DQN-TAMER: Human-in-the-Loop Reinforcement Learning with Intractable Feedback

Riku Arakawa*†, Sosuke Kobayashi †, Yuya Unno †, Yuta Tsuboi †, Shin-ichi Maeda †

***Abstract*— Exploration is a great challenge in reinforcement learning (RL), limiting its applications in robotics. Building a well-learned agent often requires many trials, due to the difficulty of matching its actions with rewards in the distant future. A remedy is to train an agent with real-time feedback from human observers who immediately gives rewards. This study tackles a series of challenges for introducing such a human-in-the-loop RL scheme. We first reformulate human observers: *Binary*, *Delay*, *Stochasticity*, *Unsustainability*, and *Natural Reaction*. We also propose an RL method called DQN-TAMER, which efficiently uses both human feedback and distant task rewards. We find that the DQN-TAMER agent outperforms the baselines in Maze simulated environment even with a limited amount of human feedback. Furthermore, through a real-world human-in-the-loop setting using a car robot with a camera, we demonstrated that natural reactions like facial expressions work well as an implicit human reward. The video attachment is available: `https://youtu.be/o25x51eHf7s`.**

I. Introduction

Reinforcement learning (RL) has potential applications for autonomous robots. However, it often requires a lot of trials until the agent reaches an optimal policy, preventing RL from spreading to real applications. This is primarily because RL agents obtain rewards only in the distant future, e.g., at the end of the task. Thus, it is difficult to propagate the reward back to actions that play a vital part in receiving the reward. Giving additional training signals by humans is a useful remedy. During training, human observers perceive the agent's actions and states and provide some feedback to the agent in real time. Such immediate rewards can accelerate learning and reduce the number of required trials. This method is called *human-in-the-loop* RL and its effectiveness has been reported [1]–[9]. However, experiments in prior studies did not or only partially consider some key factors in realistic human-robot interactions. They sometimes assumed that human observers could (1) give precise numerical rewards, (2) do so without delay (3) at every time step, and (4) that rewards would continue forever.

We first reformulate human observers with the more realistic characteristics. Next, we propose a simple but effective RL algorithm, DQN-TAMER, that leverages two different critic networks to combine task- and human-reward. Finally, we demonstrate its performance through experiments in a simulated and real-world environment using a car robot recognizing human facial expressions as implicit rewards.

* The University of Tokyo.
arakawa-riku428@g.ecc.u-tokyo.ac.jp
† Preferred Networks, Inc.
{sosk,unno,tsuboi,ichi}@preferred.jp

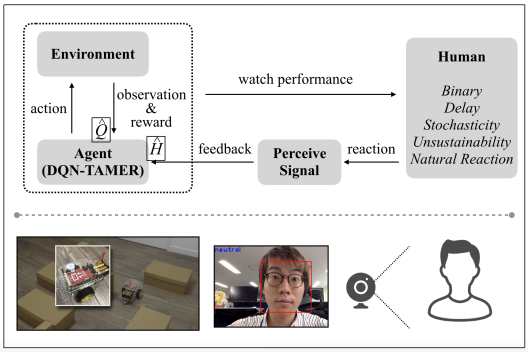

Fig. 1: Human-in-the-loop RL and our model (DQN-TAMER). The agent asynchronously interacts with a human observer in an environment. DQN-TAMER decides actions based on two critics: one ($\hat{Q}$) estimating rewards from the environment and the other ($\hat{H}$) for feedback from the human.

TABLE I: Characteristics of human observers

Study	*Binary*	*Delay*	*Stochastic*	*Unsustainable*	*Natural Reaction*
Thomaz et al. 2005 [1], [2]		✓	✓		
Joost Broekens 2007 [3]	✓			✓	✓ (face)
Knox et al. 2007 [4]	✓	✓	✓		
Tenorio-Gonzalez et al. 2010 [5]		✓	✓		✓ (voice)
Pilarski et al. 2011 [6]	✓	✓	✓		
Griffith et al. 2013 [7]	✓		✓		
MacGlashan et al. 2017 [8]	✓	✓	✓	✓	
Warnell et al. 2018 [9]	✓	✓	✓		
Ours	✓	✓	✓	✓	✓ (face)

II. Problem Formulation

We introduce five key characters to consider about human feedback when we aim to launch human-in-the-loop RL systems. Table I compares prior studies in these axes.

1) *Binary*: Requesting people give fine-grained or continuous scores is found difficult [10] and thus binary feedback is preferred, simply indicating good or bad.
2) *Delay*: Human feedback is usually delayed by a significant amount of time [11] and the delay must not be constant.
3) *Stochasticity*: It is reported that the feedback frequency varies largely among human users [12], [13].
4) *Unsustaibability*: It is very difficult to presume that humans watch an agent until it finishes learning through many episodes. Ideally, even if a human gives feedback within a limited span after learning begins, we wish it could subsequently lead to a better learning process.
5) *Natural Reaction*: When intelligent agents become more ubiquitous and we launch real human robot interaction

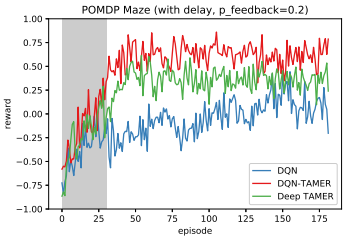

Fig. 2: The result in simulated Maze.

systems, it is preferable that the system infers implicit feedback from natural human reactions rather than what humans provide actively.

III. Method

TAMER [4] is a current standard framework in human-in-the-loop RL, where the agent predicts human feedback and takes the action that is most likely to result in good feedback. We introduce DQN-TAMER, which incorporates task-reward into this framework by using two critic deep neural networks: $\hat{Q}(s,a)$ for task-value function of the state-action pair (s,a) and $\hat{H}(s,a)$ for human-value function. Thus using weight variables α_q and α_h, the optimal policy can be written by

$$\pi(s)_{\mathrm{DQN-TAMER}} = \arg\max_a \alpha_q \hat{Q}(s,a) + \alpha_h \hat{H}(s,a). \quad (1)$$

IV. Experiment

We experimented with DQN-TAMER in a simulated and real-world environment. We compared it against two baseline algorithms including (1) DQN, which leverages reward from task using a neural network, and (2) Deep TAMER [9], which uses reward from the observer using a neural network.

A. Simulated Maze

Maze is a classical game where the agent must reach a predefined goal. We compared the sample efficiency in each algorithm through experiment, i.e., we examined how fast learning converges. We fixed the field size of a maze to 8 $\times$ 8 and the initial distance to the goal at 5. Every step, the agent can move 1 space toward north, east, south, and west. If it reaches the goal, it receives +1.0 but otherwise -0.01 as reward from the task. We simulated a human feedback as it gives a binary label whether the agent reduces the Manhattan distance to the goal, stochastically and with delay. If an agent moves closer to the goal, the human provides +1 positive feedback and -1 negative feedback otherwise.

Figure 2 shows the result of training when the human gave feedback during the first 30 episodes with low frequency (20%) along with stochastic delay. DQN-TAMER outperforms the baselines even after feedback stop. We observed similar results when we varied human parameters.

B. Real-world Maze with Natural Reaction

In a real-world Maze environment, we built a car robot with a camera and trained it with human implicit feedback, *Natural Reaction*. Reward is not directly given and, instead, inferred by recognizing a human facial expression. The intriguing question we tackle here is whether an agent can learn well from such suspicious reward including recognition errors. The agent interprets the facial expression 'happy' as positive reward (+1) and other expressions ('anger', 'contempt', 'disgust', 'fear', and 'sad') as negative (-1). We used MicroExpNet as a recognition model, which is a convolutional neural network-based model [14].

For the details, please watch the video. The facial expression classifier misclassified the facial expressions (i.e., flipping sentiment) with around 15%. However, even with the many errors, the DQN-TAMER agent successfully learned a good policy with a limited number of episodes.

V. Summary and Future Work

We introduced five key problems of human feedback in real applications. The result with a simulated human indicates the effectiveness of combining rewards from a task and a human with such intractable feedback. We also built a car robot system that exploits implicit rewards by reading human faces with a CNN based classifier. As future work, we will tackle other high-dimensional tasks by incorporating state-of-the-art RL algorithms.

References

[1] A. L. Thomaz, *et al.*, "Real-time interactive reinforcement learning for robots," in *AAAI 2005 workshop on human comprehensible machine learning*, 2005.

[2] ——, "Reinforcement learning with human teachers: Understanding how people want to teach robots," in *The 15th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, 2006, pp. 352–357.

[3] J. Broekens, "Emotion and reinforcement: affective facial expressions facilitate robot learning," in *Artifical intelligence for human computing*. Springer, 2007, pp. 113–132.

[4] W. B. Knox and P. Stone, "TAMER: Training an agent manually via evaluative reinforcement," in *2008 7th IEEE International Conference on Development and Learning*, Aug 2008, pp. 292–297.

[5] A. C. Tenorio-Gonzalez, *et al.*, "Dynamic reward shaping: Training a robot by voice," in *Proceedings of the 12th Ibero-American Conference on Advances in Artificial Intelligence*, 2010, pp. 483–492.

[6] P. M. Pilarski, *et al.*, "Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning," in *IEEE International Conference on Rehabilitation Robotics*, 2011, pp. 1–7.

[7] S. Griffith, *et al.*, "Policy shaping: Integrating human feedback with reinforcement learning," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2625–2633.

[8] J. MacGlashan, *et al.*, "Interactive learning from policy-dependent human feedback," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 2285–2294.

[9] G. Warnell, *et al.*, "Deep TAMER: Interactive agent shaping in high-dimensional state spaces," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[10] C. C. Preston and A. M. Colman, "Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences," *Acta Psychologica*, vol. 104, no. 1, pp. 1 – 15, 2000.

[11] W. E. Hockley, "Analysis of response time distributions in the study of cognitive processes." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 10, no. 4, p. 598, 1984.

[12] C. L. Isbell Jr and C. R. Shelton, "Cobot: A social reinforcement learning agent," in *Advances in Neural Information Processing Systems*, 2002, pp. 1393–1400.

[13] C. Isbell, *et al.*, "A social reinforcement learning agent," in *Proceedings of the fifth international conference on Autonomous agents*. ACM, 2001, pp. 377–384.

[14] I. Çugu, *et al.*, "Microexpnet: An extremely small and fast model for expression recognition from frontal face images," *arXiv*, vol. 1711.07011, pp. 1–9, 2017.