

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

Riku Arakawa^{*†}, Sosuke Kobayashi[†], Yuya Unno[†], Yuta Tsuboi[†], Shin-ichi Maeda[†]

Abstract—Exploration has been one of the greatest challenges in reinforcement learning (RL), which is a large obstacle in the application of RL to robotics. Even with state-of-the-art RL algorithms, building a well-learned agent often requires too many trials, mainly due to the difficulty of matching its actions with rewards in the distant future. A remedy for this is to train an agent with real-time feedback from a human observer who immediately gives rewards for some actions. This study tackles a series of challenges for introducing such a human-in-the-loop RL scheme. The first contribution of this work is our experiments with a precisely modeled human observer: BINARY, DELAY, STOCHASTICITY, UNSUSTAINABILITY, and NATURAL REACTION. We also propose an RL method called DQN-TAMER, which efficiently uses both human feedback and distant rewards. We find that DQN-TAMER agents outperform their baselines in Maze and Taxi simulated environments. Furthermore, we demonstrate a real-world human-in-the-loop RL application where a camera automatically recognizes a user’s facial expressions as feedback to the agent while the agent explores a maze.

I. INTRODUCTION

Reinforcement learning (RL) has potential applications for autonomous robots [1]. Even against highly complex tasks like visuomotor-based manipulation [2] and opening a door with an arm [3], skillful policies for robots can be obtained through repeated trials of deep RL algorithms.

However, exploration remains as one of the greatest challenges, preventing RL from spreading to real applications. It often requires a lot of trials until the agent reaches an optimal policy. 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. The estimated values of actions in given states are modified exponentially slowly over the number of remaining intervals until the future reward is received [4].

Additional training signals from a human are a very useful remedy. One direction involves human demonstrations. Using human demonstrations for imitation learning can efficiently train a robot agent [5], though it is sometimes difficult or time-consuming to collect human demonstrations.

We use real-time feedback from human observers as another helpful direction in this study. During training, human observers perceive the agent’s actions and states in the environment and provide some feedback to the agent in real time rather than at the end of each episode. Such

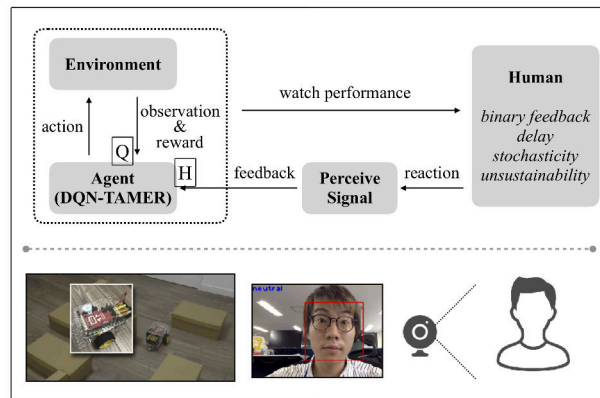


Fig. 1: Overview of human-in-the-loop RL and our model (DQN-TAMER). The agent asynchronously interacts with a human observer in the given environment. DQN-TAMER decides actions based on two models. One (Q) estimates rewards from the environment and the other (H) for feedback from the human.

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 in prior publications [6]–[15].

Human-in-the-loop RL has the potential to greatly improve training thanks to the immediate rewards. However, experiments in prior studies did not 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. In this paper, we reformulate human observers with the following more realistic characteristics: binary feedback, delay, stochasticity, and unsustainability. Furthermore, we examine the effect from recognition errors, when an agent autonomously infers implicit human reward from natural reactions like facial expressions. Table I shows a comparison with prior work.

With such a human-in-the-loop setup, we derive an efficient RL algorithm called DQN-TAMER from an existing human-in-the-loop algorithm (TAMER) and deep Q-learning (DQN). The DQN-TAMER algorithm learns two disentangled value functions for the human immediate reward and distant long-term reward. DQN-TAMER can be seen as a generalization of TAMER and DQN, where the contribution from each model can arbitrarily controlled.

The contributions of the paper are as follows:

- 1) We precisely formulate the following more realistic human-in-the-loop RL settings: (BINARY FEEDBACK,

^{*} The University of Tokyo.

arakawa-riku428@g.ecc.u-tokyo.ac.jp

[†] Preferred Networks, Inc.

{sosk, unno, tsuboi, ichi}@preferred.jp

TABLE I: Characteristics of human observers tested in prior work and this study

study	BINARY	DELAY	STOCHASTICITY	UNSUSTAINABILITY	NATURAL REACTION
Andrea et al. 2005 [6], [7]		✓	✓		
Joost Broekens 2007 [8]	✓			✓	✓ (facial expression)
Knox et al. 2007 [9]	✓	✓	✓		
Tenorio-Gonzalez et al. 2010 [10]		✓	✓		✓ (voice)
Pilarski et al. 2011 [11]	✓	✓	✓		
Griffith et al. 2013 [12]	✓		✓		
MacGlashan et al. 2017 [13]		✓	✓	✓	
Arumugam et al. 2018 [14]		✓	✓	✓	
Warnel et al. 2018 [15]	✓	✓	✓		
Ours	✓	✓	✓	✓	✓ (facial expression)

DELAY, STOCHASTICITY, UNSUSTAINABILITY, NATURAL REACTION).

- 2) We propose an algorithm, DQN-TAMER, for human-in-the-loop RL, and demonstrate that it outperforms the existing RL methods in two tasks with a human observer.
- 3) We built a human-in-the-loop RL system with a camera, which autonomously recognized a human facial expression and exploited it for effective explorations and faster convergence.

II. PROBLEM FORMULATION

We first describe the standard RL settings and subsequently introduce a human observer for human-in-the-loop RL, as shown in Figure 1. We then describe the characteristics of the human observer.

In standard RL settings, an agent interacts with an environment \mathcal{E} through a sequence of observations, actions, and rewards. At each time t , the agent receives an observation s_t from \mathcal{E} , takes an action a_t from a set of possible actions \mathcal{A} , and then obtains a reward r_{t+1} . Let $\pi(a|s)$ be the trainable policy of the agent for choosing an action a from \mathcal{A} given an observed state s . The ultimate goal of RL is to find the optimal policy which maximizes the expectation of the total reward $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$ at each state s_t , where γ is a discount factor for later rewards.

Next, we consider introducing a human into the above RL settings. At each step, a human watches the agent’s action a_t and the next state s_{t+1} , assesses a_t based on intuition or some other criteria, and gives some feedback f_{t+1} to the agent through some type of reaction. Prior work has explored modeling human feedback. This study discusses and reformulates those clearly as five components. This paper is the first study that fully integrates all components and performs experiments and analysis to test their effects.

A. Binary

Some studies consider humans giving various values as feedback to influence the agent [6], [7]. However, requesting people give fine-grained or continuous scores is found difficult [16] because it requires human have enough understanding of the task at hand and requires that human can rate the agent behavior quantitatively in an objective manner. This is why binary feedback is preferred. The feedback simply indicates whether an action is good or bad. In this way, even an ordinary person can be a desirable observer and provide

feedback as well as an expert [17]. Thus, we assume binary feedback, i.e., $f_t \in \{-1, +1\}$.

B. Delay

One may think that human feedback will surely accelerate an agent’s learning. In realistic settings, however, it is actually difficult to utilize feedback because human feedback is usually delayed by a significant amount of time [18]. In particular, the agent must perform actions in a dynamic environment where the state changes continuously. Thus, the agent cannot wait for feedback at each step. Furthermore, the delay must not be constant, implicitly depending on people’s concentration, complexity of states, and actions, etc. The randomness of delay makes the problem much more difficult. We assume that the number of feedback delay steps follows a certain probability distribution.

Surprisingly, we found that human feedback could have totally “negative” effects on the existing learning algorithms of an agent if the agent ignores this delay effect and takes the feedback as exact and immediate feedback. On the other hand, our proposed learning algorithm succeeds when such delayed human feedback is utilized even though the actual probability of delay is different from the one we assumed.

C. Stochasticity

In addition to delaying feedback, other studies missed the idea that people could not always give feedback when an agent performs an action correctly. It is also reported that the feedback frequency varies largely among human users [19], [20]. Thus, such a stochastic drop is a factor of intractable human feedback that we have to model for human-in-the-loop RL.

We introduce p_{feedback} to indicate the probability that appropriate feedback occurs in a time step (i.e. the probability of avoiding drop) to model the difficulty of random events. We vary the strength of stochasticity in the following experiments and confirm a significant effect in learning process.

D. Unsustainability

Even after introducing delay and stochasticity, the setting is still less realistic. It is very difficult to presume that humans watch an agent until it finishes learning through many episodes. The learning process might last a long time, thus a human may leave before the agent converges to an optimal policy. 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. Here we introduce the notion of feedback stop with a time step t_{stop} , where the human leaves the environment and the agent stops receiving feedback. We confirm that ending feedback degrades learning process of prior algorithms; in contrast, our proposed algorithm works robustly.

E. Natural Reaction

Finally, the method used to provide feedback is not unique or obvious. One naive method for providing binary feedback is using positive-negative buttons or levers. However, when intelligent agents become more ubiquitous and we launch real human robot interaction systems, it is preferable that the system infer implicit feedback from natural human reactions rather than humans actively providing feedback. Robots with such a mechanism would be capable of lifelong learning [21], [22] after deployment in the real world. For example, robot pets might utilize their owner's voice as feedback for directions or some toy tasks, or communication robots possibly infer feedback from a user via their facial expressions.

In this paper, we investigated the use of human facial expressions. We use a deep neural network-based classifier for facial expression recognition and we built a demo system with a camera. Note that classification errors from such a model cause an agent to misunderstand the sentiment polarity (positive or negative) associated with feedback. This is another important issue which we believe will arise in future human robot interaction applications.

III. METHODS

We first describe two existing RL algorithms. Each algorithm is a well-known deep RL method. We then propose an algorithm that generalizes both of them.

A. Deep Q-Network (DQN)

The optimal policy can be characterized as the policy that causes the agent to take an action that maximizes the action value for the action in the given state [23], [24]. An action value function $Q_\pi : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$ is a function that returns the expected total reward in a given state and for a given action when following the policy π [25]. The optimal action value is defined as the maximum action value function with respect to the policy.

$$Q^*(s, a) = \max_{\pi} Q_\pi(s, a). \quad (1)$$

Q-learning is an algorithm that estimates the optimal action value by iteratively updating the action value function using the Bellman update [23].

A deep Q-network (DQN) is a kind of approximate Q-learning that utilizes a deep neural network to represent the action value function together with some tricks in training, such as experience replay [26], reward clipping, and a target network for stabilizing training [27].

To handle the human feedback in the framework of RL, we augment an extra reward function that computes a scalar reward for human feedback in addition to the original reward

function, i.e., we employ so-called *reward shaping* [28], [29] to incorporate human feedback.

B. Deep TAMER

TAMER [9] 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. In short, TAMER is a value-based RL algorithm where the values are estimated from human feedback only. Deep TAMER [15] is an algorithm that applies a deep neural network within this TAMER framework.

In Deep TAMER, the H-function is used instead of the Q-function to show the value of an action at a certain state ($H : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$). The difference from Q-learning is that H-function estimates a binary human feedback f for each action. Similar to DQN, and given the current estimate \hat{H} , the agent policy is

$$\pi(s)_{\text{DeepTAMER}} = \arg \max_a \hat{H}(s, a). \quad (2)$$

Deep TAMER considers a certain feedback that corresponds to some recent state and action pairs, which expects DELAY. Let s and a be a sequence of states and actions, respectively. The loss function L for judging the quality of \hat{H} is defined as follows:

$$L(\hat{H}; s, a, f) = \sum_{s \in s, a \in a} \|\hat{H}(s, a) - f\|^2, \quad (3)$$

The optimal feedback estimation is the value of \hat{H} that minimizes the expected loss value, and Deep TAMER updates this using stochastic gradient descent (SGD):

$$\hat{H}_\pi^*(s, a) = \underset{\hat{H}}{\operatorname{argmin}} \mathbb{E}_{s, a} [L(\hat{H}; s, a, f)] \quad (4)$$

$$\hat{H}(s, a)_{k+1} = \hat{H}(s, a)_k - \eta_k \nabla_{\hat{H}} L(\hat{H}_k; s, a, f) \quad (5)$$

where η_k is the learning rate at update iteration k .

Also, inspired by experience replay in DQN [26], a similar technique is introduced to stabilize learning by the \hat{H} neural network. D_{local} is a set of tuples for a state, action, and feedback when a single feedback f is received, which is defined as

$$D_{\text{local}} = \{(s, a, f) \mid (s, a) \in (s, a)\}. \quad (6)$$

D_{global} stores all the past states, actions, and feedback pairs. Every time a new feedback occurs, it updates as follows:

$$D_{\text{global}} \leftarrow D_{\text{global}} \cup D_{\text{local}} \quad (7)$$

The TAMER framework (including Deep TAMER) only exploits human feedback and lacks the ability to make use of rewards from the environment. Our proposed method is described in the next subsection, where the agent successfully uses both human feedback and environmental rewards.

C. Proposed DQN-TAMER

Our motivation lies in integrating the TAMER framework into an existing value based on Q-learning and, therefore, achieves faster agent learning convergence.

Algorithm 1 Deep TAMER

Require: initialized \hat{H} , update interval b , learning rate η
Ensure: $D_{\text{global}} = \emptyset$
while NOT goal or time over **do**
 observe s
 execute $a \sim \pi(s)_{\text{DeepTAMER}}$ by (2)
 if new feedback f **then**
 prepare s, a
 obtain D_{local} by (6)
 update D_{global} by (7)
 update $\hat{H}(s, a)$ by (5) using D_{local}
 if every b steps and $D_{\text{global}} \neq \emptyset$ **then**
 update $\hat{H}(s, a)$ by (5) using mini-batch sampling from D_{global}

Algorithm 2 Proposed: DQN-TAMER

Require: initialized \hat{H} , \hat{Q} , update interval b , learning rate η , weight α_q, α_h
Ensure: $D_{\text{global}} = \emptyset$
while NOT goal or time over **do**
 observe s
 execute $a \sim \pi(s)_{\text{DQN-TAMER}}$ by (8)
 decay α_h
 if new feedback f **then**
 prepare s, a
 obtain D_{local} by (6)
 update D_{global} by (7)
 update $\hat{H}(s, a)$ by (5) using D_{local}
 if every b steps **then**
 update $\hat{Q}(s, a)$
 if $D_{\text{global}} \neq \emptyset$ **then**
 update $\hat{H}(s, a)$ by (5) using mini-batch sampling from D_{global}

DQN-TAMER trains the Q-function and H-function separately using the DQN and Deep TAMER algorithms. Given the estimated \hat{Q} and \hat{H} respectively, the agent policy is defined as

$$\pi(s)_{\text{DQN-TAMER}} = \arg \max_a \alpha_q \hat{Q}(s, a) + \alpha_h \hat{H}(s, a), \quad (8)$$

where α_q and α_h are the hyper parameters that determine the extent to which the agent relies on the reward from the environment and feedback from a human. Note that, α_h decays at every step and eventually $\alpha_h \rightarrow 0$, thus the agent initially explores efficiently by following human feedback and eventually reaches the optimal DQN policy much faster.

Since we train each network separately and combine them only when choosing actions, it is no surprise that original DQN and Deep TAMER are written in this DQN-TAMER framework. DQN is equivalent when $\alpha_h = 0$ and Deep TAMER is equivalent when $\alpha_q = 0$. Thus, DQN-TAMER can also be seen as a method for annealing DQN that is aided by including human feedback in the pure DQN algorithm.

In summary, we have four algorithms: (1) DQN, (2)

DQN with naive reward shaping where feedback is added to environmental rewards, (3) Deep TAMER, and (4) our proposed DQN-TAMER algorithm. In the following experiments, we compare these algorithms and show that DQN-TAMER outperforms the others in terms of learning speed and final agent performance.

IV. EXPERIMENTAL SETTINGS

Two experiments were performed. The first experiment aims to compare and analyze each algorithm in fair and wide settings. We prepare programs as simulated human observers based on the four requirements described in Sec II (BINARY, DELAY, STOCHASTICITY, UNSUSTAINABILITY). Following Griffith et al. [12], the simulated human gives feedback when certain conditions are satisfied for a given state and action. The simulated approaches are appreciated because we can systematically test the performance of various algorithms with hyperparameters in a consistent setting. Even with deep RL algorithms, whose performance can vary largely due to random seeds, we can fairly compare them by averaging the results from many runs. We actually used a trimmed mean of results from 30 runs in all experiments for a reliable comparison.

We trained the agents in two game environments: *Maze* and *Taxi*. As for a human observer in the simulated world, there are parameters which should be decided beforehand (p_{delay} for DELAY, p_{feedback} for STOCHASTICITY and t_{stop} for UNSUSTAINABILITY). As for the probability of the delay, p_{delay} , we assume it as $p_{\text{delay}}(0) = 0.3$, $p_{\text{delay}}(1) = 0.6$, $p_{\text{delay}}(2) = 0.1$, $p_{\text{delay}}(n) = 0$ ($n \geq 3$). Because this true probability of the delay is unknown in reality, we assume the different one during training, which is given by $p_{\text{delay}}(i) = 1/3$ for $i \in \{0, \dots, 2\}$ otherwise $p_{\text{delay}}(i) = 0$, following Warnell et al. [15].

Second, we built a real human-in-the-loop RL system to demonstrate the effectiveness of the proposed method in real applications. The system uses a camera to perceive human faces and interpret them as human feedback using a deep neural network for facial expression recognition. Even though such implicit feedback is actively inferred by the system, it learns maze navigation well. We show the results from the demo in our complementary video.

A. Maze

Maze is a classical game where the agent must reach a predefined goal (Figure 2). We compare the sample efficiency in each algorithm through experiment, i.e., we examine how fast learning converges. We fixed the field size of a maze to 8×8 and the initial distance to the goal at 5. Table II summarizes the environmental setting.

We simulate a human feedback as it gives a binary label whether the agent reduces the Manhattan distance to the goal. If an agent moves closer to the goal, the human provides +1 positive feedback and -1 negative feedback otherwise. We experimented with two different settings of observations s_t , which an agent can see from the environment. In the first setting, an agent only knows its own absolute coordinates

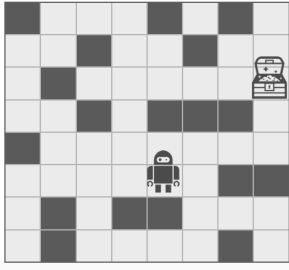


Fig. 2: Maze: an environment with walls (black squares), the agent, and the goal.

TABLE II: Maze setting

reward	every step -0.01, goal +1.0
field size	8
initial distance	5
max steps	1000
action space	(north, east, south, west)
human rule	Manhattan distance to the goal

in a maze. In the other setting, it observes the status of the surrounding areas (8 squares). In the case shown in Figure 2, an agent observe either absolute coordinate “(6, 5)” or partial observation [“space”, “space”, “space”, “space”, (“now”,) “space” ’, “wall”, “wall”, “space”] respectively in each setting. Observation of only surrounding areas follows a partially observable Markov decision process (POMDP) [30]. The POMDP framework is general enough to model a variety of real-world sequential decision processes, such as robot navigation problems, machine maintenance, and planning under uncertainty in general, but is also known it is difficult environment to train the agent.

B. Taxi

Taxi is also a moving game in a two-dimensional space (Figure 3), but it is more difficult due to its hierarchical goals [31]. In Taxi, an agent must pick up a passenger that is waiting at a certain position and move him/her to a different position. The position of the passenger and the final destination are randomly chosen from four candidate positions {R, G, B, Y}.

Thus, the optimal direction is different before and after picking up the passenger. The agent must learn such a two-staged policy to solve this task. We fix the field size of a maze to 5×5 . Table III summarizes the environment settings. The agent observes the current absolute coordinates and whether or not the passenger is currently in the taxi (agent). We simulate a human feedback as it gives a binary label whether it reduces the distance to a passenger or the destination according to the state of picking up.

C. Car Robot Demonstration

As a further demonstration, we built a demo system and trained a car agent with a real human observer. We also introduce NATURAL REACTION in this demonstration as described in Sec. II, thus bringing the system closer to real applications. Feedback is inferred by observing a person and is obtained through facial expression recognition.

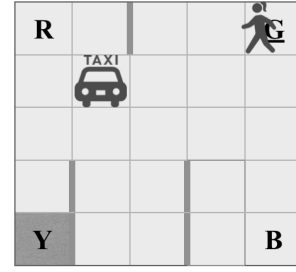


Fig. 3: Taxi: an environment with walls (| ; bold bars), the taxi agent, the passenger (at G), and the goal (Y).

TABLE III: Taxi setting

reward	every step -1, drop at right/wrong place +20/-10 pickup at the wrong place -10
field size	5
initial distance	random
max steps	1000
action space	(north, east, south, west, pickup, drop)
human rule	Manhattan distance to passenger (before pick up) Manhattan distance to goal (after pick up)

We used MicroExpNet as a recognition model, which is a convolutional neural network-based (CNN) model [32]. This model is obtained by distilling a larger CNN model, which then quickly and accurately classifies facial expressions into 8 categories: ‘neutral’, ‘anger’, ‘contempt’, ‘disgust’, ‘fear’, ‘happy’, ‘sadness’, ‘surprise’. Even such an accurate model, of course, often fails to predict the correct expression. The intriguing question we tackle here is whether an agent can learn well from suspicious feedback with errors. Figure 7 shows how we set up the environment with a car robot solving a physical maze. The agent interprets the facial expression ‘happy’ as positive (+1) and other expressions (‘anger’, ‘contempt’, ‘disgust’, ‘fear’, and ‘sad’) as negative (-1).

D. Parameter Settings

We construct every Q-function and H-function as a feed-forward neural network with a hidden layer using tanh function of 100 dimensions. Optimization is performed using RMSProp, where the initial learning rate is 10^{-3} both for the Q-function and the H-function. The probability of taking random actions for exploration is initially set to 0.3 and decayed by 0.001 at every step until it reaches 0.1. We initialize $\alpha_h = \alpha_q = 1$ of the DQN-TAMER and decay α_h by 0.9999 at every step.

V. RESULTS AND DISCUSSION

In the following, we show the averaged results over totally 30 trials for each environment, where the results were obtained from three each with ten different sets of initial conditions.

A. DELAY and STOCHASTICITY

To investigate the dependence on the delay of human feedback and the feedback occurrence probability, we conducted experiments by varying the probability of feedback

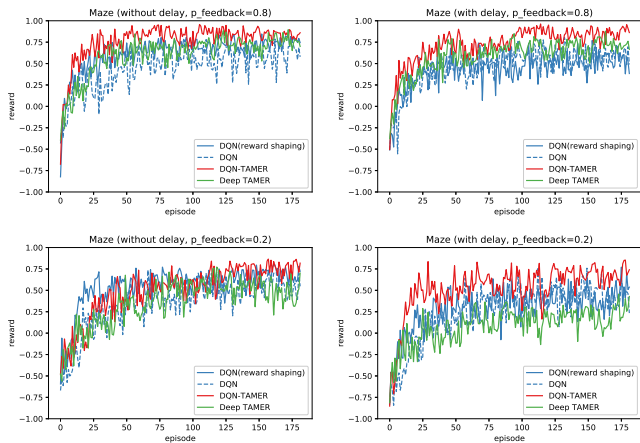


Fig. 4: Maze results (upper: high frequency, lower: low frequency, left: without delay, and right: with delay).

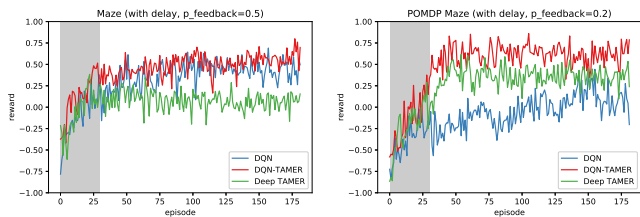


Fig. 5: Maze with feedback stop. Feedback ends after 30 episodes. left: MDP, right: POMDP

occurrence (p_{feedback}) and the existence of delay. Figure 4 shows the four results of Maze each of which corresponds the condition either the feedback frequency is high and low, and delay happens or does not.

As for DELAY, we can see that DQN with reward shaping outperforms DQN if there is no delay by comparing left and right panels of the figure. However, the performance of one with reward shaping degrades and becomes comparable with pure DQN if delay is introduced. This suggests that the human feedback does not work well by naive reward shaping.

Comparing the upper and lower figures, one can see that a learning process with more frequent feedback is faster and reaches higher rewards for all algorithms. Less frequent feedback degrades the performance of all algorithms. Deep TAMER returned a particularly poor result. Among those, DQN-TAMER is the most robust with unstable feedback since it uses a Q-function and an H-function. Therefore, it can also take advantage of rewards from the environment.

B. UNSUSTAINABILITY

We investigate the effect to learning when human feedback gets interrupted. In any case, DQN-TAMER outperforms the other methods. It is inferred that Deep TAMER becomes stagnant after feedback stops because it depends only on human feedback. In contrast, DQN-TAMER initially facilitates efficient exploration with human feedback and continues improving its policy with rewards from the environment. The result is consistent with experimental results from Maze and

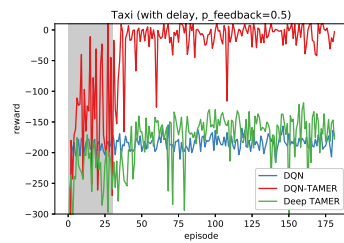


Fig. 6: Taxi with feedback stop. Feedback ends after 30 episodes.

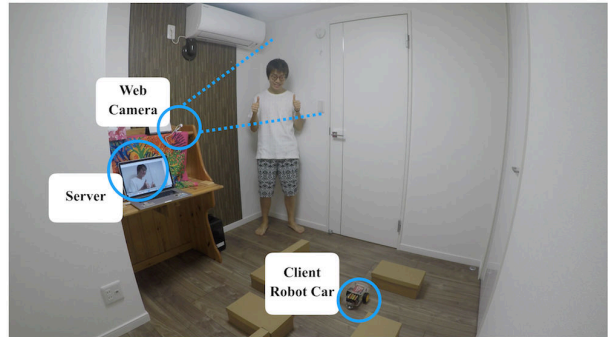


Fig. 7: Demonstration situation. We used a GoPiGo3 car robot and trained it to solve a maze using human facial expressions.

Taxi. Our proposed DQN-TAMER is very robust to various types of human feedback.

C. NATURAL REACTION

During the car robot demonstration, we found that the agent learned well from suspicious feedback with errors of the classifier efficiently. The facial expression classifier misclassified human facial expressions (i.e., flipping plus and minus of reward) with around 15%. The result demonstrated that the DQN-TAMER was robust even though such opposite feedback occur stochastically. We show the learning process in the complementary video.

VI. CONCLUSION

This study tackles a series of challenges for introducing human-in-the-loop RL into real world robotics. We discussed five key problems for human feedback in real applications: BINARY, DELAY, STOCHASTICITY, UNSUSTAINABILITY and NATURAL REACTION. The experiments results obtained from various settings show that the proposed DQN-TAMER model is robust against inconvenient feedback and outperforms existing algorithms like DQN and Deep TAMER. We also built a car robot system that exploits implicit rewards by reading human faces with a CNN based classifier. Even with classifier errors, the agent of the system efficiently learned maze navigation. These results encourage to utilize the human feedback in a real world scenario, which is difficult to handle due the instability and randomness of the delay, if we assume the randomness of the delay even when the probability function is different from the true one and combine the human feedback appropriately with the original reward given by the environment.

REFERENCES

- [1] J. Kober, *et al.*, “Reinforcement learning in robotics: A survey,” *I. J. Robotics Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [2] S. Levine, *et al.*, “End-to-end training of deep visuomotor policies,” *Journal of Machine Learning Research*, vol. 17, pp. 39:1–39:40, 2016.
- [3] S. Gu, *et al.*, “Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates,” in *IEEE International Conference on Robotics and Automation, ICRA*, 2017, pp. 3389–3396.
- [4] J. A. Arjona-Medina, *et al.*, “RUDDER: return decomposition for delayed rewards,” *CoRR*, vol. abs/1806.07857, 2018.
- [5] A. Nair, *et al.*, “Overcoming exploration in reinforcement learning with demonstrations,” *CoRR*, vol. abs/1709.10089, 2017.
- [6] A. L. Thomaz, *et al.*, “Real-time interactive reinforcement learning for robots,” in *AAAI 2005 workshop on human comprehensible machine learning*, 2005.
- [7] —, “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.
- [8] J. Broekens, “Emotion and reinforcement: affective facial expressions facilitate robot learning,” in *Artificial intelligence for human computing*. Springer, 2007, pp. 113–132.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] S. Griffith, *et al.*, “Policy shaping: Integrating human feedback with reinforcement learning,” in *Advances in Neural Information Processing Systems 26*, 2013, pp. 2625–2633.
- [13] 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.
- [14] D. Arumugam, *et al.*, “Deep reinforcement learning from policy-dependent human feedback,” 2018.
- [15] 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.
- [16] 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.
- [17] P. F. Christiano, *et al.*, “Deep reinforcement learning from human preferences,” in *Advances in Neural Information Processing Systems 30*, 2017, pp. 4302–4310.
- [18] 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.
- [19] 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.
- [20] C. Isbell, *et al.*, “A social reinforcement learning agent,” in *Proceedings of the fifth international conference on Autonomous agents*. ACM, 2001, pp. 377–384.
- [21] S. Thrun and T. M. Mitchell, “Lifelong robot learning,” *Robotics and Autonomous Systems*, vol. 15, pp. 25–46, 1995.
- [22] C. Finn, *et al.*, “Generalizing skills with semi-supervised reinforcement learning,” in *Proceedings of ICLR*, 2016.
- [23] C. J. C. H. Watkins and P. Dayan, “Q-learning,” *Machine Learning*, vol. 8, no. 3, pp. 279–292, May 1992.
- [24] R. S. Sutton, “Learning to predict by the methods of temporal differences,” *Machine Learning*, vol. 3, no. 1, pp. 9–44, Aug 1988.
- [25] S. J. Russell and P. Norvig, *Artificial intelligence - a modern approach*, 2nd Edition, ser. Prentice Hall series in artificial intelligence, 2003.
- [26] L.-J. Lin, “Self-improving reactive agents based on reinforcement learning, planning and teaching,” *Machine Learning*, vol. 8, no. 3, pp. 293–321, May 1992.
- [27] V. Mnih, *et al.*, “Playing atari with deep reinforcement learning,” *CoRR*, vol. abs/1312.5602, 2013.
- [28] A. Y. Ng, *et al.*, “Policy invariance under reward transformations: Theory and application to reward shaping,” in *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999, pp. 278–287.
- [29] W. B. Knox and P. Stone, “Learning non-myopically from human-generated reward,” in *18th International Conference on Intelligent User Interfaces*, 2013, pp. 191–202.
- [30] L. P. Kaelbling, *et al.*, “Planning and acting in partially observable stochastic domains,” *Artif. Intell.*, vol. 101, no. 1-2, pp. 99–134, 1998.
- [31] T. G. Dietterich, “Hierarchical reinforcement learning with the maxq value function decomposition,” *Journal of Artificial Intelligence Research*, vol. 13, pp. 227–303, 2000.
- [32] 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.