2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Shaping Social Robot to Play Games with Human Demonstrations and Evaluative Feedback

Chuanxiong Zheng[1], Lei Zhang[1], Hui Wang[1], Randy Gomez[2], Eric Nichols[2], Guangliang Li[1*]

***Abstract*— In this paper, building on recent advances in the fields of gaming AI and social robotics, we present a new approach to facilitate the social robot Haru to imitate game strategies from human players' demonstrated trajectories and evaluative feedback in a real-time two-player game. Our research shows that Haru is able to learn and imitate human different game strategies from human players in a human time scale. In addition, our results show that human evaluative feedback plays an important role in allowing Haru to obtain a better performance via our method than human player's demonstrations. Finally, results of our user study indicate that Haru imitating human player's game strategies via our method is perceived to be more human-like and have better game performance and experience than self-learning from pre-defined reward functions via traditional deep reinforcement learning.**

I. INTRODUCTION

In recent years, research on social robots has attracted a lot of attention due to the rapid development of artificial intelligence (AI). As social robots move from lab to human-inhabited environments, they are no longer just tools for humans, but companions or partners of humans [1]. As human-human interaction, in human's daily lives with social robots, games will play an important and indispensable role in the interaction between social robots and humans.

For example, as one of the favorite activities of children, many researchers have tried to develop social robots that can play games with children to promote their learning skills. Strohkorb et al. [2] conducted an interdisciplinary study in which pairs of children played interactive build-a-rocket games with social robots to promote the growth and use of children's collaborative skills: increased attention to task and enhanced interpersonal cohesiveness. Gordon et al. [3] developed an integrated experimental paradigm and an integrated affection tutoring system in which children work with fully autonomous social robot learning partners to play second language learning games on a tablet computer. The system combines educational content, emotion sensing, and expressive social robots deployed in real-world, long-term interaction studies, and demonstrates that the emotional individualization of social robot tutors can positively influence students' emotions in constructive and meaningful ways. Gillet et al. [4] proposed Cozmo, a social robot mediator that plays music-based puzzles with children to help some children overcome intergroup prejudice, especially for some children that are new to a country. However, in most of their work, social robots have fixed game playing levels and strategies that cannot be adapted, and human partners cannot guide the game strategies of social robots either.

On the other hand, games have always been a valuable test bed for artificial intelligence (AI) research as benchmarking tasks. For example, Mnih et al. [5] presented the first deep reinforcement learning (RL) model that successfully learned control policies directly from high-dimensional sensory inputs in the Atari games. Vinyals et al. [6] proposed AlphaStar using deep neural networks trained directly from raw game data through supervised learning and reinforcement learning, which defeat the world's most professional Starcraft players. However, although agents can learn to achieve grand-master or superhuman performance in the above games via reinforcement learning, it still remains a great challenge for them to learn in a human time scale to play these kinds of games because of sparse rewards or complexity. Moreover, since most robots will operate in human-inhabited environments, to become companions of humans, it is of great importance for them to interact and learn from human users [7], [8]. In this paper, we proposed to allow a social robot Haru to learn game strategies from human players' demonstrations and evaluative feedback in real-time two-player games based on recent advances in the fields of gaming AI and social robotics, as shown in Fig. 1. We hypothesize that with our method, the social robot Haru can learn and imitate the strategies of human players as it plays the game with them and will be preferred compared to self-learning with traditional RL method.

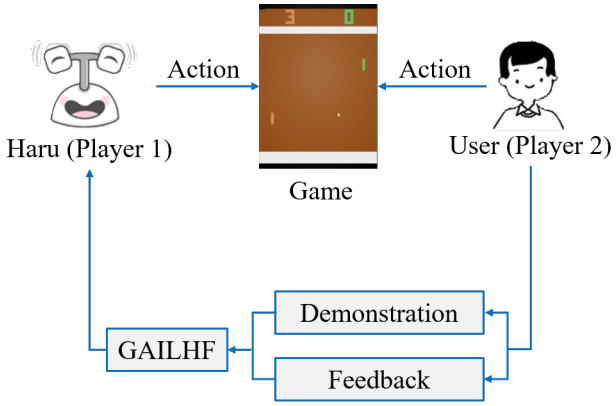


Fig. 1: An illustration of Haru learning game strategies from human demonstrated trajectories and evaluative feedback via our proposed generative adversarial imitation learning from human feedback (GAILHF) method.

[1]School of Information Science and Engineering, Ocean University of China, {guangliangli}@ouc.edu.cn
[2]Honda Research Institute Japan Co., Ltd, Wako, Japan. {r.gomez, e.nichols}@jp.honda-ri.com
* Corresponding author. This work was supported by Natural Science Foundation of China (under grant No. 51809246), Qingdao Municipal Natural Science Foundation (under grant No. 23-2-1-153-zyyd-jch) and Honda Research Institute Japan Co., Ltd.

II. Related Work

A. Reinforcement Learning in Games

For a long history, games have been used as benchmarking domains to drive the progress in the development of artificial intelligence [9]. Tesaruso developed TD-Gammon which is able to teach itself to play backgammon solely by playing against itself via reinforcement learning in 1979 [10]. Mnih et al. proposed deep Q network which can learn successful policies directly from high-dimensional sensory inputs using end-to-end reinforcement learning, and achieve a level comparable to that of a professional human games tester across a set of 49 games in the challenging domain of classic Atari 2600 games [5]. In 2016, Silver et al. [11] proposed AlphaGo using a novel combination of supervised learning from human expert games and reinforcement learning from games of self-play, which defeated a human professional player in the full-sized game of Go. Superhuman performance has also been demonstrated in imperfect information games—Texas hold'em no-limit poker [12]. Schrittwieser et al. [13] proposed MuZero which can master Atari, chess, Go, or shogi by planning with a learned model. In addition, reinforcement learning has also been extended to multi-agent setting to tackle more difficult video games, such as the real-time strategy games StarCraft II [6] and DotA 2 [14]. Although grand-master or superhuman performance was achieved in the above games via reinforcement learning, it still remains a great challenge for them to learn in a human time scale to play these kinds of games because of sparse rewards or complexity. It is difficult or even unpractical to design an efficient reward function for complex and various game tasks, which makes applying traditional deep reinforcement learning methods to real-world robots a great challenge. Moreover, since most robots will operate in human-inhabited environments, to become companions of humans, the ability to interact and learn from human users will be key to their success.

B. Agent Learning from Human

Many methods that allow agents to learn from human demonstrations and/or evaluative feedback instead of predefined reward functions have been proposed [15], [7], [16]. Studies show that learning by interacting with a human teacher can greatly reduce the sample complexity compared to traditional RL [17], [7]. For example, behavioral cloning (BC) [18] allows agents to learn the mapping from states to optimal actions with expert demonstrations. Silva et al. [19] created a controller for a companion team member in a computer role playing game by utilize a case-based behavior engine that employs BC trained offline. However, BC needs large amounts of demonstration data and cannot generalize well to unseen states [20], [21]. An inverse reinforcement learning (inverse RL) agent can learn a policy via reinforcement learning using a cost function extracted from expert demonstrations [22] and allow the learner to generalize expert behaviors to unseen states more effectively [23]. However, most inverse RL algorithms need a model and are expensive to run in large environments. Generative adversarial imitation learning (GAIL) can directly learn policies from the expert demonstrations in a model-free way, and extends inverse RL to large environments. However, in many tasks, optimal demonstrations are hard to obtain practically, and if the demonstrations are suboptimal or far from optimal, GAIL can seldom surpass the performance of demonstrations. Fortunately, an agent via interactive reinforcement learning (interactive RL) from human evaluative feedback can generally surpass the trainer's performance in the task [24], [25], [7]. Ibarz et al. combined imitation learning and learning from trajectory preferences, in which humans compare pairs of short trajectory segments of an agent's behaviour and label those closer to the intended goal to train a reward model that acts as a preference predictor [26]. In addition, Li et al. proposed allowing an agent to learn from human demonstrations first via inverse RL and use evaluative feedback to further improve the agent's performance [25], [27]. To our knowledge, our work is the first to allow a social robot to imitate human game strategies online by learning from demonstrations and evaluative feedback of a human player simultaneously via GAIL.

III. Methodology

In this section, we introduce our generative adversarial interactive imitation learning from human feedback method (GAILHF), which allows social robot Haru to learn and imitate human strategies from demonstrated trajectories and evaluative feedback in a real-time two-player game. Fig. 2 shows the mechanism by which Haru imitates and learns from human players with our method.

As shown in Fig. 2, Haru is set to play a game with a human player. While playing the game, the game trajectories of Haru and the human player are stored to their respective buffers. Meanwhile, to ensure that Haru has the same state and action space as the human player, the coordinates of the game characters' location will be extracted as state represenation of both Haru and human player in the preprocessing. Then, as in the original generative adversarial imitation learning (GAIL) method [28], Haru uses a generator (i.e., policy) to control the game character while playing against a human player, and obtains trajectories τ_i consisting of state-action pairs. Meanwhile, trajectories $\hat{\tau}_i$ of the human player during game playing will be stored and taken as demonstrations. The game trajectories of Haru and the human player in the two buffers will be used to train the discriminator. The discriminator will output a value $D(s',a)$ for the input state-action pair (s',a) and judge whether it is from a demonstration trajectory of the human player or a trajectory by interacting with the game environment. If it is from a demonstrated trajectory of the human player, $D(s',a)$ is expected to be as large as possible, and if it is from a trajectory of Haru, $D(s',a)$ is expected to be as small as possible. The generator is trained using traditional reinforcement learning algorithms — proximal policy optimization (PPO) [29]. The discriminator outputs the value $D(s',a)$ as a reward to update Haru's generator, so that the

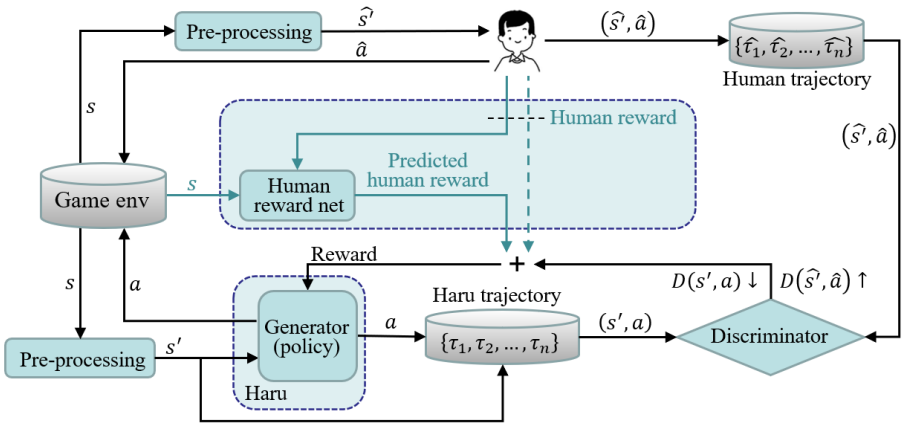


Fig. 2: Illustration of our proposed framework for Haru learning game strategies from human players' demonstrated trajectories and evaluative feedback. Note that, the dotted arrow indicates the provided human evaluative feedback was directly used before the human reward network converges.

distribution of state-action pairs generated by the generator is as close as possible to the distribution of the human player's demonstrated trajectory. Finally, the discriminator will not be able to distinguish the state-action pairs in Haru's trajectory from those in the human demonstrated trajectory.

Different from the original GAIL, in our GAILHF method, the human user can evaluate Haru's behavior by observing it in the game, which will be used as rewards together with the output of the learned discriminator to update Haru's policy. After experimental testing, we chose to allow the human player to provide a $+0.1$ or -0.1 reward when she thinks Haru is performing a good or bad game action respectively, and a 0 reward when the human player does not make an evaluation. The human reward r_h is directly added to the discriminator's output to update the generator in the beginning of the training process.

In addition, to reduce the human player's burden, we trained a human reward network (HRN) to predict human rewards. The human reward r_h together with the labelled state s' of the game character controlled by Haru will be taken as a sample $[s', r_h]$ and stored in a buffer. Then, a mini-batch of samples will be randomly sampled from the replay buffer to update HRN by minimizing the loss function:

$$L_{hrn} = \frac{1}{n}\sum_{i=1}^{n}\left(r_h - H\left(s'\right)\right)^2, \tag{1}$$

where r_h is the human reward, n is the number of samples, and $H\left(s'\right)$ is the trained human reward network. After testing, we found that a minimum number of 200 samples is needed to obtain HRN with high enough prediction accuracy. After that, the predicted human reward by HRN will be added to the discriminator's output to update the generator, instead of the original human evaluative feedback.

IV. Experiments

A. Experimental Platform

We use the social robot Haru [30], [31] as an experimental platform in our studies. Haru is a desktop robot designed to explore the problem of human-robot interaction from a multidisciplinary perspective. Haru consists of two large square eyes and a body base, each using a TFT screen display with a size of 3 inches. Inside the body, there is a matrix of addressable LEDs (the mouth), along with built-in stereo speakers for sound generation. Haru also has five degrees of freedom (i.e., basic rotation, neck tilt, eye movement, eye rotation, and eye tilt), which together with the eyes and mouth can make a range of emotional expressions (see Fig. 3). In our study, Haru will interact with the human player by expressing emotions and speech. For example, when Haru scores or loses points, it will make happy or sad expressions and speak to express its emotion.

Fig. 3: The experimental platform — social robot Haru.

B. Game Task

In our study, we chose the Pong game from Atari, which has been widely used as benchmarking task in the research field of deep reinforcement learning [5], [32]. In addition, we adopted the PettingZoo framework [33] that is widely used in research on multi-agent reinforcement learning to integrate Atari with Haru, so as to enable the human player to play the game with Haru at the same time. In the game, Haru and the human player each control a racket to hit the ball (Haru uses the orange racket on the left and the human player uses the green racket on the right). Each player will move the racket up and down to hit the ball. If the ball goes over the opponent's paddle, the player will receive +1 point. Both players' scores are displayed at the top of the game screen on the player's own side. The player scoring 21 points will win the round (see Fig. 4).

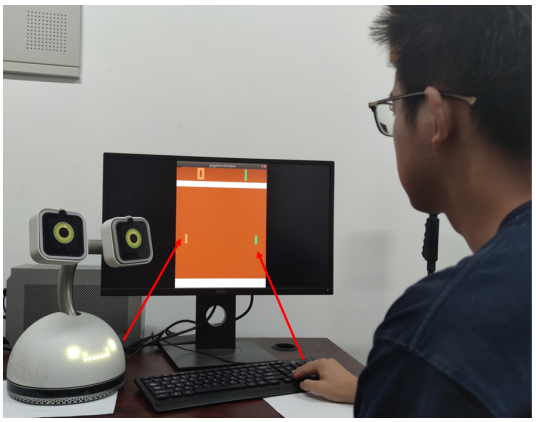


Fig. 4: A human player is playing the Pong game with Haru in our study. Note that the numbers on the top of the game screen are scores of two players.

C. Experimental Setup

1) Preliminary Study: To verify whether Haru can successfully learn strategies for playing the game from a human player's demonstration and evaluative feedback, the first author played the game with Haru using two different strategies, with each strategy for four rounds:

- Strategy 1: The human player controls the racket to move up and down while following the ball's movement, and tries to return to a fixed position each time she serves the ball.
- Strategy 2: The human player does not move the racket with the ball, rapidly serves the ball when she judges the ball is approaching her side, and does not return the racket to the initial position but keeps at the point she serves the ball.

The first author also evaluated Haru's behavior during the training process, which was used as feedback to train a human reward network. The game playing trajectories of the first author were stored. Haru will learn from the saved trajectories as demonstrations and trained human reward network. In addition, to investigate the contribution of the demonstrations and human reward in our GAILHF method, we also trained Haru via the original GAIL method learning from only demonstrated game trajectories as comparison.

Then, to investigate whether Haru can learn from naive users who has little knowledge about the system, we recruited five participants aged from 23 to 27 from our university (3 males and 2 females). Each participant played four rounds of the game with Haru after a brief introduction of the game rule. They can also evaluate Haru's behavior during the training process, which was used as feedback to train a human reward network for predicting human rewards. The number of samples labelled by provided human evaluative feedback for each subject is 200. The game trajectories of each subject were also stored. Then, Haru can learn from the saved trajectories by taking them as demonstrations and trained human reward network.

2) User Study: To test the acceptance of Haru learning to play games via our method, we conducted a larger user study by recruiting 18 participants (10 male, 8 female) aged from 20 to 26 from a university campus (5 are novice master students in robotics but know nothing about the system, 13 are bachelor students who are ignorant of machine learning and robotics). Two experimental conditions were set in our within-subjects study:

- Imitated Haru Condition: Haru plays the game with strategy learned from human demonstration and evaluative feedback via our GAILHF method;
- Self-learned Haru Condition: Haru plays the game with policy learned using PPO [29].

In the Imitated Haru Condition, Haru tries to learn and imitate the human player, while in the Self-learned Haru Condition, Haru self-learned from the original reward function for the game from OpenAI Gym via PPO, in which Haru will receive a $+1$ reward if the ball goes over the opponent's racket and a -1 reward otherwise.

All participants were briefed on the rules of the game, and then each participant played the game with Haru in all two conditions. The order in which each participant played with Haru in the two conditions was randomly assigned. In both conditions, participants were required to play the game for two rounds. All participants filled out a questionnaire at the end of game playing in each condition, and the purpose of the study was revealed at the end of the study. The questionnaire consisted of three sections. The first section with three questions includes assessing whether Haru's actions were human-like or AI-like, the overall evaluation of the entire game process, and willingness to continue playing the game. The second part with four questions is to evaluate Haru's game performance in both conditions. The third section with four questions is to evaluate the human player's game experience. All questions in the questionnaire are on a 7-point scale and are derived from [34], [35].

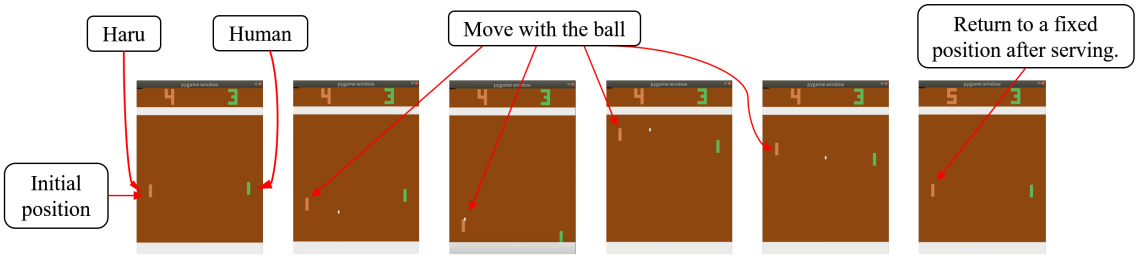
V. Results and Discussion

A. Imitating Strategies of Human Player

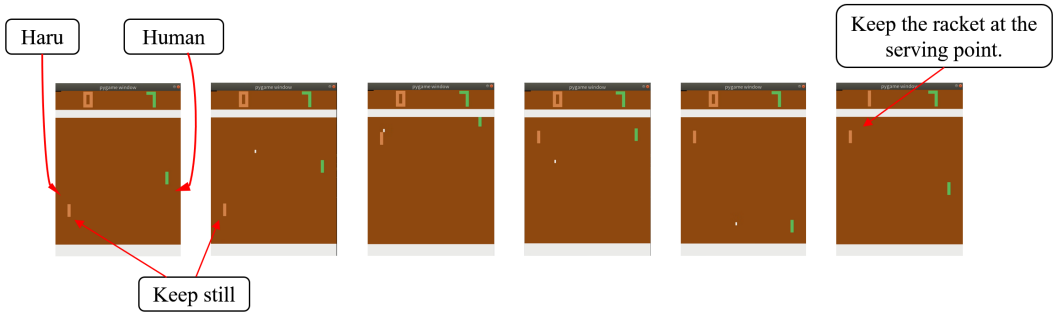
Fig. 5 (a) and (b) show the screenshots of Haru's learned game strategies from playing with the human player via our GAILHF method in the preliminary study. As shown in Fig. 5 (a), Haru successfully imitated human player's Strategy 1 which moved its racket to the predicted position of the ball and returned to the initial position after serving it. Fig. 5 (b) shows that Haru also successfully imitated Strategy 2 from the human player. As we can see, Haru kept the racket still, rapidly moved it to the predicted position of the ball to serve it when the ball is approaching its side and remained the racket at the serving point afterwards.

B. Ablation Study

To investigate the contribution of the demonstrations and human evaluative feedback for Haru's game playing performance with our GAILHF method, we trained Haru via the original GAIL method learning from only demonstrated game trajectories as comparison. The original reward function for the PONG game from OpenAI Gym was used for evaluating the performance of Haru's learning performance



(a) Haru successfully imitated Strategy 1 of the human player



(b) Haru successfully imitated Strategy 2 of the human player

Fig. 5: Screenshots of Haru's successful imitated strategies from the human player in the preliminary study. Haru controls the left racket (orange one) and the human player controls the right racket (green one) to catch the ball (white) served by the opponent.

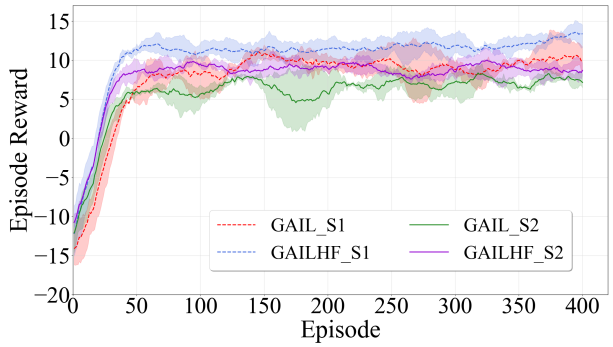


Fig. 6: Haru's learning curves from Strategy 1 (S1) and Strategy 2 (S2) of the human player, via GAIL from demonstrated trajectories and our method GAILHF from both demonstrated trajectories and human evaluative feedback.

via GAILHF and GAIL, but not for learning. The performance metric is the cumulative rewards received within one episode during learning. Fig. 6 shows Haru's learning curves from Strategy 1 (S1) and Strategy 2 (S2) of the human player, via GAIL from only demonstrated trajectories and our method from both demonstrated trajectories and human evaluative feedback, respectively. The shaded area is the 0.95 confidence interval and the bold line is the mean performance over three experimental trials. From Fig. 6 we can see that, Haru can learn a good performance via both GAILHF and GAIL from Strategy 1 and 2 of human players. However, with addition human evaluative feedback, Haru learning via GAILHF can generally obtain a better performance than via GAIL for both strategies.

We also compared the mean game score of Haru trained via GAILHF and GAIL to the demonstration performance of five human players in our preliminary study. A mean score is computed over 50 testing rounds with final learned policies of Haru trained via GAILFH and GAIL. Student's t-test was used to analyze the significance of difference between Haru's learning performance via GAILFH, GAIL and human player's demonstrations. As shown in Table I, our results indicate that there is no significant difference between Haru's score learned from demonstration via GAIL and scores of human player's demonstrations ($t = 0.06, p = 0.96$), which is consistent with that a GAIL agent's learning performance can seldom surpass demonstrations [28]. However, Table I shows that Haru's learning performance via GAILHF is significantly better than that of GAIL ($t = 3.29, p = 0.01$).

TABLE I: Haru's learning performance (mean score in the game) via GAILHF and GAIL, in comparison to the mean score of demonstrations by each subject playing the game in the preliminary study.

Player	Demonstrations	GAIL	GAILHF
P1	-18.00±1.23	-12.78±2.33	-2.65±0.43
P2	-15.50±2.06	-10.63±0.45	-4.87±0.19
P3	-11.25±2.28	-8.15±0.77	4.9±0.50
P4	-3.75±5.54	-7.45±1.76	-4.83±3.50
P5	3.00±3.54	-5.34±2.13	-0.03±1.42

In summary, our experimental results show that Haru learning via our GAILHF method can successfully imitate human player's game strategies. Moreover, while Haru' learning performance via GAIL is limited by human player's demonstrations, our GAILHF method learning from demonstrations and additional humane evaluative feedback facilitates Haru to obtain a better performance than both GAIL and human player's demonstrations in terms of both accumulated rewards and game score.

C. Imitated Haru vs. Self-learned Haru

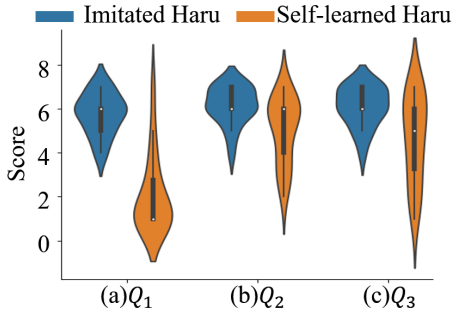


Fig. 7: The imitated Haru was more likely to be perceived as human-like, more satisfactory and preferred for further playing by participants in our user study. Note: the first section of the questionnaire: (a) Q1: Whether Haru's actions were human-like, (b) Q2: Whether the overall experience of playing the game is satisfactory, (c) Q3: Would you like to continue playing games with Haru.

We analyzed the distribution of scores provided by participants in the first section of our designed questionnaire in the user study, as shown in Fig. 7. Student's t-test (Significance level: $p < 0.05$) was used to test the difference over the distribution of scores in the two experimental conditions. Our results show that, imitated Haru by learning from demonstrated game trajectories and evaluative feedback of human player via our GAILHF method was more likely to be perceived as human-like by participants than the self-learned Haru via the traditional reinforcement learning PPO method ($t = 8.01, p < 0.01$). Moreover, participants were also more satisfied with imitated Haru than self-learned Haru ($t = 2.71, p = 0.01$), and preferred to continue playing with imitated Haru learning with our GAILHF method than self-learned Haru with traditional deep reinforcement learning method ($t = 2.79, p < 0.01$).

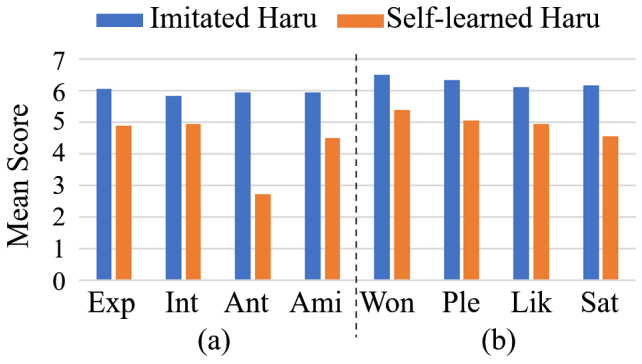


Fig. 8: (a) Imitated Haru was perceived to have better "game performance" but same 'intelligence' as self-learned Haru; (b) participants had significantly better "game experience" with imitated Haru learning via our GAILHF method than with self-learned Haru. Note that: Exp – 'Expression', Int – 'Intelligence', Ant – 'Anthropomorphism', Ami – 'Amity', Won – 'Wonderfulness', Ple – 'Pleasantness', Lik – 'Like' and Sat – 'Satisfaction'.

In addition, we analyzed the mean scores of the four questions in the second and third sections of our designed questionnaire for both imitated Haru and self-learned Haru in the user study, as shown in Fig. 8. In the second part of the questionnaire, both imitated and self-learned Haru were evaluated along four dimensions: Expression, Intelligence, Anthropomorphism, and Amity. In the third section of questionnaire, participants' game experience with both imitated and self-learned Haru was also evaluated in four dimensions: Wonderfulness, Pleasantness, Like, and Satisfaction. Student's t-test was used to analyse the significance of differences between scores of imitated and self-learned Haru, as shown in Table II. From Fig. 8 and Table II we can see that, there was a significant performance difference between imitated Haru learning with our GAILHF method and self-learned Haru with PPO, except in terms of 'Intelligence'. This indicates imitated Haru learning from human demonstrations and evaluative feedback via our GAILHF method was perceived to have the same intelligence level with self-learned Haru with traditional RL method, but was perceived to have significantly more 'Expression', 'Anthropomorphism' and 'Amity' than self-learned Haru. Moreover, our results show that participants had significantly better game experience with imitated Haru learning via our GAILHF method than with self-learned Haru via traditional RL method.

TABLE II: Significance of the differences between imitated Haru and self-learned Haru perceived by participants in terms of "game performance" and "game experience" in our user study. Significance level: t-test, $p < 0.05$.

Game performance		Game Experience	
Exp	t=2.36, p=0.02	Won	t=2.28, p=0.03
Int	t=1.74, p=0.09	Ple	t=2.69, p=0.01
Ant	t=7.02, p<0.01	Lik	t=2.14, p=0.04
Ami	t=2.90, p<0.01	Sat	t=3.17, p<0.01

VI. CONCLUSIONS

In this paper, we proposed a novel method allowing a social robot Haru to imitate human gaming strategies by learning from a human player's demonstrated game trajectories and evaluative feedback. Our experimental results show that Haru is able to learn and imitate different game strategies of human players. Moreover, further results show that human evaluative feedback facilitates Haru to obtain a better performance than learning from only demonstrated trajectories and human player's demonstration performance. Finally, our user study shows that Haru imitating human player's game strategies via our method is perceived to be more human-like and have better game performance and experience than self-learning from pre-defined reward functions via traditional deep reinforcement learning method.

ACKNOWLEDGMENT

We thank Paulo Alvit and Gonçalo Dias from IDMIND for their support and help with our experiment.

REFERENCES

[1] C. Breazeal, "Toward sociable robots," *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 167–175, 2003.

[2] S. Strohkorb, E. Fukuto, N. Warren, C. Taylor, B. Berry, and B. Scassellati, "Improving human-human collaboration between children with a social robot," in *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 551–556, 2016.

[3] G. Gordon, S. Spaulding, J. K. Westlund, J. J. Lee, L. Plummer, M. Martinez, M. Das, and C. Breazeal, "Affective personalization of a social robot tutor for children's second language skills," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 30, pp. 3951–3957, 2016.

[4] S. Gillet, W. van den Bos, and I. Leite, "A social robot mediator to foster collaboration and inclusion among children.," in *Robotics: Science and Systems*, pp. 1–9, 2020.

[5] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[6] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[7] G. Li, R. Gomez, K. Nakamura, and B. He, "Human-centered reinforcement learning: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 4, pp. 337–349, 2019.

[8] J. Lin, Z. Ma, R. Gomez, K. Nakamura, B. He, and G. Li, "A review on interactive reinforcement learning from human social feedback," *IEEE Access*, vol. 8, pp. 120757–120765, 2020.

[9] P. R. Wurman, P. Stone, and M. Spranger, "Improving artificial intelligence with games," *Science*, vol. 381, no. 6654, pp. 147–148, 2023.

[10] G. Tesauro, "Td-gammon, a self-teaching backgammon program, achieves master-level play," *Neural Computation*, vol. 6, no. 2, pp. 215–219, 1994.

[11] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[12] N. Brown and T. Sandholm, "Superhuman ai for multiplayer poker," *Science*, vol. 365, no. 6456, pp. 885–890, 2019.

[13] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, *et al.*, "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[14] M. Hutson, "Ai takes on video games in quest for common sense," 2018.

[15] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009.

[16] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 297–330, 2020.

[17] W. B. Knox and P. Stone, "Combining manual feedback with subsequent mdp reward signals for reinforcement learning.," in *Proceedings of Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 5–12, 2010.

[18] S. Ross and D. Bagnell, "Efficient reductions for imitation learning," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 661–668, JMLR Workshop and Conference Proceedings, 2010.

[19] M. Silva, S. McCroskey, J. Rubin, M. Youngblood, and A. Ram, "Learning from demonstration to be a good team member in a role playing game," in *Proceedings of the 26th International FLAIRS Conference*, 2013.

[20] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Proceedings of Robotics: Science and Systems (RSS)*, 2018.

[21] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7559–7566, IEEE, 2018.

[22] A. Y. Ng, S. Russell, *et al.*, "Algorithms for inverse reinforcement learning.," in *Proceedings of International Conference on Machine Learning (ICML)*, vol. 1, p. 2, 2000.

[23] J. Ho, J. Gupta, and S. Ermon, "Model-free imitation learning with policy optimization," in *Proceedings of International Conference on Machine Learning (ICML)*, pp. 2760–2769, 2016.

[24] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone, "Deep TAMER: Interactive agent shaping in high-dimensional state spaces," in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 1545–1553, 2018.

[25] G. Li, B. He, R. Gomez, and K. Nakamura, "Interactive reinforcement learning from demonstration and human evaluative feedback," in *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 1156–1162, 2018.

[26] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in atari," in *Proceedings of Advances in Neural Information Processing Systems*, pp. 8011–8023, 2018.

[27] J. Huang, J. Hao, R. Juan, R. Gomez, K. Nakamura, and G. Li, "Gan-based interactive reinforcement learning from demonstration and human evaluative feedback," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4991–4998, IEEE, 2023.

[28] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Proceedings of Advances in Neural Information Processing Systems*, 2016.

[29] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[30] R. Gomez, D. Szapiro, K. Galindo, and K. Nakamura, "Haru: hardware design of an experimental tabletop robot assistant," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 233–240, IEEE, 2018.

[31] Y. Vasylkiv, Z. Ma, G. Li, H. Brock, K. Nakamura, I. Pourang, and R. Gomezv, "Shaping affective robot haru's reactive response," in *Proceedings of the 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 989–996, IEEE, 2021.

[32] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, *et al.*, "Model-based reinforcement learning for atari," *arXiv preprint arXiv:1903.00374*, 2019.

[33] J. Terry, B. Black, N. Grammel, M. Jayakumar, A. Hari, R. Sullivan, L. S. Santos, C. Dieffendahl, C. Horsch, R. Perez-Vicente, *et al.*, "Pettingzoo: Gym for multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15032–15043, 2021.

[34] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[35] Z. Ashktorab, Q. V. Liao, C. Dugan, J. Johnson, Q. Pan, W. Zhang, S. Kumaravel, and M. Campbell, "Human-ai collaboration in a cooperative game setting: Measuring social perception and outcomes," in *Proceedings of the ACM on Human-Computer Interaction*, pp. 1–20, ACM New York, NY, USA, 2020.