

“© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

Weak Human Preference Supervision For Deep Reinforcement Learning

Zehong Cao, *Member, IEEE*, KaiChiu Wong, Chin-Teng Lin, *Fellow, IEEE*

Abstract—The current reward learning from human preferences could be used to resolve complex reinforcement learning (RL) tasks without access to a reward function by defining a single fixed preference between pairs of trajectory segments. However, the judgement of preferences between trajectories is not dynamic and still requires human input over thousands of iterations. In this study, we proposed a weak human preference supervision framework, for which we developed a human preference scaling model that naturally reflects the human perception of the degree of weak choices between trajectories and established a human-demonstration estimator via supervised learning to generate the predicted preferences for reducing the number of human inputs. The proposed weak human preference supervision framework can effectively solve complex RL tasks and achieve higher cumulative rewards in simulated robot locomotion – MuJoCo games – relative to the single fixed human preferences. Furthermore, our established human-demonstration estimator requires human feedback only for less than 0.01% of the agent’s interactions with the environment and significantly reduces the cost of human inputs by up to 30% compared with the existing approaches. To present the flexibility of our approach, we released a video (<https://youtu.be/jQPe1OILT0M>) showing comparisons of the behaviours of agents trained on different types of human input. We believe that our naturally inspired human preferences with weakly supervised learning are beneficial for precise reward learning and can be applied to state-of-the-art RL systems, such as human-autonomy teaming systems.

Keywords—Deep Reinforcement Learning, Weak Human Preferences, Scaling, Supervised Learning.

I. INTRODUCTION

Reinforcement learning (RL) [1] typically uses a reward function to train an agent’s behaviours for a specified task. Nevertheless, constructing an effective reward function in complicated scenarios can often be challenging. If the design of a reward function is too simple, then the behaviours of the trained agent may not match our expectations, i.e., the results may exhibit misalignment between our expectations and the actual testing [2]. To achieve more effective RL, having a communication pathway between the RL agent and our expectations during the training process is valuable [3].

In autonomous vehicle research, robotic controls with RL have already demonstrated tremendous potential in improving transportation systems for related problems such as localisation, path planning, and collision avoidance [4] [5]. Investigating human-autonomy teaming using RL can enable the

Z. Cao is with the School of Information and Communications Technology (ICT), University of Tasmania, Australia. (E-mail: zehong.cao@utas.edu.au.)

K. Wong was with the University of Tasmania and is now with the MyState Bank, Australia.

C.T. Lin is with the Australian Artificial Intelligence Institute (AAIL) and the School of Computer Science, University of Technology Sydney, Australia.

Manuscript received on 27 July, 2020; revised on 06 December, 2020.

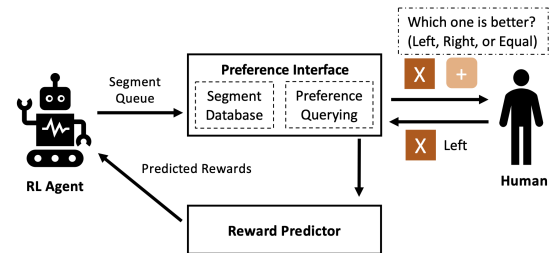


Fig. 1: Illustration of Deep RL from Human Preferences

development of a reliable reward function in the human-robot interaction process. Some recent work showed challenges in training an intelligent robot to complete task objectives [6] [7] [8] and in multi-agent interactions [9] [10], addressing the difficulties of alignment between human expectations and the final training outcome. Although approaches such as inverse RL [11] and imitation learning [12] were suggested to extract the reward function and mimic the actions of human experts to ensure an expected outcome, these approaches are not direct enough to train the desired behaviour. Moreover, the degree of movement of a robot could be larger than that of a human, as a human demonstration for imitation learning may not be available in some cases [13].

A novel study proposed by the DeepMind group [14] first addressed deep RL from human preferences, in which a comparison measurement between pairs of trajectory segments was designed to replace the reward function and preference inputs from an expert were allowed, as shown in Figure 1. This approach requests advice from an expert to ensure that the RL training is on the correct track, implying that agents could be assisted when tackling problems in some highly complicated scenarios. However, the judgement of preferences between trajectories is not dynamic. For example, the candidate preferences include only fixed left, right, or equal options, represented by the preference values 1, 0, and 0.5, respectively; thus, we assume that this approach cannot reflect natural human intentions. In addition, the current approach still requires human input over thousands of iterations, which requires a substantially large amount of time from humans.

Inspired by the above DeepMind study and the uncertainties in decision making [14] [15], in this study, we propose a new framework: weak human preference supervision for deep RL. In particular, we develop a scaling model to support dynamic and weak human preferences, instead of single fixed preferences, for RL. Moreover, we use a database of human

preference scaling values to establish a human-demonstration estimator via supervised learning to predict the preference scaling values based on initial human inputs to reduce the amount of human effort.

The contributions of our weak human preference supervision framework for RL are as follows:

- 1) Our proposed weak human preference supervision for deep RL allows humans to input dynamic and weak preference levels via our developed human preference scaling model to reflect human behaviour and decrease the number of human inputs for our established human-demonstration estimator.
- 2) Based on 5 experiments with the robotic physical simulator MuJoCo [16], our developed human preference scaling model for RL can achieve higher cumulative reward values than those of the current fixed human preference model, and our established human-demonstration estimator support for RL can reduce the amount of human input for dynamic and weak preferences by up to 30% without significantly sacrificing the reward values.

The rest of this paper is organised as follows. The related work is briefly introduced in Section II. Then, the preliminaries and our proposed framework are illustrated in Section III. Afterwards, experiments involving MuJoCo games and the relevant settings are addressed in Section IV. Finally, we present our findings and comparison results in Section V and conclude this work in Section VI.

II. RELATED WORK

A. Preference Learning

An agent may often not be able to learn expected actions from rewards when using the traditional RL strategy, but preference learning can potentially minimise the gap between the missing agent information and the behaviour desired by humans. A recent study [17] stated that preference learning could facilitate the adaptation of robotic movement and noted that the human demonstration is particularly easy for task-dependent goals. An earlier study [18] also showed that human preferences are more effective in acquiring better actions than are agent rewards in RL. They implied that the feedback from a human could leverage RL in qualitative policy models and that manipulation from preference learning could be a new strategy to assist agents in achieving the desired human behaviour in robotic control.

Additionally, [19] provided a human behaviour decision model to increase the accuracy in preference learning, but it focused only on adjusting the attitude and the importance of relative criteria without improving the human preference coverage. Another recent work [20] developed a DemPref model to provide efficient preference-based learning in robotic movements, but this approach requires collecting a large number of purely human-demonstrated actions prior to training.

B. Learning by Pairwise Comparisons

Learning by pairwise comparisons allows the application of human preferences to RL. Frnkranz's work [18] revealed that

the preferences could help an agent carry out label ranking in RL so that humans could provide feedback to the agent. To distinguish between preference learning and RL, a survey study [21] clarified that an agent can receive feedback from two options in preference learning, while in RL, only one human feedback option is accepted. This restriction facilitates RL agent label ranking when performing a task.

$z_i \succ z_j$ indicates that z_i is the preferred choice over z_j , where Z represents a set of options. Kreps [22] offered some notations that could be defined from preference learning while comparing a pair of options:

- $z_i \succ z_j$: option z_i is *absolutely preferred*.
- $z_i \prec z_j$: option z_j is *absolutely preferred*.
- $z_i \sim z_j$: options z_i and z_j are *the same*, as no preferred option can be distinguished.
- $z_i \succeq z_j$: option z_i is *weakly preferred*.
- $z_i \preceq z_j$: option z_j is *weakly preferred*.

C. RL from Human Preferences

To link preference learning to RL, according to the communication pathway proposed by Christiano [14], the basic flow of RL from human preferences comprises three main modules: agent, preference interface, and reward predictor. The agent continually trains and explores the environment as RL progresses. The novelty of the communication pathway is the addition of a preference interface, which randomly generates some episodes in which human judgement is requested. An expert therefore inputs which options he/she prefers by inputting the preference database into the preference interface. These preferences are sent to a reward predictor to perform further training to assess the relative rewards from these preferences. These data are then used return the trained preferences as the agent's observations to perform the policy training.

D. Motivations

The settings of Christiano's study [14] cover only three conditions among the definitions of preference learning from Kreps [22], i.e., $z_i \succ z_j$, $z_i \prec z_j$ and $z_i \sim z_j$. Christiano's study did not consider $z_i \succeq z_j$ and $z_i \preceq z_j$, which are two weak but important conditions we encounter in real-life decision making based on preferences. This lack motivates us to develop a new RL-based human preference setup to cover the weak preference conditions $z_i \succeq z_j$ and $z_i \preceq z_j$.

Furthermore, Christiano's study [14] required a large number of human preference labels, which may require a long period of time dedicated to human inputs. In particular, at least 5,500 human inputs are required in the training setup. We assume that some human preference labels may not be accurate, as a human may experience a high level of fatigue, potentially resulting in an increase in the error rate after a long period of performing the preference judgement task. This problem motivates us to develop a human-demonstration estimator via supervised learning to reduce the number of human inputs and improve learning performance.

III. PRELIMINARIES AND METHODS

A. Settings and Goal

An RL agent interacts with the environment for a number of steps, during which the agent inspects the environment according to the observation $o_t \in O$, receives an instant reward r at each timestep t and then performs action $a_t \in A$ based on the observation. The agent aims to maximise the cumulative rewards during the training and receives the predicted instant reward values during each timestep $r_t \in R$ via the reward predictor and preference interface modules. When the agent takes an action in the environment, a video clip from a trajectory segment is pushed into the list in the preference interface. The preference interface module randomly draws two video clips from the list and asks a human which clip he/she would prefer. Once the human inputs his/her preference, this information is passed to the reward predictor to generate predicted rewards for the agent for training.

In this study, we follow the preference interface design from Christiano's work [14], which accepts the generated segments from the RL agent and places them into a queue, as shown in Figure 1. Two segments are randomly selected from the queue, and the preference interface asks a human for his/her preference between the two candidate options. After the preference interface collects the preferred option from the human, this option is saved in the preference queue of the reward predictor for the later training of RL policies. Because the currently used approach accepts only a fixed preference, i.e., left, right or equal, as its input, it lacks dynamic inputs of the preferred range. In this study, we modify the preference interface design to consider weak human preference conditions and propose a synthetic preference scaling model, as shown in the following section.

B. Weak Human Preferences: Synthetic Preference Scaling

In this study, we assume that the human always chooses the trajectory segment that has the most potential to return a high reward value; thus, we use the synthetic oracle, a Bayesian approach for policy learning from trajectory preference queries, to mimic the preference of the human, whose preference over several trajectories precisely reflects the reward [23]. For the synthetic human preference, when the agent queries for comparisons, the synthetic human can immediately reply by indicating a preference for whichever trajectory segment yields a higher reward in the underlying task.

Based on the synthetic human preference [23], we develop a weak human preference model called synthetic preference scaling, as presented in Algorithm 1.

We elaborate on the proposed Algorithm 1 in the following. Because the range of z is set as $[0.0, 1.0]$, a z value of 1.0 means that the human absolutely prefers the left trajectory segment, while a z value of 0.0 means that the human absolutely prefers the right trajectory segment. A z value of 0.5 indicates that the human cannot judge between the two trajectory segments. For preference scaling, the reward list R is collected in the memory in each iteration, and the synthetic preferences are calculated based on the reward values with a normalisation measurement. In particular, based on the 90%

Algorithm 1 Weak Human Preferences: Synthetic Preference Scaling

Require: synthetic preference scaling represented as \hat{z} .

Require: \hat{z} from a given reward set (R_{left}, R_{right}) in the range $[0.0, 1.0]$, where equal preference corresponds to a value of 0.5.

Input: R_{left} , the reward of the left trajectory segment.

Input: R_{right} , the reward of the right trajectory segment.

Input: R , the reward list composed of all n reward sets.

Output: \hat{z}

- 1: $R \leftarrow [R_{left_1}, R_{right_1}, R_{left_2}, R_{right_2}, \dots, R_{left_n}, R_{right_n}]$
 - 2: $\text{sort}(R)$, sort the list, in ascending order by default.
 - 3: $N \leftarrow$ number of elements in R .
 - 4: $R_{min} \leftarrow (\lceil \frac{10}{100} \times N \rceil)^{th}$ element from R .
 - 5: $R_{max} \leftarrow (\lceil \frac{90}{100} \times N \rceil)^{th}$ element from R .
 - 6: **if** $(R_{left} > R_{right})$ **then**
 - 7: normalise $\hat{R}_{left} \leftarrow \max\left(0, \min\left(\frac{R_{left} - R_{min}}{R_{max} - R_{min}}, 1\right)\right)$
 - 8: $\hat{z} \leftarrow 0.5 + 0.5 \times \hat{R}_{left}$
 - 9: **else if** $(R_{left} < R_{right})$ **then**
 - 10: normalise $\hat{R}_{right} \leftarrow \max\left(0, \min\left(\frac{R_{right} - R_{min}}{R_{max} - R_{min}}, 1\right)\right)$
 - 11: $\hat{z} \leftarrow 0.5 - 0.5 \times \hat{R}_{right}$
 - 12: **else**
 - 13: $\hat{z} \leftarrow 0.5$
 - 14: **end if**
 - 15: **return** \hat{z}
-

confidence interval level, from the sets in the reward list R , we first remove the lowest 10% and highest 10% of reward values to ensure that outliers or anomalous reward values will not affect the preference scaling calculation. Then, all collected reward values are normalised to values between 0.0 and 1.0, and all sets in the reward list R are ranked in ascending order.

C. Human Preference Scaling for Deep RL

Following the basic preference settings for RL as addressed in Section III-A, at each time t , we maintain the policy $\pi : O \rightarrow A$, where the agent interacts with the environment according to observation $o_t \in O$ and then performs particular action $a_t \in A$ based on instant observation o_t . During the training, the agent tries to estimate the reward function $\hat{r} : O \times A \rightarrow R$ from a deep neural network, which is updated as follows:

- (1) A set of trajectories $\{\gamma^1, \gamma^2, \dots, \gamma^i\}$ is generated by policy π . The parameters of policy π are updated by the traditional RL to ensure that the maximum sum of predicted rewards $r_t = \hat{r}(o_t, a_t)$ that could be achieved from observation o and action a is obtained.
- (2) A pair of segments (σ^1, σ^2) is randomly selected from a set of trajectories $\{\gamma^1, \gamma^2, \dots, \gamma^i\}$. This pair of segments is sent to the preference interface and allows the human to perform the comparison.
- (3) Based on Algorithm 1, the preference scaling z is collected from the human and linked to the pair of segments (σ^1, σ^2) .

Please note that the above updating processes occur in the asynchronous mode: process (1) passes the trajectories $\{\gamma^1, \gamma^2, \dots, \gamma^i\}$ to process (2), process (2) passes the human preferences to process (3), and process (3) passes the parameters of \hat{r} back to process (1).

Preference Elicitation: In Figure 2-A, we show a human preference scaling structure for the RL agent to make these processes easily understandable. To reflect natural human intent, we modify the format of human preferences from fixed preferences to scale-based preferences in process (3). In terms of the current fixed preferences for a pair of segments [14] (as shown in Figure 1), this approach allows inputting only the left, right or equal option, and the judgement is saved in the preference database D with the data format (σ^1, σ^2, z) , where σ^1 and σ^2 are the extracted paired segments and $z = [0, 0.5, 1]$ is the fixed preference input from the human. Thus, the approach does not provide information regarding how much a human prefers a particular segment. For example, the left segment could be better than the right segment, but the left one may not be perfect.

Our proposed scale-based preferences provide a scaling model (as shown in Algorithm 1) with which the human can input a dynamic score for the preferred segment by assigning any value between 0.0 and 1.0. The value of \hat{z} could be in the range of $\{0.0, 1.0\}$ to specify the dynamic judgement of the human. If \hat{z} is input as 0.0, then the condition of σ^1 is absolutely preferred, while if \hat{z} is input as 1.0, then the condition of σ^2 is absolutely preferred. A \hat{z} value of 0.5 indicates that σ^1 and σ^2 are not different; hence, the above conditions do not hold. Then, we can further address the additional conditions, which are in the weakly preferred categories. To be more specific, any values between 0.0 and 0.5 (excluding the margin values 0.0 and 0.5) indicate that σ^1 is weakly preferred, while any values between 0.5 and 1.0 (excluding the margin values 0.5 and 1.0) indicate that σ^2 is weakly preferred. For example, a human could input a preference value of 0.87 or any value he/she likes in the range between 0.0 and 1.0 to specify that they weakly prefer the left or right segment. Our contribution aims to supply more accurate information regarding the human preferences to the RL agent, not simply indicating whether the chosen option is better than the other but rather indicating how much better it is.

Fitting the Reward Function: If the reward function estimate \hat{r} is the predicted reward from the reward predictor, as shown in Figure 2-A, then we consider \hat{r} as a latent factor explaining the human judgements and assume that the human probability of preferring a segment σ depends exponentially on the value of the latent reward summed over the length of the clip, which follows Christiano’s design process [14], as follows.

$$\hat{P}[\sigma^1 \succ \sigma^2] = \frac{\exp \sum \hat{r}(o_t^1, a_t^1)}{\exp \sum \hat{r}(o_t^1, a_t^1) + \exp \sum \hat{r}(o_t^2, a_t^2)}$$

\hat{r} minimises the cross-entropy loss between these predictions and the actual labels of human inputs:

$$\begin{aligned} \text{loss}(\hat{r}) = & \\ - \sum_{(\sigma^1, \sigma^2, \hat{z}) \in \mathcal{D}} & \hat{z}(1) \log \hat{P}[\sigma^1 \succ \sigma^2] + \hat{z}(2) \log \hat{P}[\sigma^2 \succ \sigma^1] \end{aligned}$$

Optimising the Policy: After the reward function \hat{r} computes the rewards, we can meet the need for traditional RL. As the reward function, \hat{r} may be non-stationary, which leads us to prefer RL algorithms that are robust to changes in the reward function, such as policy gradient methods [24]. In this study, we use proximal policy optimisation (PPO) [25] to perform simulated robotics tasks and apply the same parameter settings as in Christiano’s work [14].

D. Human Preference Supervision: Human-Demonstration Estimator for Deep RL

Because a large number of human preference scaling values (generally over 1,000 preference inputs) must be stored in the preference database to train the fitting of predicted rewards, we assume that reducing the number of human preference inputs by training a preference estimator is worthwhile. As shown in Figure 2-B, we develop a preference estimator based on the previous human inputs and use a regression model with supervised learning that we call the human-demonstration estimator to predict some of the human preferences. We expect this human-demonstration estimator to not only reduce the number of human inputs n but also maintain good performance without sacrificing the cumulative rewards.

The human-demonstration estimator is an extended version of the previous human preference scaling for deep RL presented in Section III-C. In particular, the collected database includes n -fold preference scalings of a preference estimator based on previous human inputs, and a regression model with supervised learning that we call the human-demonstration estimator is used to predict some of the human preferences with the data format $(\sigma^1, \sigma^2, \hat{z})$, as discussed earlier. Specifically, the initial preferences are collected at the initial stage and stored in the preference queue. These preferences are used as the base of the preference estimation. In our study, two reliable supervised learning models, i.e., linear regression and support vector regression (SVR) with a radial basis function (RBF) kernel, are considered as the estimator to predict the human preference based on the collected initial preferences. To construct a prediction model and fit the parameters of the human preference scaling estimator, the database is separated into two parts, i.e., a training dataset and a testing dataset, and we implement two types of data splitting: i) 50% training data and 50% testing data and ii) 70% training data and 30% testing data. As a result, 30-50% of the human inputs will be replaced by the agent’s estimation.

With linear regression, the objective function for ordinary least squares with one preferred segment σ in the set (σ^1, σ^2) is as follows:

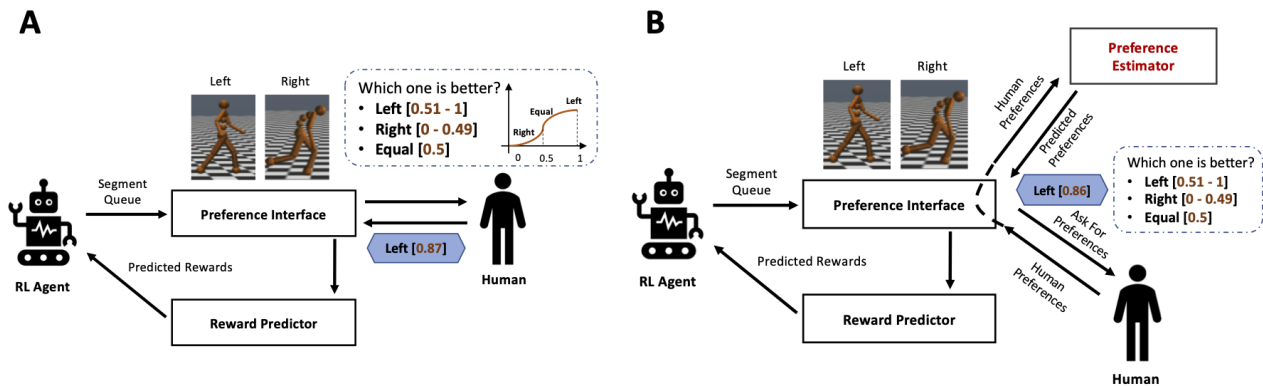


Fig. 2: Proposed (A) Human Preference Scaling Model and (B) Human-Demonstration Model for Deep RL

$$\text{Min} \sum_{i=1}^n (\hat{z}'_i - w_i \sigma_i)^2$$

where \hat{z}'_i is the estimated preference scaling value and w_i is the coefficient.

SVR gives us the flexibility to define errors and finds an appropriate hyperplane in higher dimensions to fit the data. The objective function of SVR is to minimise the coefficients: the l_2 -norm of the coefficient vector. The error term is instead handled in the constraints, where we set the absolute error less than or equal to a specified margin, called the maximum error ε . We can tune ε to gain the desired accuracy of our model. The updated objective function of SVR and the constraints are as follows:

Minimise:

$$\text{Min} \frac{1}{2} \|\mathbf{w}\|^2$$

Constraints:

$$|\hat{z}'_i - w_i \sigma_i| \leq \varepsilon$$

where \hat{z}'_i is the estimated preference scaling value and w_i and σ_i are the coefficients and the preferred segment, respectively.

To evaluate the performance of the estimators, the mean squared error (MSE) is applied to measure the average of the squares of the errors, i.e., the average squared difference between the estimated preference scaling values \hat{z}' and the ground truth of preference scaling \hat{z} .

IV. EXPERIMENT

We implemented the existing models and our proposed models for deep RL and performed experiments in 5 scenarios from MuJoCo [26] with TensorFlow [27] under the OpenAI Gym platform [28]. The collected results were consolidated under the TensorBoard package from TensorFlow.

A. Robotic Control Scenarios

OpenAI Gym provides baseline environments to train the agent on the RL algorithms [28]. MuJoCo is one of the popular continuous control tasks in OpenAI Gym, with a

physics engine that can be used to simulate model-based control [16]. MuJoCo [26] contains diverse scenarios with robot control, where the agent moves different joints with continuous control instead of intermittent control to achieve a particular goal [29]. The agent performs different types of actions to achieve the maximum cumulative reward value to reach the target goal. This process could be challenging, as the MuJoCo environment involves high exploration dimensions for the agent.

As shown in Table I, in this study, our testing environments include 5 scenarios: Walker, Hopper, Swimmer, Cheetah, and Ant. The existing approaches, such as traditional RL (PPO) and RL from human preferences (RLHP), and our proposed weak human preference supervision framework composed of two models, i.e., 1) weak human preferences: RL from human preference scaling (RLHPS) and 2) human preference supervision: RL from human preference scaling with demonstrations (RLHPS with Demo), are applied to compare their performances in these 5 scenarios.

B. Parameter Settings

Generally, in the policy gradient strategy, e.g., PPO, the agent starts with the initial policy, interacts with the environment, obtains a predicted reward from human feedback or from using the pre-defined reward function, and then uses the reward to improve the policy. Here, we need to know how many transitions (sequences of states, rewards, and actions) the agent should gather before updating the policy and how to use the transitions for updating under the new policy.

First, we need to address the collection of experience (horizon, mini-batches, and epochs) prior to updating the policy. For example, PPO collects trajectories up to the time horizon (T) limit and then performs a minimum batch size stochastic gradient descent (SGD) update on all collected trajectories within a specified epoch. Second, to update the new policy from the old policy, PPO uses a surrogate loss function to keep the step from the old policy to the new policy within a safe range; here, consideration of the discount factor γ and the GAE parameter λ is necessary. In addition, the remaining parameters are general hyperparameters that can be used in many deep learning experiments to determine, e.g., the

TABLE I: List of 5 MuJoCo Scenarios

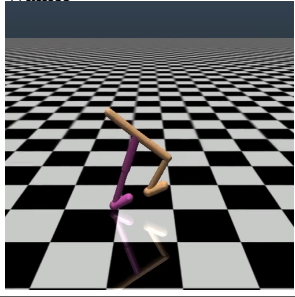
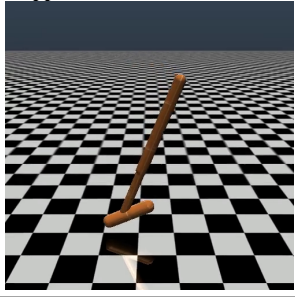
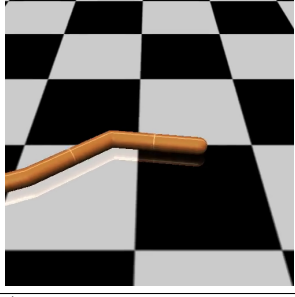
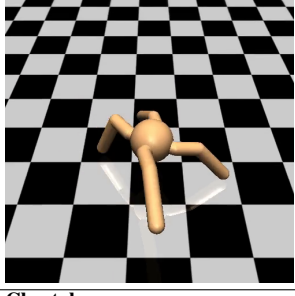
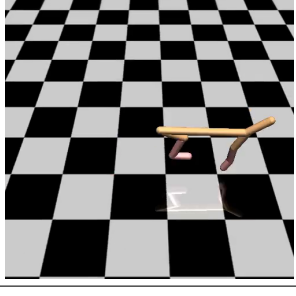
MuJoCo Scenario	Observation	Task Summary
Walker 	(18, 6, 24)	A planar walker tries to roll forward and walk as quickly as possible. The reward depends on the velocity v and the torso height h .
Hopper 	(14, 4, 15)	A one-legged robot is required to move forward and attain as high a torso height as possible. The reward depends on the velocity v and the torso height h .
Swimmer 	(10, 2, 13)	A robot tries to reach a random target by swimming. The reward is given when the nose of the robot touches the random target.
Ant 	(29, 8, 67)	A 4-legged robot attempts to learn to walk as quickly as possible. The reward is based on the velocity v and the body height h .
Cheetah 	(18, 6, 17)	A robot has to learn to move forward as quickly as possible. The reward R is based on the velocity v , the formula for which is $R(v) = \max(0, \min(\frac{v}{10}, 1))$.

TABLE II: Hyperparameters of PPO used in MuJoCo Scenarios

Hyperparameter	Value
Horizon (T)	2048
Mini-batch size	64
Number of epochs	10
γ	0.99
GAE parameter (λ)	0.95
Adam step size	3×10^{-4}
Learning rate	1×10^{-4}
Number of steps	3×10^7
Number of hidden units	64

learning rate, number of steps, and number of hidden units. In this study, the experiments in the MuJoCo scenarios are trained for 3×10^7 timesteps over 10 iterations. All MuJoCo scenarios are trained under the PPO strategy with or without human feedback with the hyperparameters given in Table II.

C. Baselines

Traditional RL: The baseline of each scenario is training with the traditional RL (PPO) without any human involvement. The agent has to learn, based on the scenario goal, only from the rewards they receive. The learning performance is the same as that of the traditional RL process and relies on the design of the reward function of each scenario. The details of the reward design of each scenario are specified in Table I. Our goal for this setup is to set the baseline algorithm and evaluate the performance in each scenario to check what reward values could be achieved without human input.

RL from Human Preferences (RLHP): We consider RLHP as another baseline that contains the basic human preferences to replicate the results under advice from a human [14]. The experimental setup emphasises a preference interface to ask for human preferences, and the user interface shows the rewards and the video clip information to let each human input a large number of preferences. The interface allows the user to input either a left, right or equal option into the preference interface.

V. RESULTS

In this study, 5 scenarios (Walker, Hopper, Swimmer, Cheetah and Ant) are used as our experimental environments to evaluate the training performance among 4 types of RL model by comparing the traditional RL baselines (without human preferences), i.e., PPO and RLHP, with our two proposed models, i.e., RLHPS and RLHPS with Demo. The RL algorithms involved in this study are summarised in Table III, where human preferences require an input of 700 to 1,400 labels, which correspond to less than 0.01% of the training timesteps.

A. RL from Human Preference Scaling (RLHPS)

Based on the experiments in the 5 MuJoCo scenarios, the performance in terms of the cumulative reward values of the two baselines, i.e., the traditional RL (PPO) and RLHP, and our proposed approach (RLHPS) are compared, as shown in Figure 3, which presents the training of our agent by learning

TABLE III: Summary of the RL Algorithms

RL Type	Algorithm	No. of Inputs (1)	No. of Inputs (2)	Note
RL without Human Inputs	Traditional RL (PPO)	N/A	N/A	
RL with Human Inputs (Preferences)	RLHP	1,400	700	RL from human preferences.
	RLHPS (ours)	1,400	700	RL from human preference scaling.
	RLHPS with Demo (ours)	980	700	RL from human preference scaling with demonstrations; 30-50% of the 1,400 inputs (420-700 inputs) are generated from the estimated preferences.

from two types of human preference input (700 and 1,400 labels, amounting to less than 0.01% of the training timesteps) for RLHP and our proposed RLHPS. The cumulative reward values from our proposed RLHPS are always higher than those from RLHP or PPO, except in the Hopper scenario, where all the RL algorithms achieve similar reward values after 2×10^7 timesteps. From another perspective, the use of 1,400 human input labels generally yields higher rewards than obtained from 700 human input labels, suggesting that more human effort may benefit the training of a robust RL agent.

Cumulative Reward Values: Particularly in the Walker scenario, our proposed RLHPS can achieve an approximately 1,500 higher reward value than the traditional RL PPO setup and an approximately 1,000 higher reward value than RLHP for the case of 1,400 human preference inputs. For the Swimmer scenario, the reward learning from RLHPS with the 1,400-label setup can achieve an approximately 350 reward value at the end of the experiment, while RLHP with the 1,400-label setup can achieve a reward value of only approximately 300 at the end of the experiment. Both human preference setups (RLHP and RLHPS) are much better than the traditional RL setup (PPO), as PPO achieves a reward value of only 150. In terms of the Cheetah scenario, our proposed RLHPS can also acquire higher rewards, a reward value of approximately 4,000, compared to the range of 3,000-3,500 when trained by RLHP and PPO.

Regarding the special case of the Ant scenario, complete three-dimensional movement and observation are required for the robot to achieve learning. Continuous movement of the agent is very difficult to achieve, as the ant robot has to find a way to balance its body and walk. From our observation, during the initial period of training, the robot does not know how to walk and balance and always flips over, which causes the reward values in this period to always be negative. The experimental results show that the scale-based preferences, highlighted in red, could yield a good performance compared to the fixed preferences. This outcome is good evidence that our proposed RLHPS can attain higher rewards (either with 700 or 1,400 labels) and quickly achieve positive reward values compared with RLHP or PPO.

Instant Reward Distributions: For our proposed RLHPS, we also investigated the instant reward distributions between the initial and final training periods in the 5 MuJoCo scenarios. The initial training period is that in which the human has not input a preference label, generally within the first 300 timesteps. The final training period includes the final 2,500-

3,000 timesteps, during which the human inputs 1,400 preference labels and the RL agent is approaching completion of the training process. This scenario was considered because the experimental setup requires use of the initial stage (the initial training period) to generate segments for the segment queue for the preference interface so that the agent can perform the initial RL training before the preference interface can select some segments from which the human can choose. **We also intend to replicate the performance from Christiano’s work [14] as our baseline; thus, our proposed method uses exactly the same settings of hyperparameters in PPO to contrast with the existing results.**

As shown in Figure 4, the instant reward distributions are sampled for these two training periods (initial vs. final) in the 5 MuJoCo scenarios. The blue data distribution represents the beginning of training before acquiring any preferences; hence, most of the instant rewards are still negative. The grey data distribution shows the situation at the end of training, where the RL agent can achieve more positive rewards from human preferences. Our findings from the instant reward distributions confirm that our proposed RLHPS has a positive effect on RL and that the agent can use self-management to learn the values of the human preference scaling.

In summary, our findings show that the setting with human preference scaling is always much better than the baseline with fixed human preferences, which confirms that our scaling model enables the agent to learn in higher-dimensional environments, as our setting did reflect natural human intent. Additionally, from the comparison results of the instant reward distributions, RLHPS is confirmed to have excellent training performance, as instant reward values are made more favourable than those in the initial stages prior to the agent receiving any preferences.

B. RL from Human Preference Scaling with Demonstrations (RLHPS with Demo)

After the experiments on RLHPS, we implemented the estimator interface to link the preference interface and the human based on our proposed approach – RLHPS with Demo. As addressed in Section III-D, we performed two types of data splitting, with 30% and 50% of the data from the preference database used for testing. Because 1,400 human preference inputs could achieve better performance across the different scenarios according to the previous experiment, we kept the same number of labels – 1,400 – to test the performance of RLHPS with Demo. As shown in Figure 5, the experiment employing RLHPS with Demo (30%) indicates that we can

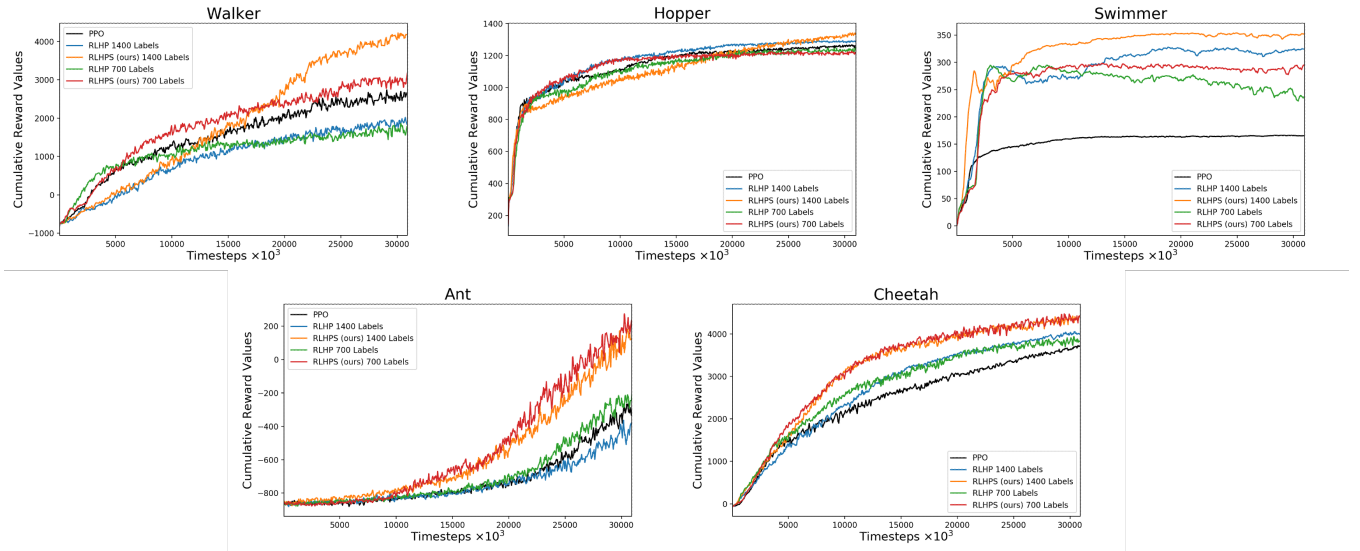


Fig. 3: Experimental Results of PPO, RLHP, and RLHPS in the 5 MuJoCo Scenarios

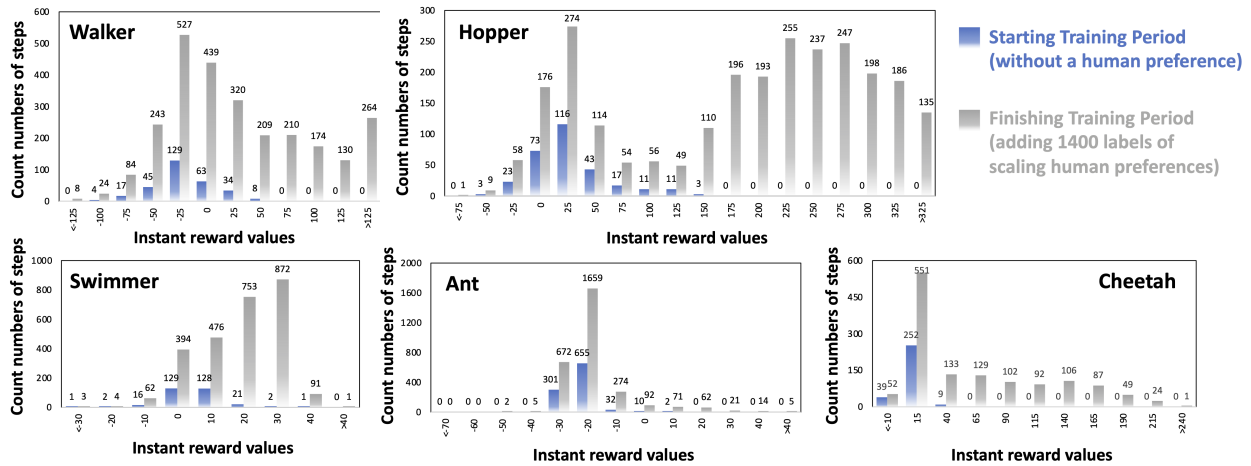


Fig. 4: RLHPS: Instant Reward Distributions Between the Initial and Final Training Periods in the 5 MuJoCo Scenarios

reduce the human preference inputs by 30%; thus, humans need to input only 70% of the 1,400 (980) preference labels, with the remaining 30% predicted by the regression model. To be specific, during the training process, the training dataset of 980 preference labels is randomly selected from all (1,400) preference labels. The remaining 420 preference labels (30% of 1,400 preference labels) are considered the testing dataset. The experiment employing RLHPS with Demo (50%) aims to provide further observations of the influence of reducing the number of human preferences on the cumulative reward values, indicating that 50% of the 1,400 (700) preference labels need to be input by the human, while 50% of the 1,400 preferences can be predicted by linear regression or the SVR model with the smallest MSE, as shown in Table IV, which presents the prediction accuracies achieved using the

averaged MSE trained by linear regression or SVR (with the RBF kernel) in the 5 scenarios.

Cumulative Reward Values: Figure 5 generally indicates that RLHPS with Demo (30% estimated human preferences) can achieve similar or superior cumulative reward values compared to the other approaches, i.e., RLHPS with Demo (50% estimated human preferences) and RLHPS (without Demo), which excludes the human-demonstration stage. For RLHPS with Demo (50% estimated human preferences), where we reduce the number of human preference inputs by half, only a performance similar to that of PPO can be achieved, far from the achievement when using RLHPS. The main reason is that removing 50% of the human inputs will cause higher MSEs, which may lead to a negative impact on training performance.

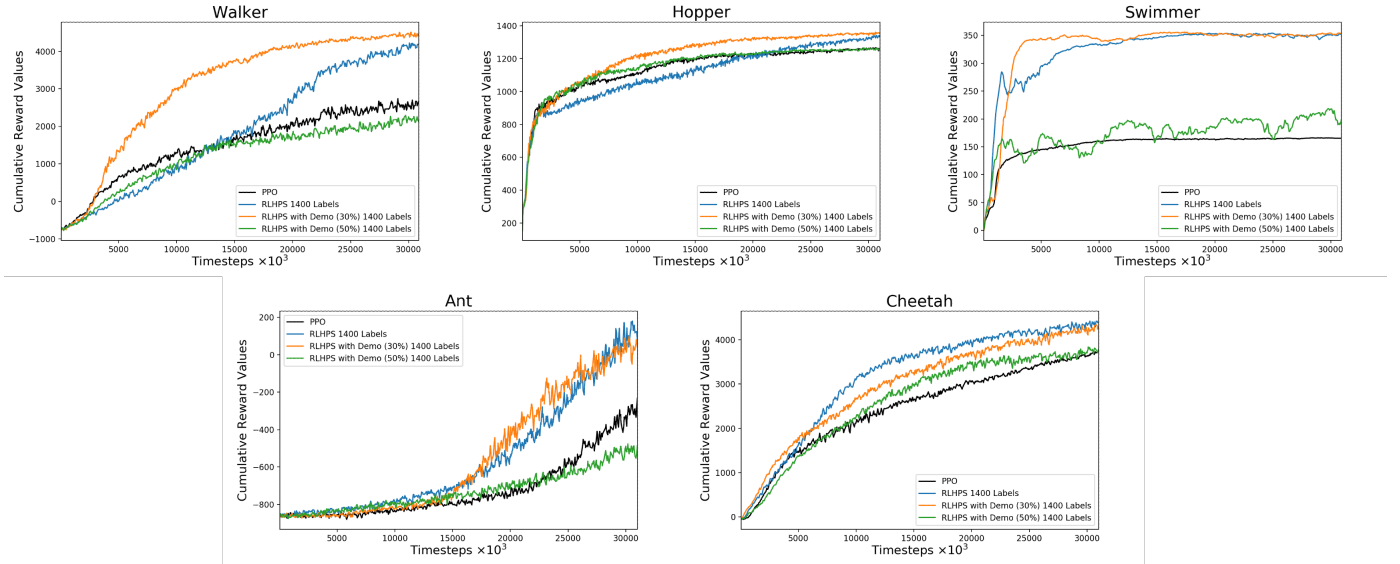


Fig. 5: Experimental Results of PPO, RLHPS, and RLHPS with Demo in the 5 MuJoCo Scenarios

In terms of the Walker and Hopper scenarios, RLHPS with Demo (30% estimated human preferences) yields the highest cumulative reward values. For the Swimmer and Ant scenarios, RLHPS with and RLHPS without Demo have similar performances, suggesting that we can employ 30% less human inputs to achieve similar outcomes. Only in the Cheetah scenario was RLHPS with Demo unable to achieve a comparable performance to that of RLHPS, suggesting that the estimated preferences may not be of benefit to walking guidance for the Cheetah scenario.

TABLE IV: MSEs of Preference Estimations

Scenario	Linear Regression		SVR (RBF kernel)	
	Mean	Standard Deviation	Mean	Standard Deviation
Walker	0.0514	0.0752	0.0099	0.0184
Hopper	0.0688	0.0383	0.0122	0.0716
Swimmer	0.0239	0.0397	0.0737	0.0665
Ant	0.0356	0.0846	0.0637	0.1210
Cheetah	0.0607	0.0769	0.0725	0.0870

Although the 30% reduction may not show a significant improvement in terms of the percentage, it could reduce human input by a notable amount (from 1,400 to 980 inputs) without impacting the training performance. In short, it removes the need for approximately 420 human inputs per experiment. Assuming that a human needs 1 second per input on average, this reduction will amount to 7 minutes saved per experiment under the MuJoCo scenarios. A total of 25 minutes per experiment could also be saved under the Atari scenarios. If we repeat the experiments 25 times, we could save 175 minutes and 625 minutes for the MuJoCo and Atari scenarios, respectively, which are significant time reductions when considering complex scenarios.

C. Observation of Testing Behaviours

We released a video (<https://youtu.be/jQPe1OILT0M>) to demonstrate the behaviours of the trained agent in all Mu-

JoCo scenarios. Generally, our proposed method, RLHPS or RLHPS with Demo can perform more appropriate behaviours to achieve the expected goal than can RLHP or PPO. In only a few exceptional cases, the goal of the Walker scenario is for the agent to roll forward and walk. Nevertheless, by observing the behaviour of the trained agent, we find that the agent operated very slowly to achieve this goal. In the Ant scenario, our RLHPS with Demo seemed to take more time to learn to walk around than required by RLHP or PPO.

D. Limitations

This study has some limitations. Our findings provide insights into some human preference labels that could be generated by the prediction model, which could reduce the number of human inputs without sacrificing the training performance while maintaining an excellent reward return. However, this reduction process is still limited. From the experimental results, when half of the human preference labels are predicted, the training performance is dramatically impacted. Therefore, under the current settings, using up to 30% of the estimated preference inputs in place of human inputs can prevent this negative impact and maintain the training performance.

Furthermore, the frame selections are made on a random basis, which may influence the human preferences for some cases close to the equal option. We suggest that this could be refined by having an intelligent selection model select segments that have diverse rewards or at time points when the prediction model is unable to judge between the two segments. In addition, the normalisation step of our human preference scaling model could be further investigated since the distribution of rewards could affect preference levels for the estimator.

VI. CONCLUSIONS

Our study proposed a weak human preference supervision framework involving a human preference scaling model with

demonstrations that aims to effectively solve complex RL tasks and achieve higher cumulative rewards in simulated robot locomotion – MuJoCo games. We attempted to optimise RLHP in two ways: enhancement of the weak human preference details by scaling the preference levels and reduction of the number of human preference inputs by replacing some inputs with preference labels generated by an estimator. Remarkably, our two developed models, RLHPS and RLHPS with Demo, achieve higher cumulative reward values and significantly reduce the cost of human inputs up to 30% compared to PPO and RLHP. To present the flexibility of our approach, the released video shows comparisons of the behaviours of agents trained on different types of human input.

Given the high scalability of deep RL, we believe that our proposed weak human preference supervision framework, which includes RLHPS and RLHPS with Demo, can help an agent learn natural human preferences with fewer inputs to enhance the training performance. In this work, our contribution to the improvement of RL-based robotic movement potentially approaches human thinking in more complex situations and focuses on the new human-robot interaction scheme by proposing a weak human preference supervision framework in deep RL for robotic controls. **In further studies, we aim to investigate the developed weak human preference supervision in deep RL agents on the further sample-complexity environment, and explore the human preference scaling in human-agent cooperation tasks that require considering trajectories with diverse returns and individual differences of queues, such as a broader core set of robotics challenges concerning the deployment of automated driving systems.**

ACKNOWLEDGEMENTS

We thank Prof. Jun Wang (University College London) and Prof. Haifeng Zhang (Chinese Academy of Sciences) for providing suggestions to improve the presentation of the manuscript. This work was supported in part by the CoSE incentive grant scheme at the University of Tasmania and the US Office of Naval Research Global, under Cooperative Agreement Number: ONRG-NICOP-N62909-19-1-2058.

APPENDIX

The code of this paper can be found at GitHub <https://github.com/kaichiuwong/rlhps>

REFERENCES

- [1] 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.
- [2] S. Nikolaidis, S. Nath, A. D. Procaccia, and S. Srinivasa, “Game-theoretic modeling of human adaptation in human-robot collaboration,” in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, 2017, pp. 323–331.
- [3] K. Bogert, J. F.-S. Lin, P. Doshi, and D. Kulic, “Expectation-maximization for inverse reinforcement learning with hidden data,” in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 2016, pp. 1034–1042.
- [4] Y. Lu, C. Wen, T. Shen, and W. Zhang, “Bearing-based adaptive neural formation scaling control for autonomous surface vehicles with uncertainties and input saturation,” *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
- [5] S. Xu and H. Peng, “Design, analysis, and experiments of preview path tracking control for autonomous vehicles,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 48–58, 2019.
- [6] S. Russell, “Should we fear supersmart robots?” *Scientific American*, vol. 314, no. 6, pp. 58–59, 2016.
- [7] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in ai safety,” *arXiv preprint arXiv:1606.06565*, 2016.
- [8] Z. Ke, Z. Li, Z. Cao, and P. Liu, “Enhancing transferability of deep reinforcement learning-based variable speed limit control using transfer learning,” *IEEE Transactions on Intelligent Transportation Systems*, 2020.
- [9] Z. Cao and C.-T. Lin, “Reinforcement learning from hierarchical critics,” *arXiv preprint arXiv:1902.03079*, 2019.
- [10] Z. Cao, K. Wong, Q. Bai, and C.-T. Lin, “Hierarchical and non-hierarchical multi-agent interactions based on unity reinforcement learning,” in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 2020, pp. 2095–2097.
- [11] A. Y. Ng and S. J. Russell, “Algorithms for inverse reinforcement learning,” in *Icml*, vol. 1, Conference Proceedings, pp. 663–670.
- [12] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, “Imitation learning: A survey of learning methods,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.
- [13] Y. Schroecker and C. L. Isbell, “State aware imitation learning,” in *Advances in Neural Information Processing Systems*, 2017, pp. 2911–2920.
- [14] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” in *Advances in Neural Information Processing Systems*, 2017, Conference Proceedings, pp. 4299–4307.
- [15] F. Xiao, Z. Cao, and A. Jolfaei, “A novel conflict measurement in decision making and its application in fault diagnosis,” *IEEE Transactions on Fuzzy Systems*, 2020.
- [16] E. Todorov, T. Erez, and Y. Tassa, “Mujoco: A physics engine for model-based control,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, Conference Proceedings, pp. 5026–5033.
- [17] B. Woodworth, F. Ferrari, T. E. Zosa, and L. D. Riek, “Preference learning in assistive robotics: Observational repeated inverse reinforcement learning,” in *Machine Learning for Healthcare Conference*, 2018, pp. 420–439.
- [18] J. Fürtkranz, E. Hüllermeier, W. Cheng, and S.-H. Park, “Preference-based reinforcement learning: a formal framework and a policy iteration algorithm,” *Machine learning*, vol. 89, no. 1-2, pp. 123–156, 2012.
- [19] M. Aggarwal and A. Fallah Tehrani, “Modelling human decision behaviour with preference learning,” *INFORMS Journal on Computing*, vol. 31, no. 2, pp. 318–334, 2019.
- [20] M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, “Learning reward functions by integrating human demonstrations and preferences,” *arXiv preprint arXiv:1906.08928*, 2019.
- [21] C. Wirth, R. Akrouf, G. Neumann, and J. Fürtkranz, “A survey of preference-based reinforcement learning methods,” *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4945–4990, 2017.
- [22] D. Kreps, *Notes on the Theory of Choice*. Westview press, 1988.
- [23] A. Wilson, A. Fern, and P. Tadepalli, “A bayesian approach for policy learning from trajectory preference queries,” in *Advances in neural information processing systems*, 2012, pp. 1133–1141.
- [24] J. Ho, J. Gupta, and S. Ermon, “Model-free imitation learning with policy optimization,” in *International Conference on Machine Learning*, 2016, pp. 2760–2769.
- [25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [26] E. Todorov, T. Erez, and Y. Tassa, “Mujoco: A physics engine for model-based control,” in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.
- [27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, “Tensorflow: A system for large-scale machine learning,” in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 2016, pp. 265–283.
- [28] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, “Openai gym,” *arXiv preprint arXiv:1606.01540*, 2016.
- [29] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. d. L. Casas, D. Budden, A. Abdolmaleki, J. Merel, and A. Lefrancq, “Deepmind control suite,” *arXiv preprint arXiv:1801.00690*, 2018.