-RL. In context-based meta-RL, the task-specific context collected from a task can be used to adapt to that task. Notable algorithms include PEARL [28], variBAD [29] and ELUE [30].

A critical issue in meta-RL is that many algorithms have only been proven to work on datasets that have little variations between tasks. Meta-World [16] is a meta-RL benchmark that contains a large variety of tasks. There are 50 robot manipulation tasks with significant differences between them, e.g. pick and place, pushing, opening doors, etc. Moreover, each task has randomized parameters, giving an even larger distribution. Therefore, only a highly adaptable RL agent would be able to solve Meta-World. However, each Meta-World task is limited by simplistic randomized parameters that only change the location of objects or goals. Contrary to this, the parameters of each of our tasks are given by complex human behavior.

III. METHODS

A. Meta-Reinforcement Learning

RL problems are usually represented as Markov Decision Processes (MDPs). An MDP is solved by finding an optimal policy that maximizes the expected accumulated reward that an agent gets when navigating the MDP. The dynamics of the environment are not known to the agent, so finding an optimal policy has to be done through trial and error, by interacting with the environment.

Meta-RL extends this definition by using a distribution of tasks, where each task is an MDP. While there are differences between these tasks, there are also similarities, as all MDPs in the distribution have shared features. A meta-RL agent trains on multiple tasks during the meta-training phase. This way, it learns about the shared structure of the MDPs and

finds a policy that can quickly adapt to the task-specific structure of each MDP. Testing is performed on new tasks from the same distribution during meta-testing. The agent is expected to adapt efficiently, after training on only a few samples from each new task.

In contrast to meta-RL, a standard RL agent trained on a task distribution might try to find a policy that is globally optimal for all tasks. However, this optimal policy might fail to generalize to new tasks, if they are too different from those seen during training. Moreover, the assumption that there exists a policy that is optimal on all tasks from the distribution is not usually true, in which case only a policy that adapts can be optimal [31].

B. The Environment

The environment we propose is a set of tasks, containing a theoretically infinite amount of MDPs. All tasks are represented as human-robot collaboration scenarios with one robot and two humans. The scenario is a construction game with colored blocks, in which the three agents work together to build a structure. A simulated version of the environment is shown in Fig. 1. The role of the RL agent is to choose which colored block the robot should hand to one of the humans, based on the game's rules and the humans' objectives.

Indirect interaction is a central feature of the environment. Let one of the humans be called h_1. The robot only interacts directly with h_1, and h_1 interacts directly with the other human, h_2. Therefore, the robot and h_2 interact only through h_1. Additionally, h_2 is affected by the outcome of the task, and this outcome can be influenced by the robot, so h_2 does depend on the robot. This dependence is also indirect, as h_2 can only be affected by the robot's actions by interacting with the other human, h_1. Consequently, in the case that either the robot or the other human behaves sub-optimally, it is difficult for h_2 to assign blame. This means that the robot's behavior impacts both the task performance and the trust between the two humans. The relationships between the three agents are compactly shown in Fig. 2.

The agent that controls the robot is trained using RL,

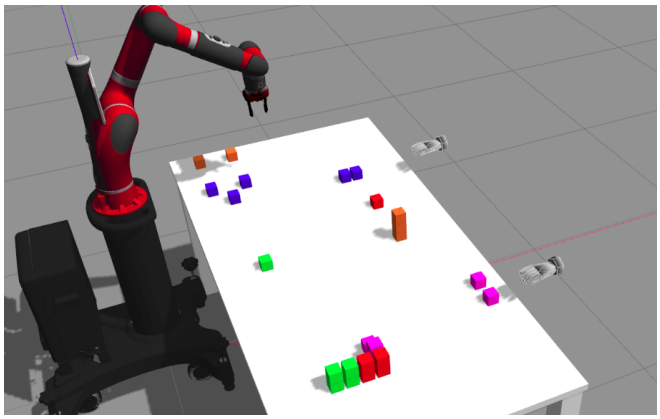

Fig. 1. A task in our simulated environment. The robot and the two humans (represented by hand models) work together to build a structure with the blocks provided.

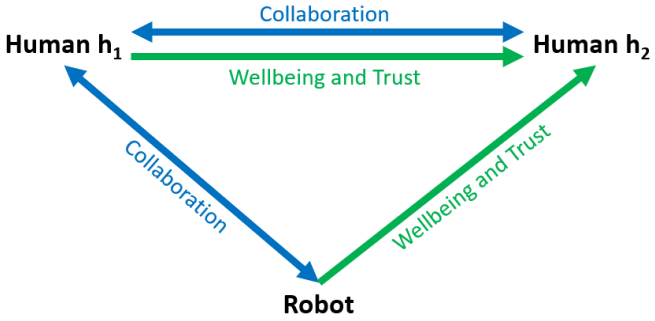

Fig. 2. The relationships between the robot and the two humans, h_1 and h_2. Blue arrows indicate direct interactions. Green arrows show which members affect h_2's wellbeing and trust. The robot targets to influence aspects of h_2, but only interacts directly with h_1.

so designing an appropriate reward function is important. Rewards define the goal of each task. The reward function is part of the MDP, so a poorly designed reward function leads to an optimal agent that solves an MDP we might not be interested in. Recall that we propose an environment for analyzing the effects of the robot's behavior on the trust between h_2 and the human-robot team, and on the hedonic WB of h_2. In this initial work, we approximate these two social constructs through overall task performance. Task performance is affected not only by the robot but by h_1 as well. The human h_1 can make errors while performing the task and the error probability varies across different humans. The error probability also depends on the robot's behavior (i.e. h_1 makes more errors in some states than in others). Allowing h_1 to make mistakes gives meaning to the human-robot interactions. If h_1's behavior was deterministic, the optimal RL agent would only have to focus on the task and ignore the humans. However, this also means that the agent now has to adapt to human behavior, which is why we propose using meta-RL.

Additionally, each task contains a set of sub-goals that h_2 wants to complete as fast as possible and in a specific order. These sub-goals are completed correctly only if both the robot and the human h_1 behave optimally. Therefore, we assume that h_2's trust and WB depend on how the sub-goals are completed. Then, for a task with n sub-goals, we define the reward at time-step t when working on the i-th sub-goal as

$$r_{t,i} = -\frac{t - k_i}{i}, \tag{1}$$

where $i \in \{1, 2, ..., n\}$ and k_i is the time-step at which the i-th sub-goal was started. The accumulated reward increases when sub-goals are completed in the correct order and there are no delays caused by robot or human errors. Thus, the reward cannot be maximized by an RL agent that ignores the humans' goals and behaviors.

Each task in the distribution can be manually defined or randomly generated. Differences between tasks include the color and number of blocks, the number of sub-goals, the correct order of the sub-goals, and the requirements for completing each sub-goal. Meta-RL algorithms from recent

literature are usually tested on very simple manipulation or navigation tasks, where the position of the goal changes across tasks [28], [29], [32]. These environments are, by definition, distributions of tasks, and they have no global optimal policy that can solve all MDPs. However, there usually is a global optimal policy for adapting to them. In our proposed environment, there can be large differences between the optimal policies for different tasks. This forces the agent to find task-specific strategies for exploring each MDP efficiently. Therefore, our proposed environment contains a wider task distribution, with significant differences between tasks. Additionally, each task can be seen as a distribution itself, since a task depends on the humans' behavior, which varies. This hierarchical approach to defining the tasks is similar to Meta-World [16].

IV. RESULTS

We will now present a preliminary experiment where we attempt to solve the proposed environment using RL. We use PEARL [28] to solve the tasks. For simplicity, these experiments were performed on a single type of task, and the task distribution required for meta-RL was created by varying the human behavior. The main goals of these initial results are to show that RL can solve the environment and to highlight the importance of having an adaptable agent.

The improved sample efficiency meta-RL has over standard RL only holds for adapting to new tasks during meta-testing. During meta-training, meta-RL still requires a large number of samples. As a consequence, in these initial experiments, we used simulated humans and a simulated robot. Varied human behavior is still a central part of the proposed task collection. Therefore, whenever a task is created, the simulated human h_1's probabilities of making different types of errors are sampled from a multivariate normal distribution.

Before PEARL can be applied to the proposed environment, some slight modifications have to be made. PEARL uses Soft Actor-Critic (SAC) [33], which is a sample-efficient off-policy RL algorithm. SAC assumes a continuous action space, so the first modification is to extend SAC to the discrete action space [34] used by our proposed environment. The second change is related to the temperature parameter, which controls the exploration-exploitation trade-off during training. PEARL uses a version of SAC where the temperature hyperparameter is manually set. The temperature has a significant effect on training, but finding the optimal value can be difficult. Therefore, we decided to use the improved version of SAC, where the temperature is set automatically, as presented in [35].

Our initial results are shown in Fig. 3. Training performance is measured by meta-testing the agent on the same tasks used for meta-training. Validation performance is measured by meta-testing the agent on tasks it has never seen before. While the training and evaluation tasks are sampled from the same distribution, we decided to only sample evaluation tasks from a region of the distribution that was intentionally left out when sampling the training tasks. This ensures that high performance is the result of an agent that can generalize to new tasks, and not a consequence of the similarity between training and validation tasks. The results are reported as the average return of meta-testing on all tasks. The performance on each task is the average return of meta-testing the agent on several episodes from that task. Running several episodes on each task allows the agent to adapt.

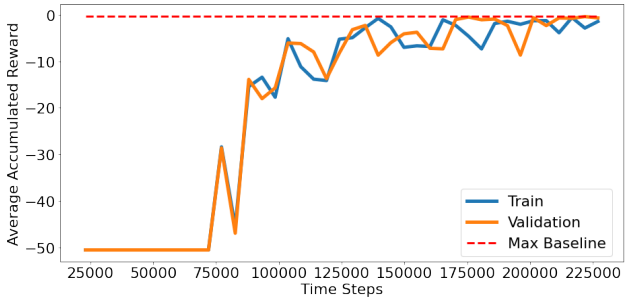

Fig. 3. The agent's validation performance is close to the theoretical maximum baseline, showing that our environment is learnable

For comparison, we provide a baseline that records the maximum reward that can be obtained when both the RL agent and the humans behave optimally. Since the RL agent has only limited control over the humans' behavior, the optimal agent will not always reach this baseline. Results show that the agent can learn the goals of the task. Moreover, the agent's performance on new tasks is similar to its performance on tasks encountered during training, which can be an early sign of the agent's ability to adapt.

The results presented are promising. However, it is important to remember that they are not sufficient to claim that the agent can adapt to new tasks. More work has to be done to define the parameters of an environment where adapting is the only way to be optimal. If such a carefully-crafted environment is used during meta-training, then during meta-testing, the trained agent should be able to adapt to tasks that contain any new human, simulated or real.

V. CONCLUSIONS

We have presented an HRI environment focused on indirect interactions, and preliminary results on using RL to solve this environment. A current limitation of our work is that training is done in a simulation. Extending this to real tasks, with real humans, is an important next step. Fortunately, meta-learning can be used for sim-to-real transfer [36]. Performing experiments with real humans also means we can measure the effects indirect interactions have on the participants' trust and WB. Finally, we want to use meta-RL to solve our proposed environment and create an RL agent that can adapt to new tasks and humans. This will likely only be possible if we also provide an improvement to current meta-RL algorithms.

REFERENCES

[1] J. Schmidtler, V. Knott, C. Hölzel, and K. Bengler, "Human centered assistance applications for the working environment of the future," *Occupational Ergonomics*, vol. 12, no. 3, pp. 83–95, 2015.

[2] Z. R. Khavas, S. R. Ahmadzadeh, and P. Robinette, "Modeling trust in human-robot interaction: A survey," in *International Conference on Social Robotics*. Springer, 2020, pp. 529–541.

[3] P. A. Hancock, T. T. Kessler, A. D. Kaplan, J. C. Brill, and J. L. Szalma, "Evolving trust in robots: specification through sequential and comparative meta-analyses," *Human factors*, vol. 63, no. 7, pp. 1196–1229, 2021.

[4] T. Sander, "Positive computing," in *Positive psychology as social change*. Springer, 2011, pp. 309–326.

[5] R. A. Calvo and D. Peters, "Promoting psychological wellbeing: loftier goals for new technologies [opinion]," *IEEE Technology and Society Magazine*, vol. 32, no. 4, pp. 19–21, 2013.

[6] A. Gaggioli, P. Cipresso, S. Serino, D. M. Campanaro, F. Pallavicini, B. K. Wiederhold, and G. Riva, "Positive technology: a free mobile platform for the self-management of psychological stress," *Annual Review of Cybertherapy and Telemedicine*, vol. 199, pp. 25–29, 2014.

[7] C. Botella, A. Mira, I. Moragrega, A. García-Palacios, J. Breton-Lopez, D. Castilla, A. R. L. Del Amo, C. Soler, G. Molinari, S. Quero, *et al.*, "An internet-based program for depression using activity and physiological sensors: efficacy, expectations, satisfaction, and ease of use," *Neuropsychiatric Disease and Treatment*, vol. 12, p. 393, 2016.

[8] R. M. Baños, E. Etchemendy, D. Castilla, A. García-Palacios, S. Quero, and C. Botella, "Positive mood induction procedures for virtual environments designed for elderly people," *Interacting with Computers*, vol. 24, no. 3, pp. 131–138, 2012.

[9] A. Steptoe, A. Deaton, and A. A. Stone, "Subjective wellbeing, health, and ageing," *The Lancet*, vol. 385, no. 9968, pp. 640–648, 2015.

[10] C. D. Fisher, "Conceptualizing and measuring wellbeing at work." 2014.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[12] N. Akalin and A. Loutfi, "Reinforcement learning approaches in social robotics," *Sensors*, vol. 21, no. 4, p. 1292, 2021.

[13] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[14] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester, "Challenges of real-world reinforcement learning: definitions, benchmarks and analysis," *Machine Learning*, vol. 110, no. 9, pp. 2419–2468, 2021.

[15] M. Thabet, M. Patacchiola, and A. Cangelosi, "Sample-efficient deep reinforcement learning with imaginary rollouts for human-robot interaction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5079–5085.

[16] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*. PMLR, 2020, pp. 1094–1100.

[17] Y. Gao, E. Sibirtseva, G. Castellano, and D. Kragic, "Fast adaptation with meta-reinforcement learning for trust modelling in human-robot interaction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 305–312.

[18] A. Ballou, C. Reinke, and X. Alameda-Pineda, "Variational meta reinforcement learning for social robotics," *arXiv preprint arXiv:2206.03211*, 2022.

[19] S. Vinanzi, A. Cangelosi, and C. Goerick, "The collaborative mind: intention reading and trust in human-robot interaction," *Iscience*, vol. 24, no. 2, p. 102130, 2021.

[20] T. Sanders, A. Kaplan, R. Koch, M. Schwartz, and P. A. Hancock, "The relationship between trust and use choice in human-robot interaction," *Human factors*, vol. 61, no. 4, pp. 614–626, 2019.

[21] A. Xu and G. Dudek, "Towards modeling real-time trust in asymmetric human–robot collaborations," in *Robotics Research*. Springer, 2016, pp. 113–129.

[22] A. Colim, R. Morgado, P. Carneiro, N. Costa, C. Faria, N. Sousa, L. A. Rocha, and P. Arezes, "Lean manufacturing and ergonomics integration: Defining productivity and wellbeing indicators in a human–robot workstation," *Sustainability*, vol. 13, no. 4, p. 1931, 2021.

[23] C. Y. C. Chang, M. Díaz, and C. Angulo, "The impact of introducing therapeutic robots in hospital's organization," in *International Workshop on Ambient Assisted Living*. Springer, 2012, pp. 312–315.

[24] V. Sauer, A. Mertens, J. Heitland, and V. Nitsch, "Exploring the concept of passenger well-being in the context of automated driving," *International Journal of Human Factors and Ergonomics*, vol. 6, no. 3, pp. 227–248, 2019.

[25] H. Claure, Y. Chen, J. Modi, M. Jung, and S. Nikolaidis, "Multi-armed bandits with fairness constraints for distributing resources to human teammates," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 299–308.

[26] M. El-Shamouty, X. Wu, S. Yang, M. Albus, and M. F. Huber, "Towards safe human-robot collaboration using deep reinforcement learning," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4899–4905.

[27] S. Thrun and L. Pratt, "Learning to learn: Introduction and overview," in *Learning to learn*. Springer, 1998, pp. 3–17.

[28] K. Rakelly, A. Zhou, C. Finn, S. Levine, and D. Quillen, "Efficient off-policy meta-reinforcement learning via probabilistic context variables," in *International conference on machine learning*. PMLR, 2019, pp. 5331–5340.

[29] L. Zintgraf, S. Schulze, C. Lu, L. Feng, M. Igl, K. Shiarlis, Y. Gal, K. Hofmann, and S. Whiteson, "Varibad: Variational bayes-adaptive deep rl via meta-learning," *Journal of Machine Learning Research*, vol. 22, no. 289, pp. 1–39, 2021.

[30] T. Imagawa, T. Hiraoka, and Y. Tsuruoka, "Off-policy meta-reinforcement learning with belief-based task inference," *IEEE Access*, vol. 10, pp. 49 494–49 507, 2022.

[31] K. Arndt, M. Hazara, A. Ghadirzadeh, and V. Kyrki, "Meta reinforcement learning for sim-to-real domain adaptation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2725–2731.

[32] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[33] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[34] P. Christodoulou, "Soft actor-critic for discrete action settings," *arXiv preprint arXiv:1910.07207*, 2019.

[35] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, *et al.*, "Soft actor-critic algorithms and applications," *arXiv preprint arXiv:1812.05905*, 2018.

[36] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 737–744.