Personalised Interactive Reinforcement Learning with Multi-Task Pre-training

TARAKLI, Imene and DI NUOVO, Alessandro <http://orcid.org/0000-0003-2677-2650>

Personalised Interactive Reinforcement Learning with Multi-Task Pre-training

Imene Tarakli[1] and Alessandro Di Nuovo[1]

Sheffield Hallam University, Sheffield, UK,
i.tarakli@shu.ac.uk

Abstract. Personalised robots have immense potential to enhance daily life through tailored interactions, yet achieving efficient personalisation remains challenging. This paper introduces a Multi-task Interactive Reinforcement Learning (MIRL) framework aimed at improving the efficiency of interactive learning with evaluative feedback. We demonstrate that pre-training the robot across diverse tasks significantly reduces the learning steps required during fine-tuning, thereby enhancing sample efficiency. Our approach effectively aligns robot behaviours with user preferences, as evidenced by experimental results. These advancements promise to advance the usability and effectiveness of personalised robotics in diverse applications.

Keywords: Interactive Reinforcement Learning, Pre-training, Personalisation

1 Introduction

Personal robots have emerged as a highly promising technology for performing interactive tasks in a variety of settings, including domestic, healthcare, and public environments [17] [6]. These robots have the potential to significantly enhance quality of life by assisting with daily activities, providing companionship, and supporting individuals with special needs. A critical factor for the widespread adoption and acceptance of these robots is their ability to achieve a high level of personalisation, which allows them to be tailored to the specific needs, preferences, and routines of individual users. Personalisation is essential for making interactions with robots more natural and effective, ultimately leading to higher user satisfaction and engagement [12] [5].

Achieving such a degree of personalisation, however, presents several challenges. One major challenge is enabling robots to learn from each user without requiring the users to have any technical knowledge or expertise in programming or robotics. This means the learning process must be intuitive and accessible to all users, regardless of their background. Another challenge is the speed at which the robot can learn; it is important that the robot adapts its behaviour to align with the user's intent swiftly. This is not only beneficial for the user's experience, making the robot more immediately useful, but it also reduces the burden on the user to continuously teach the robot, thereby lowering the overall effort

required for personalisation. One effective method for enabling users to teach robots is Interactive Reinforcement Learning (RL). In Interactive RL, the robot learns through a process of trial and error, guided by feedback from the user [11]. The user observes the robot's actions and provides evaluative feedback, such as approval or disapproval, based on the correctness and desirability of the actions taken [16]. This feedback is then used to adjust and improve the robot's policy, helping it to learn behaviours that align with the user's preferences and goals. Interactive RL has been successfully applied in various real-world scenarios, with a potential to yield positive user experiences and favourable perception of the robot [15].

However, Interactive RL approach requires a substantial amount of effort from users, particularly during the initial stages of learning [3]. At the beginning, the robot lacks any contextual understanding of the task and acts randomly. It learns primarily through the user's feedback, which means the user must provide constant guidance and correction. This extensive teaching process can be time-consuming and demanding, making it less feasible for users who may not have the time or patience to engage in prolonged training sessions. Consequently, while Interactive RL holds promise, its practicality and scalability for widespread use in social robots remain limited by the high initial effort required from users.

To address the feedback efficiency of interactive RL models, previous studies have explored the incorporation of expert demonstrations provided by users at the initial stages of the learning process [9]. These demonstrations allow the robot's policy to be initialized with some task-specific information, thereby preventing the robot from starting with completely random actions. Users can provide demonstrations in two primary ways: through teleoperation or kinesthetic teaching. In teleoperation, the user takes control of the robot remotely, guiding it to perform the desired tasks. This method can be effective as it allows the user to demonstrate complex actions precisely, however, it requires the user to have a certain level of technical skill to control the robot accurately without causing harm to the robot or the surroundings [10]. Kinesthetic teaching, on the other hand, involves the user physically manipulating the robot's joints to move them into the desired positions. This hands-on approach enables the robot to learn through direct physical interaction, making it easier to convey the specific movements required for a task, however, the physical interaction with the robot might be challenging for some users without the necessary strength, dexterity, or familiarity with the robot's mechanics [2].

In this study, we investigate a novel pre-training method aimed at enhancing the efficiency of Interactive Reinforcement Learning (RL) from human evalautive feedback. Specifically, our approach focuses on multi-task pre-training, wherein the robot is initially trained to learn a general policy across a subset of predefined environments. This initial training phase is designed to provide the robot with a foundational understanding of various tasks, allowing it to develop an adaptable and robust policy. By leveraging diverse tasks during pre-training, we hypothesise that the robot can acquire transferable skills and knowledge that can be easily personalized to new, unseen environments. While multi-task pre-

training has been explored in previous studies, particularly in the context of preference-based learning [7] [14], our objective is to assess its effectiveness in improving the sample efficiency of Interactive RL through evaluative feedback. In our approach, users provide feedback to the robot to fine-tune the pre-trained policy, ensuring it aligns with their specific preferences and requirements without the need to start from scratch. We hypothesise that this feedback mechanism enables the robot to quickly adapt to individual user needs, enhancing both the learning speed and the overall user experience.

1.1 Preliminaries

Reinforcement Learning (RL) aims to solve tasks that are modeled as Markov Decision Processes (MDPs) [13]. An MDP is defined by the tuple $< S, A, T, R, \gamma >$, where S represents the set of states and A represents the set of actions. The transition function $T : S \times A \rightarrow S$ specifies the probability of moving to a new state given the current state and action. The reward function $R : S \times A \rightarrow R$ assigns a reward for performing an action in a given state, and $\gamma \in [0, 1]$ is the discount factor that determines the importance of future rewards.

The objective in RL is to find policies $\pi : S \rightarrow A$ that maximize the total expected discounted rewards over time. The expected return from each state-action pair, known as the action-value and denoted by $Q(s, a)$, is given by $Q(s, a) = E_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s, a)\right]$. Optimal policies, denoted by π^*, are those that maximize these returns, directing the agent to achieve the highest possible rewards.

2 The MIRL Framework

To address the challenge of sample efficiency in existing interactive RL methods, we propose the Multi-task Interactive Reinforcement Learning (MIRL) framework. The framework consists of two phases, as illustrated in figure 1:

1. **Multi-task pretraining:** the robot is first pre-trained on a diverse set of tasks with pre-define reward functions to acquire a generalised policy.
2. **Fine-tuning with evaluative feedback:** The users provide feedback to personalise the behaviour of the robot, acquired in the previous stage, in new, unseen environments.

2.1 Multi-task pre-training:

The initial phase of our model involves multi-task pre-training, which is designed to provide the robot with a comprehensive representation across a variety of tasks. During this phase, the robot is exposed to a diverse set of pre-defined environments, each representing a different task within the same domain. These tasks are selected to cover a broad spectrum of scenarios that the robot might encounter in real-world applications. The goal is to enable the robot to develop

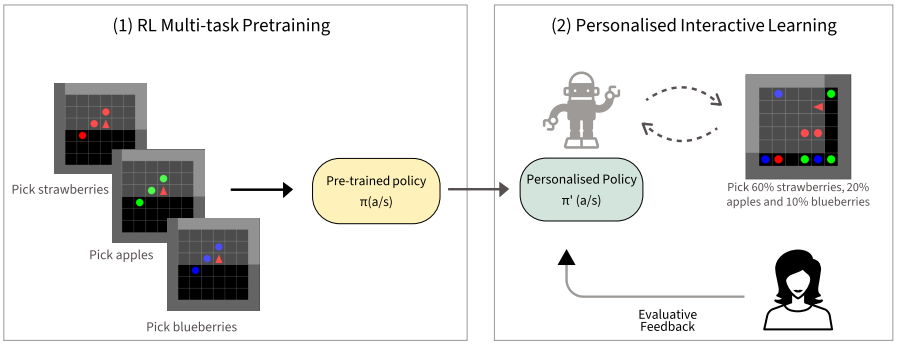


Fig. 1: Illustration of the framework. (1) The robot first learns a basic policy by performing a multi-task RL pretraining on different environments. (2) A user then personalises the robot's behaviour on a new task through evaluative feedback.

a unified generalised policy, Q, that aims to maximise the long-term returns of each environment. For each episode, the robot is randomly presented to a task from among the predefined environments, allowing it to learn generalisable skills across tasks.

2.2 Evaluative Feedback Finetuning

In this phase, the robot is introduced to new, unseen environments where it needs to adapt its behaviour to align with the preferences and needs of individual users. This fine-tuning is achieved through evaluative feedback provided by the users. Specifically, we use the TAMER framework [8], where users provide their binary feedback to assess the correctness of the action taken by the robot. The feedback is then used to update the policy of the robot until it converges to a behaviour aligned with the human's intent. By initialising the robot's behaviour with a generalised policy obtained from the pre-training phase, the robot should quickly adapt to new tasks, requiring less load on the user.

3 Experiments

3.1 Experimental setup

We structure the methodology to address the following research question: Does multi-task pre-training accelerate learning from human feedback?

Environment We consider the Fruit-Picking environment [1] [4], which consists of a gridworld containing different types of fruits that the robot needs to collect. Similar to previous work [1], during the multi-task pre-training phase, the robot

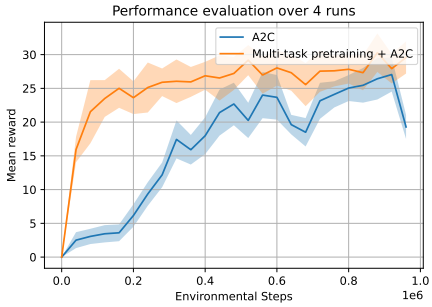
learns to pick one type of fruit per task. This is represented with a reward function that provides a positive reward if the robot picks the correct type of fruit and a negative reward otherwise. In the personalisation phase, users teach the agent a different task within the same domain. Specifically, we consider a task where the robot must pick different types of fruits according to user preferences (e.g., 60% preference for strawberries, 30% preference for apples, and 10% preference for blueberries).

Implementation details We train the agent using a deep Advantage Actor-Critic (A2C) architecture. Since the goal of this paper is to demonstrate the feasibility of using multi-task pre-training to accelerate learning, we simulate human feedback using a reward function that represents the intended fruit distribution. For the aforementioned distribution, the reward function is as follows:

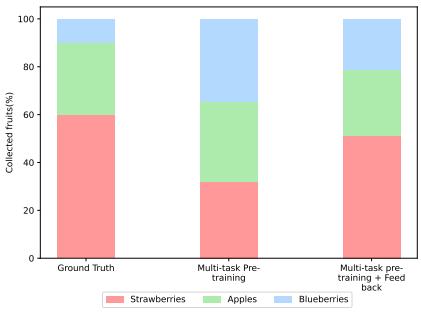
$$R(s,a) = \begin{cases} 0.6 & \text{if } a = \text{pick strawberry} \\ 0.3 & \text{if } a = \text{pick apple} \\ 0.1 & \text{if } a = \text{pick blueberry} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Baseline To evaluate the effectiveness of our model, we compare it against an A2C agent without pre-training. This comparison helps to isolate the impact of multi-task pre-training on learning efficiency and performance.

3.2 Experimental results



(a) Learning curve of A2C agent with and without pre-training on the Fruit Picker task.



(b) Mean Preference Distribution of the collected fruits on 1000 episodes of the Fruit Picker task.

Fig. 2: Evaluation of the framework on a new unseen environment.

Training Performance Comparison Figure 2a illustrates the training results of an A2C agent in the Fruit-Picking environment, comparing the performance with and without multi-task RL pre-training. Our approach (represented by the orange line) achieves comparable performance to the baseline (blue line) using 8 times fewer environmental steps. This highlights the effectiveness of multi-task pre-training in significantly enhancing the sample and feedback efficiency of interactive learning with evaluative feedback.

Preferences distribution Figure 2b shows the preference distribution achieved after training the robot to collect fruits based on user preferences. Initially, during the multi-task pre-training phase, the robot demonstrates an equal collection rate across all types of fruits. Subsequently, after training with simulated feedback, the robot learns a policy resulting in a collection distribution of 51.4% strawberries, 27.3% apples, and 21.3% blueberries. This distribution closely approximates the intended distribution of 60% strawberries, 30% apples, and 10% blueberries.

4 Discussion & Conclusion

The results presented in this study demonstrate the efficacy of multi-task pre-training in enhancing the efficiency of Interactive Reinforcement Learning (RL) with evaluative feedback within the MIRL framework. Our approach achieved comparable performance to the baseline A2C agent while requiring significantly fewer environmental steps (8 times fewer). This substantial reduction in training steps highlights the capability of multi-task pre-training to accelerate the learning process, making it more sample-efficient. Multi-task pre-training enabled the robot to develop a generalised policy across a diverse set of tasks during the initial phase. This foundational knowledge provided the robot with a broader understanding of the environment and tasks, facilitating quicker adaptation to new, unseen tasks during the fine-tuning phase. By initialising the robot with a general policy, derived from multi-task pre-training, we minimised the initial learning curve and cognitive load on users. This approach is particularly advantageous in applications where users may have limited time or expertise to guide the robot extensively. Moreover, we showcased that MIRL had the ability to align with the user's intent. While a the robot first acquired a general policy that allowed it to understand the dynamic of the environment, the subsequent fine-tuning phase, allowed the robot to adjust its behaviour to achieve a collection distribution closely matching the intended user preferences. This adaptation illustrates the effectiveness of interactive RL with evaluative feedback in personalised robot behaviour, allowing to adapt to the user needs without technical programming. The findings of this study have several implications for the development and deployment of personal robots in real-world settings. By enhancing the sample efficiency and aligning with user preferences, our approach improves the overall usability of interactive robots. This advancement could lead to more seamless integration of robots into everyday environments, such as homes or

healthcare facilities, where personalised assistance and interaction are crucial. Future research directions could explore further optimisations and extensions of the multi-task pre-training framework. Investigating the scalability of this approach to more complex environments and tasks, as well as incorporating real-time user feedback mechanisms, could enhance its applicability and adaptability. Additionally, studying the impact of task opposition, where the final task may require behaviours conflicting with those learned during pre-training, would deepen understanding of adaptation dynamics.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955778. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

1. Marwa Abdulhai, Natasha Jaques, and Sergey Levine. Basis for intentions: Efficient inverse reinforcement learning using past experience. *arXiv preprint arXiv:2208.04919*, 2022.
2. Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.
3. Mohamed Chetouani. Interactive robot learning: an overview. *ECCAI Advanced Course on Artificial Intelligence*, pages 140–172, 2021.
4. Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. Babyai: A platform to study the sample efficiency of grounded language learning. *arXiv preprint arXiv:1810.08272*, 2018.
5. Nikhil Churamani, Paul Anton, Marc Brügger, Erik Fließwasser, Thomas Hummel, Julius Mayer, Waleed Mustafa, Hwei Geok Ng, Thi Linh Chi Nguyen, Quan Nguyen, et al. The impact of personalisation on human-robot interaction in learning scenarios. In *Proceedings of the 5th international conference on human agent interaction*, pages 171–180, 2017.
6. Daniela Conti, Grazia Trubia, Serafino Buono, Santo Di Nuovo, and Alessandro Di Nuovo. An empirical study on integrating a small humanoid robot to support the therapy of children with autism spectrum disorder and intellectual disability. *Interaction Studies*, 22(2):177–211, 2021.
7. Donald Joseph Hejna III and Dorsa Sadigh. Few-shot preference learning for human-in-the-loop rl. In *Conference on Robot Learning*, pages 2014–2025. PMLR, 2023.
8. W Bradley Knox and Peter Stone. Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance. *Artificial Intelligence*, 225:24–50, 2015.

9. Guangliang Li, Bo He, Randy Gomez, and Keisuke Nakamura. Interactive reinforcement learning from demonstration and human evaluative feedback. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1156–1162. IEEE, 2018.
10. MD Moniruzzaman, Alexander Rassau, Douglas Chai, and Syed Mohammed Shamsul Islam. Teleoperation methods and enhancement techniques for mobile robots: A comprehensive survey. *Robotics and Autonomous Systems*, 150:103973, 2022.
11. Anis Najar and Mohamed Chetouani. Reinforcement learning with human advice: a survey. *Frontiers in Robotics and AI*, 8:584075, 2021.
12. Michał Stolarz, Alex Mitrevski, Mohammad Wasil, and Paul G Plöger. Learning-based personalisation of robot behaviour for robot-assisted therapy. *Frontiers in Robotics and AI*, 11:1352152, 2024.
13. Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
14. Adrien Ali Taiga, Rishabh Agarwal, Jesse Farebrother, Aaron Courville, and Marc G Bellemare. Investigating multi-task pretraining and generalization in reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.
15. Imene Tarakli, Georgios Angelopoulos, Mehdi Hellou, Camille Vindolet, Boris Abramovic, Rocco Limongelli, Dimitri Lacroix, Andrea Bertolini, Silvia Rossi, Alessandro Di Nuovo, et al. Social robots personalisation: At the crossroads between engineering and humanities (concatenate). In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 920–922, 2023.
16. Andrea Lockerd Thomaz, Cynthia Breazeal, et al. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai*, volume 6, pages 1000–1005. Boston, MA, 2006.
17. Roberto Vagnetti, Nicola Camp, Matthew Story, Khaoula Ait-Belaid, Joshua Bamforth, Massimiliano Zecca, Alessandro Di Nuovo, Suvo Mitra, and Daniele Magistro. Robot companions and sensors for better living: defining needs to empower low socio-economic older adults at home. In *International Conference on Social Robotics*, pages 373–383. Springer, 2023.