Interactive Reinforcement Learning for Navigation Task in Service Robots

Ahmed Elbary
Department of Computer Science
University of Lincoln
Lincoln, United Kingdom
29385647@students.lincoln.ac.uk

***Abstract*— This paper investigates the role of feedback timing in interactive reinforcement learning (IRL) by comparing two agents' performance one receiving post-action evaluative feedback (IRL), and the other following the TAMER framework with prior approval. The task simulates a restaurant delivery robot operating in a constrained 4×4 grid-world restaurant environment, where the robot must learn to navigate from the kitchen to a designated customer table while avoiding incorrect table and obstacles. Both agents are trained using tabular Q-learning and required a human-like scalar feedback, while the performance is evaluated over multiple episodes using the total reward, number of steps, and the success rate. The results show that the TAMER agent achieves faster convergence, greater stability, and higher efficiency, especially in early training, while the IRL agent demonstrates less consistent learning due to delayed corrective signals. These findings highlight the significance of feedback timing in human feedback learning systems.**

Keywords*— *Reinforcement Learning, Interactive Learning, TAMER, Learning from Human, Service Robotics

I. INTRODUCTION

Human robot interaction (HRI) is a major application in the modern autonomous systems, especially in the applications where robots operate in a shared spaces with humans or are required to align with human goals and expectations. In such applications, the learning process must not only be efficient but also safe, and responsive to human feedback. Traditional reinforcement learning methods while they are successful in simulation environments, often face limitations in these human contexts due to their trial-and-error nature.

To bridge this gap, the interactive learning strategies between the robots and humans have been explored which let the robots to get benefit from a real time human input in the training process. In contrast with the offline supervised learning or fixed policy optimization, the interactive learning frameworks allow the agents to adjust their behaviour based on the human feedback to each move. This application is really close with how humans teach each other or teachers in schools teach young students through feedback, approval, and corrections on the robot itself and presents a practical way for developing robots that are autonomous and flexible to new users and tasks.

In service environments such as restaurants, hospitals, or retail stores, the delivery robots are expected to navigate dynamic layouts, to avoid obstacles, and to interact properly with people. However, designing a comprehensive reward function that implement all the aspects of correctness can be highly complex. For instance, delivering the items to the wrong table in the restaurants may be technically successful in terms of position tracking, but incorrect in the task outcome. Similarly, taking a shorter path that shortcuts through high restricted areas or customer zones may violate the social rules. In such scenarios, the learning from human feedback can help the agents to distinguish between a functionally acceptable and the outcome appropriate behaviour.

The task addressed in this study simulates such a challenge using a delivery robot operating in a constrained indoor restaurant. Using a grid-world environment the robot is trained to start the journey from a fixed starting point to a designated target representing a customer table while avoiding incorrect tables and navigational hazards. The learning process is informed by the human feedback provided during the robot's decision-making loop.

II. LITERATURE REVIEW

Recent studies in interactive learning systems showed the need for balancing the human effort with the learning performance in general. When human provided feedback can for sure accelerate the learning process, as such, the researchers have explored some strategies to minimize the reliance on continuous human feedback while also maintaining a learning efficiency. One method involves leveraging human feedback to initialize the reward function early in training, followed by autonomous fine-tuning through the reinforcement learning. This hybrid approach allows the agents to move from being depended on human knowledge only while gradually increasing autonomy.

One of the most contributions in this area is the TAMER framework that was proposed by Knox and Stone [1], which allows the agents to learn optimal behaviour by taking into account the human provided evaluative feedback instead of relying on the basic environmental rewards. The agent learns a reward function that tries to approximate the human preferences and then uses it to guide the next action selection. This framework showed that a real-time human input could help for a sufficient and more effective learning signal in the environmental rewards, by letting the foundation for more exploration into human feedback during the learning process.

A wider perspective on IRL is provided in the well-known survey by Arumugam et al. [2], where in this this survey it categorizes the IRL systems by feedback type, timing, frequency, and the integration strategy, it also mentioned that the timing whether feedback occurs before or after an action works as a critical role in shaping the learning dynamics. The systems that run the feedback prior to the execution can prevent unsafe or unwanted actions, while those using post action rating feedback method enable the policy correction.

Another major contribution relevant to the real time training is the study by Duret-Lutz et al. [3], which investigates the application of experience replay to reinforcement learning under a real-time constraints.

Experience replay involves a storing and the reuse of the prior transitions to improve the efficiency and the learning stability. This study demonstrates that experience replay can be effectively used in interactive learning contexts, where the human feedback is limited or happens in a long time period between each feedback and the other which helps to maintain learning without requiring constant supervision.

Recent studies also presented the importance of balancing human effort compared with learning performance, where human provided feedback can accelerate learning. One approach to reducing the cognitive cost involves using prior human feedback to shape the reward function, which allow the agent to refine its policy autonomously in later episodes. The concept of shaping policies based on the prior action was further explored through methods such as policy shaping [4] in which human feedback is used to adjust the agent's policy rather than its reward function.

The temporal context in which the feedback is given also affects the learning efficiency. For instance, the COACH algorithm proposed by MacGlashan et al. [5] treats the human feedback as a policy gradient value, which enables the convergence in real-time learning scenarios. This and similar approaches have been applied to tasks such as robotic navigation, manipulation, and social interaction, where learning from real time human feedback improves the alignment with task requirements and user preferences.

As robotic agents are deployed in shared human environments, the ability to adapt them based on feedback is important. Studies in service robotics suggest that real-time evaluative feedback allows robots to learn the social compliant behaviour which cannot be easily done through static reward functions. Despite these challenges remain in feedback consistency, the scalability to a larger state spaces, and effective user interface design for human trainers. The IRL systems must continue to explore adaptive feedback strategies and efficient policy learning mechanisms in application that balance the automation with human supervision.

III. METHODOLOGY

This section outlines the experimental setup, the learning models, and evaluation strategy used to investigate the impact of human feedback on reinforcement learning in a simplified delivery problem. A grid-world environment is designed to simulate a restaurant scenario, and two interactive agents: one based on a post action evaluation and another agent is the TAMER framework are implemented.

A. Environment and Task Setup

The learning environment for this study consists of a 4×4 grid world that simulates a basic layout of a restaurant, where this abstraction is commonly used in reinforcement learning research due to its suitability for experimentation. In the simulated scenario, a delivery robot is required to navigate from a fixed starting location which is the kitchen in this case, to a predefined target which is the correct customer table. The layout also includes an incorrect table as a distractor, also two static obstacles that the robot must avoid during navigation.

The starting position is located at the bottom left corner of the grid, while the correct delivery table is positioned at the top right corner. The incorrect table is placed on the same column but in a different row, requiring the agent to distinguish between the correct table and the wrong table to avoid losing rewards. The obstacles are placed to introduce additional navigational complexity in the middle of the grid, where each action by the robot consists of a single move in one of four directions: up, down, left, or right. The actions that would cause the agent to move off the grid or into the obstacle results in no change of position. A successful delivery to the correct table leads to a reward of +1, while reaching the wrong table results in a penalty of –1. This design reflects a real-world challenge in indoor delivery robotics, where the agents must reach a specific destination, avoid obstacles, and to learn from human feedback.

B. Interactive Learning Approaches (IRL)

To investigate the influence of the feedback timing on learning performance, two interactive learning agents were developed and evaluated: a traditional Interactive Reinforcement Learning (IRL) agent and a TAMER style agent. Both agents share the same task and the learning objectives, but they differ in the timing and mechanism by which human feedback is requested.

The IRL agent operates in a post action feedback paradigm system after executing each action, the agent queries a human trainer for an evaluative point on the resulting state. The feedback can be positive, negative, or neutral, and is represented as a scalar shaping reward that affect the environmental signal. For example, a "good" evaluation might contribute an additional +0.2 to the total reward, whereas a "bad" rating subtracts a similar value, and finally the neutral has no impact on the reward value. This model simulates a reactive feedback loop, where the human observes the agent's behaviour and adjusts its learning based on the outcomes rather than intentions.

In contrast, the TAMER style agent asked the human input before executing an action at each decision point, the agent proposes a move and requests approval from the human. If the human rejects the proposed action, then the agent would generate an alternative suggestion until an acceptable action is found, or a rejection limit is reached. After executing the approved action, the resulting state is still subject to be evaluated by the human again allowing the agent to learn from the prior signal and the feedback from human. This dual feedback structure aligns closely with the TAMER framework proposed in earlier literature, where the human trainers guide agents in a real time by shaping their decision-making process through direct interaction.

The distinction between these two agents lies in the timing and point of human involvement. The IRL agent learns from observation-driven evaluation, while the TAMER agent incorporates prior guidance and correction. Comparing these agents allows for a direct analysis of how early versus late feedback affects policy quality, convergence speed, and behavioural outcomes.

C. Learning Algorithm and Parameter Settings

Both agents employ tabular Q-learning as their underlying learning algorithm. This method maintains a table of Q-values representing the expected utility of performing each action in each state. After each action, the Q-value is updated using the Bellman equation, combining the received reward with the estimated value of the next state. The update rule is defined as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)]$$

In this equation, $Q(st, at)$ is the current Q-value for state action pair (s_t, a_t), where α is the learning rate, γ is the discount factor, and r_t is the reward at the time t. The reward itself consists of two factors: one from the environment and one from human feedback as follows:

$$r_t = r_t^{env} + r_t^{human}$$

This formulation let the agents to integrate the human guidance directly into the learning process in a structured and mathematically way.

The learning rate α is set to 0.5, while the discount factor γ is set to 0.9. The exploration is handled via an $\epsilon = 0.3$ greedy strategy, which enables the agent to balance between the exploitation of known policies and the exploration of less familiar ones. The tabular structure of Q-learning is suited to this grid world setup because of the state space is small and discrete. It also allows for fast updates and clear visualization of learned policies, serving both the debugging and performance analysis.

D. Simulation Procedure

The training is conducted over multiple episodes, each beginning with the agent in the kitchen position. During each episode the agent interacts with the environment step by step, using the learned Q-values and receiving the feedback from the human according to its designated learning model. The episode concludes when the agent reaches the correct or incorrect table or when a maximum number of steps is exceeded.

Throughout training, the human acts as the source of guidance providing the feedback in real time via a command-line interface. This setup mimics the in-person training scenarios and reflects the realistic constraints where human attention is limited and the feedback must be concise. The interaction between the agent and the human trainer is essential to shape the learned policy and to understand how a different feedback strategies affect learning behaviour.

E. Evaluation Metrics

The performance of each agent is assessed using several quantitative metrics such as the total reward accumulated per episode that reflects how well the agent balances efficient movement with accurate goal. The number of steps taken before reaching a terminal state indicates the learning efficiency and the path quality. Additionally, the frequency of successful deliveries to the correct table is used as a measure of task accuracy. The cumulative reward across all the episodes is also tracked to evaluate the overall convergence and learning stability over time.

These metrics provide a foundation for comparing the IRL and TAMER agents under consistent conditions. By observing how these values are changing throughout the training process, insights can be gained into the relative advantages and limitations of each feedback model in terms of learning speed, the reliability, and the human involvement requirements. These metrics are recorded across multiple episodes and used to generate performance plots for each agent, including the total reward per episode, steps per episode, and the cumulative success rate.

IV. Results and Analysis

To evaluate and compare the performance of the IRL and the TAMER agents, a series of 100 training episodes were conducted in the simulated restaurant delivery environment. The agents were evaluated based on the total reward per episode, the number of steps taken to reach the terminal state, the cumulative success rate, and finally the average performance across all episodes. The results are presented through four figures illustrating learning behaviour.

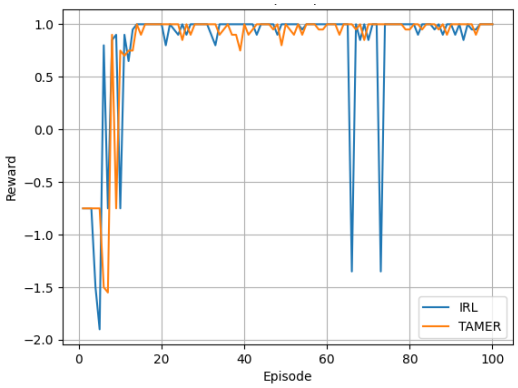

Figure 1 Rewards per episode for IRL and TAMER in 100 episodes

Figure 1 shows the reward obtained per episode for both agents, in the early episodes both IRL and TAMER agents experience unstable performance due to the exploration and the low Q-value confidence. However, the TAMER agent converges more quickly more than the IRL, achieving the rewards close to +1 after approximately the 15th episode. In contrast, the IRL agent suffers from the negative rewards even in the last episodes, suggesting some failures in the task execution. These fluctuations can happen to its reactive to the feedback mechanism which allow actions to be executed before even receiving any correction.

Figure 2 illustrates the number of steps taken per episode where both agents start with inefficient navigation, often reaching the step limit of 15 steps. Over the time, both agents show an improved efficiency, but the TAMER agent demonstrates a more stable system in terms of the fewer number of steps with fewer spikes in step count. In the episode 30 the TAMER consistently completes the task in fewer steps than the IRL, which indicates that the prior feedback helps in selecting more effective paths.

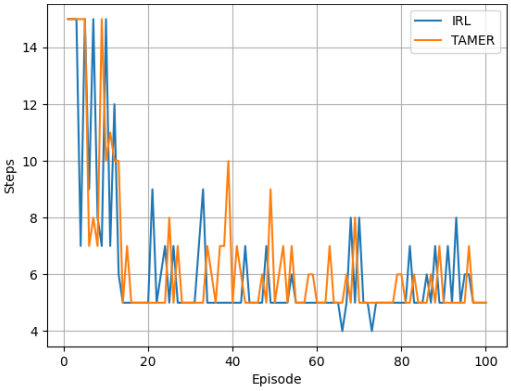

Figure 2 Number of steps per episode for IRL and TAMER agents

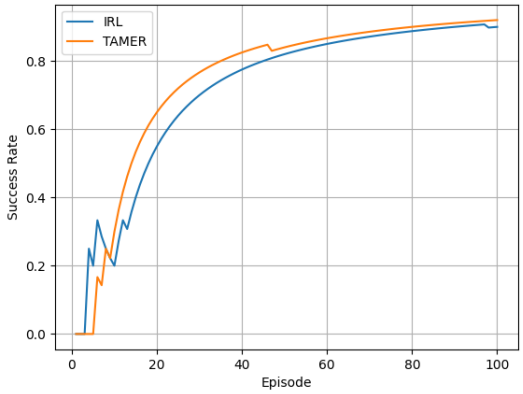

Figure 3 The cumulative success rate across the training

Figure 3 presents the cumulative success rate across the training where the TAMER achieving a success curve reaching around 90% success by the episode 40, while the IRL progresses more gradually through the 100 episode. By the end of training, both agents reach a similar success rate, but TAMER agent achieved this success earlier with fewer errors.

Finally, Figure 4 below compares the two agents' average performance across the three metrics: the reward, steps, and success rate per episodes. The TAMER agent outperforms the IRL in all three categories. Where it achieves a higher average reward and completing the episodes in fewer steps, and scored a slightly higher overall success rate than IRL. These results align with expectations from the literature suggesting that the prior action filtering based on human input in TAMER leads to faster and higher performance policy learning.

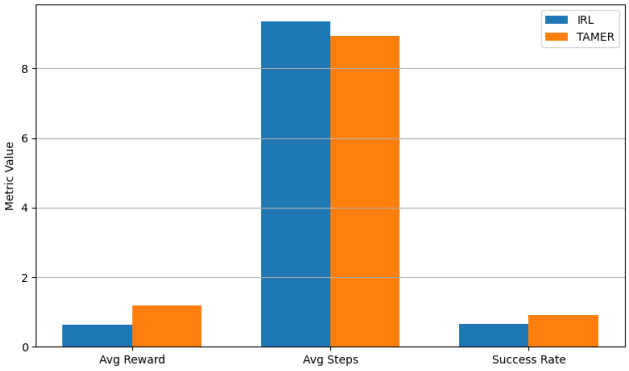

Figure 4 Average performance comparison across three metrics

V. Discussion

The comparative analysis showed several ideas into the impact of the feedback timing on agent learning behaviour, the TAMER agent which has the human-rate approval before doing the actions, consistently outperforms the IRL agent that only receives post-action evaluative feedback. This advantage is especially clear in the early episodes of learning, where the feedback before action help to prevent any incorrect deliveries and encourages for more effective exploration.

Although both agents converge in the end of training, but the TAMER approach shows a smoother and faster convergence across all the metrics. Also, its ability to avoid the unneeded actions early helped to reduce the frequency of the negative rewards and unnecessary steps, leading to more stable Q-value updates. This is more valuable in high stakes domains where any error correction after execution may be costly.

However, the TAMER strategy also implies a greater dependency on the frequent human involvement during the action selection phase. While this is simulated in the current setup, a real implementation in real world application would require more effort from the trainer. In contrast, the IRL agent requires less engagement from the trainer, which may be preferable in some scenarios where the rabidly feedback is impractical.

Overall, the results validate the existing findings in the reinforcement learning literature: the feedback timing affects the learning efficiency, and early guidance mechanisms can accelerate convergence and improve overall the performance. These outcomes support the use of TAMER systems in the environments where the precision and efficiency are critical.

VI. Conclusion

This study evaluated the effect of the feedback timing on learning performance in a grid-world delivery task simulating a service robot scenario. Two agents were implemented: the IRL, which receives feedback after each action, and the TAMER, which incorporates prior evaluation before actions. Both agents were trained using the tabular Q-learning and tested over 100 episodes.

The results demonstrated that the TAMER agent achieved better performance across multiple metrics, including total reward, number of steps, and the success rate. Its feedback structure allowed a faster convergence and fewer policy errors, particularly in the early training. The IRL agent also reached reasonable performance but required more time and was less stable due to its reactive feedback. These findings highlight the importance of feedback structure in interactive learning systems and support the adoption of the prior feedback

Acknowledgement

I extend my deepest gratitude to Dr Pual Baxter for his guidance and support in carrying out this project. Sincerely thank the authors of the referenced papers for their inspiring research.

References

[1] W. B. Knox and P. Stone, “TAMER: Training an agent manually via evaluative reinforcement,” in Proc. IEEE Int. Conf. on Development and Learning, 2008.

[2] R. Arumugam, S. Krening, M. Cakmak, et al., “A Survey on Interactive Reinforcement Learning: Design Principles and Open Challenges,” in Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI), 2019.

[3] M. Duret-Lutz, C. Lefort, and C. Ollion, “Experience Replay for Real-Time Reinforcement Learning Control,” IEEE Access, vol. 9, pp. 79479–79489, 2021.

[4] S. Griffith, K. Subramanian, J. Scholz, C. Isbell, and A. Thomaz, “Policy shaping: Integrating human feedback with reinforcement learning,” in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[5] R. MacGlashan, M. Littman, et al., “Convergent Actor-Critic by Humans (COACH),” in *Proc. Int. Conf. on Intelligent Robots and Systems (IROS)*, 2017, pp. 4020–4027.