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

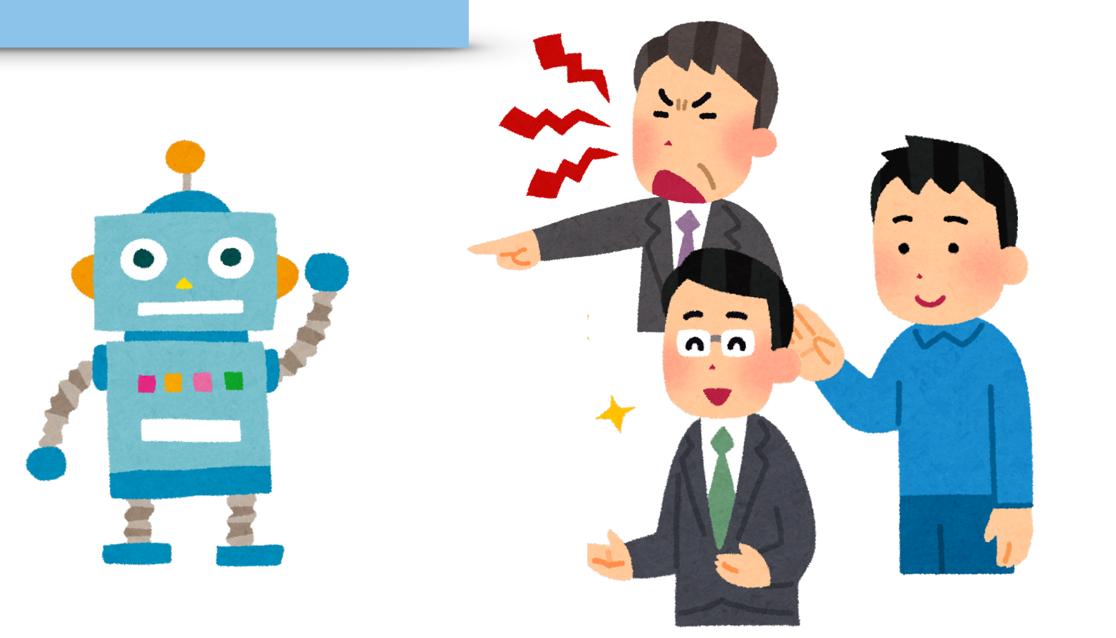
Riku Arakawa[1], *Sosuke Kobayashi*[2], *Yuya Unno*[2], *Yuta Tsuboi*[2], *and Shin-ichi Maeda*[2]

[1] *The University of Tokyo*, [2] *Preferred Networks, Inc.*

* This work is done in Preferred Networks, Inc.

* Full paper is available at *https://arxiv.org/abs/1810.11748*

Motivation



- Reinforcement learning (RL) is promising for robotics, but it requires a great deal of time.
- One reason is that agent can get rewards from environments often at the end of the task, long after some of their actions.

Can robots learn from human through immediate natural response such as their facial expression or behavior?

1. What's the difficulty of human feedback?

→*reformulate human observers with more realistic characteristics.*

2. How can we mix human feedback and task reward?

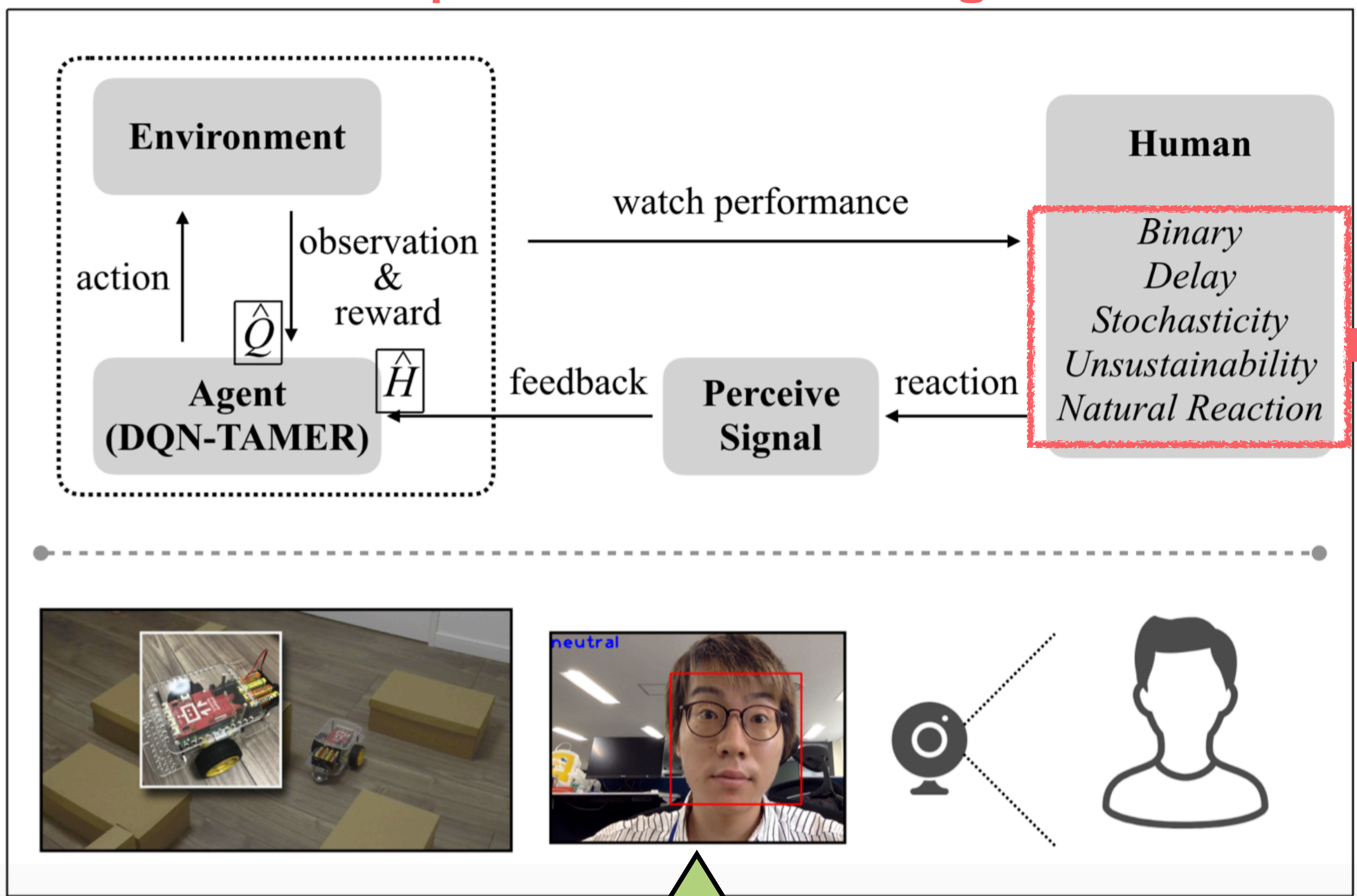
→*apply a simple RL algorithm that utilizes rewards from both human and environments.*

3. Can the agent read human natural response as feedback?

→*demonstrate in the real world setting with human facial expression as rewards.*

1. Problem Formulation

Human-in-the-Loop reinforcement learning



CNN-Based classifier for facial expression

factor	description
binary	Binary feedback is preferred, simply indicating good or bad.
delay	Human feedback is usually delayed by a significant amount of time and the delay must not be constant.
stochastic	It is reported that the feedback frequency varies largely among human users.
unsustainable	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.
natural reaction	It is preferable that the system infers implicit feedback from natural human reactions rather than what humans provide actively.

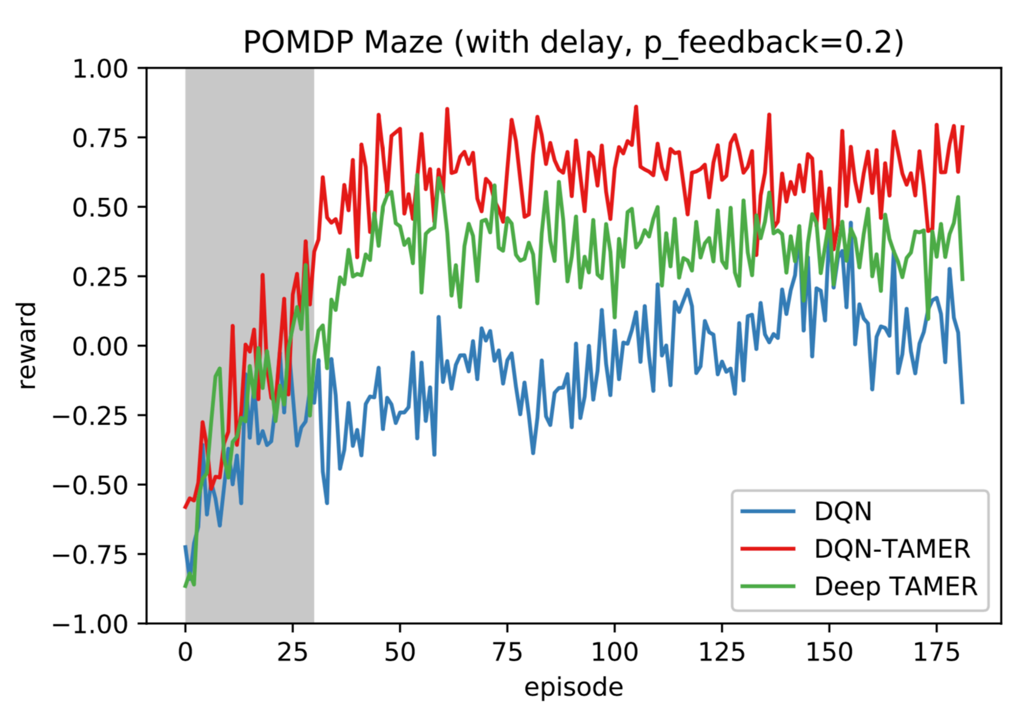
Ours is the first to consider all of these points. →

Study	*Binary*	*Delay*	*Stoch-astic*	*Unsust-ainable*	*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)

2. Method: DQN-TAMER

Policy:

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



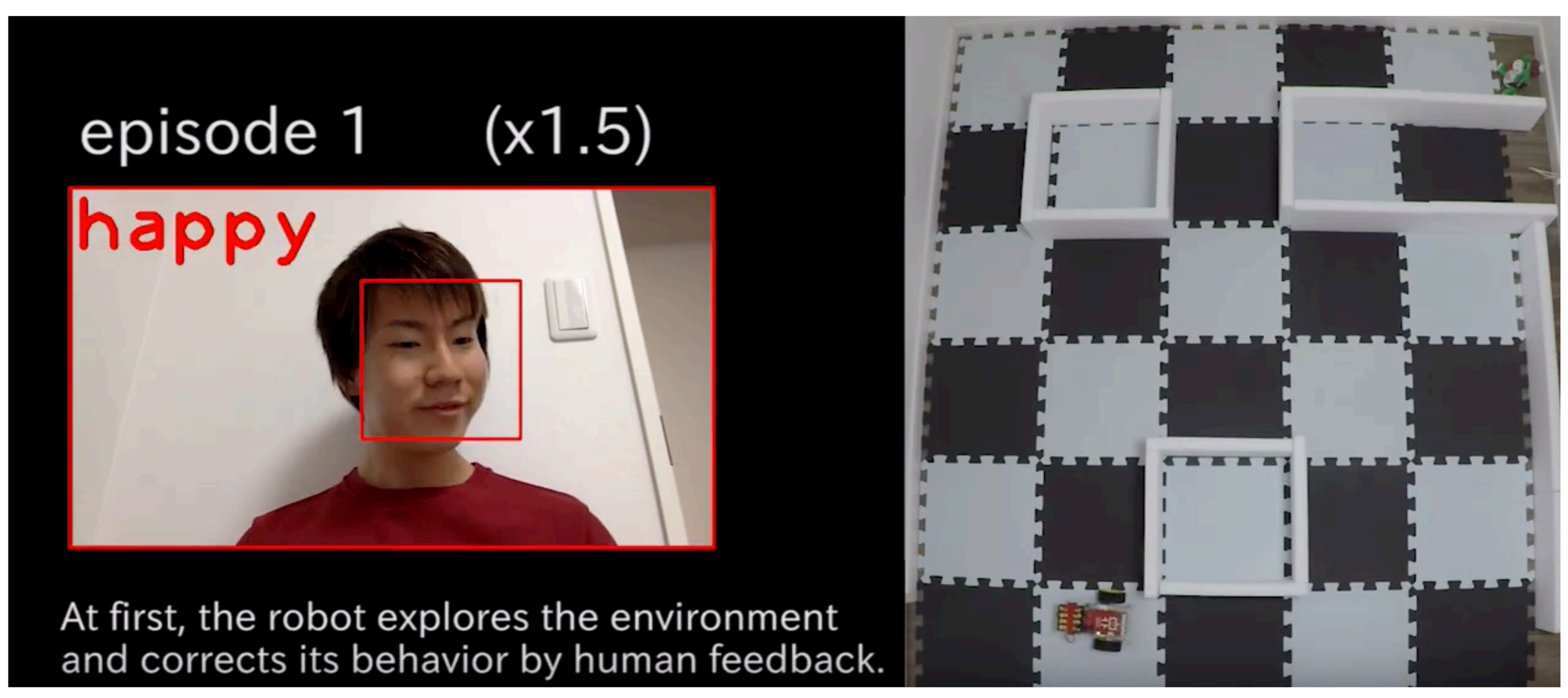
Evaluate its robustness in various settings of simulated human.

<u>DQN-TAMER outperforms baselines.</u>

3. Demonstration

Car robot solving a grid maze

✓ <u>The agent could utilize human facial expression, even though its recognition sometimes failed.</u>



[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 Process- ing 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.