A Appendix

COVID-19 Effects. We were fortunate to be able to perform an in-person experiment on a real-world robot and human subjects, yet the subject recruiting process was particularly challenging during the current COVID pandemic. With limited access to potential experiment participants, we were only able to run a pilot test with two people prior to the actual experiment reported here. Solely from the pilot study data, we were not able to disqualify our eventual experimental assumption that there is a significant learning effect among participants even within only five games. Instead, as described in Sec. 5.1, it was not until obtaining our full experimental results that we were able to discard this assumption due to observing a large variance in the degree of human learning across participants. Future research that aims to understand the human learning effect in competitive-HRI tasks should collect more data for each individual and facilitate the participant's ability to learn. In particular, increased learning can be achieved by reducing environmental complexity, such as decreasing the robot's joint velocity, reducing the manipulator's reachability in task space, using a robot with a lower degree of freedom (DoF) and so on.

A.1 The Fencing Game

Algo. 2 summarizes the scoring mechanism for the Fencing Game described in Sec. 2 [48].

Algorithm 2: The Fencing Game Scoring Mechanism

```
Initialize: Game score $s = 0$; Timestep = 0.01 Sec; Game horizon = 20 Sec
bat_a → Antagonist's bat
bat_p → Protagonist's bat
target → Target Area
for every timestep in this game do
    if bat_a in target then
        if bat_a contacts bat_p then
            $s$ -= 10
        else
            $s$ += 1
        end
    if bat_p in target for more than 200 consecutive timesteps then
        $s$ += 10
end
```

A.2 System Implementation Details

Hardware Details. The PR2 robot is a popular general purpose robotic platform with two 7 degree of freedom (DoF) arms and its overall form-factor is similar to a human adult [49, 50, 51]. It is comparable to a human player in a competitive game in terms of body size and arm flexibility. The human player's bat is attached to an infrared photo-diode array tracker, and its position and orientation is perceived by the robot via two tracking base stations. An audible scoring feedback system is created to report the scoring situation of each game in real-time. The participants will hear a higher frequency (440 Hz) signal sound when scoring, and a lower frequency (300 Hz) signal sound when receiving penalty from the robot. The system implementation pipeline is shown in Fig. 5.

Algorithm Details. This paragraph provides extra technical details on the two-phase iterative co-evolution algorithm described in Sec. 4. There are two major differences between phase one and two training in Algo. 1. **(1)** Phase one training uses a continuous reward to facilitate agents' development of basic motor skills. In phase two training, the agents are solely rewarded by the game scores. **(2)** In phase one training, each agent is always learning against the latest version of the opponent. But in phase two training, an agent would constantly and randomly load a previous version of opponent from the history, after a short period of training. The use of continuous reward encourages the robot to quickly explore the task space, which makes the phase one training good for quickly initializing the policies for the antagonist and protagonist. However, continued used of the iteration strategy in phase one training would likely trap both agents in a low quality local equilibrium and/or chasing each other in circles in parameter space in a long sequence of training [52]. In contrast, the reward and iteration mechanisms of phase two training create a high variance learning environment which effectively mitigates the circling problem. In addition, because of the high variance nature of the phase two training, agents are more likely to learn emergent behaviors and converge to more sophisticated policies.

Training Details. As shown in Fig. 2, at the beginning of the training (i.e., phase 1 - itr 1) the game scores tend to be largely biased toward whichever agent is currently learning. This is because both agents' policies are simple/naive at the beginning phase and the opponent can easily find a counter strategy to dominate the game during learning. After the warm-start training (i.e., phase 1 - itr 2), the two agents converged to an area in their policy space where both agents are challenging each other without completely dominating the game.

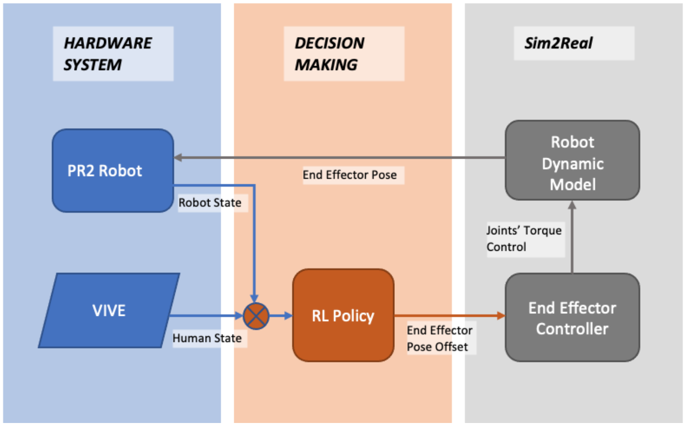

Figure 5: A block diagram demonstrating the pipeline of the proposed robotic system. Human motion tracking is achieved via a HTC VIVE VR system.

A.3 User Studies

Demographics There were 16 human subjects (10 male, 6 female, of age $M = 28.8$ years, $SD = 5.56$) recruited for the first human-subject studies. Nine out of 16 subjects reported doing more than three hours of physical exercise per week, and seven subjects reported less than 3 hours of weekly exercise. Jogging, walking, cardio, and weight training were the most common exercises chosen among all subjects. Ten subjects from this population participated the second user study.

Before the Experiment. Before the experiment, each subject was asked to sit for 3 minutes and then walk for 1 minute to record two average heart rate baseline values. In the experiment, both the robot and a subject held a polystyrene bat to play the games. A safety line was drawn on the ground, and every subject stood behind it to prevent a potential collision. Since the robot was not mobile, each subject also kept his/her feet planted on the ground during a game. The target area was not directly visible; an audible scoring feedback system notified a subject when his/her bat was placed within the target area. A high-frequency signal sound indicated that the human player was scoring, and a low-frequency signal sound indicated that the human player was getting penalization. Before the experiment, a subject had 5 minutes to explore this target area with their bat, so that the target area's location and the scoring mechanism were clear to the subject. Afterward, the subject played two warm-up games with the robot to get further familiarized with the system. In these warm-up games, the robot's actions were slowed down and no data was collected.

Experiment Procedure for User Study One. The experiment contains four sections. In each section, a subject will play five consecutive games with the robot (with a fixed robot policy) and rest for approximately 30 seconds between games. Subjects' heart rates will be recorded during each 20 second game. Due to the short duration of the games, our discussions in Sec. 5.1 use peak heart rate as a summarizing statistic of this recorded data. In the first section of the experiment, the robot will use the warm-start policy resulting from the phase one training. For the rest of the sections, the robot will load one of the three selected characterized policies in each section by following a random order. This ensures that all subjects will first learn to play with the robot at a regular speed, and then play with the robot with a new game style in each of the sections. After each section, a subject will be asked to describe the interaction in the last 5 games by selecting one or more of the following adjectives: ‘Exciting’, ‘Joyful’, ‘Frustrating’, ‘Motivating’, ‘Amusing’, ‘Intimidating’, ‘Physically Demanding’, ‘Cognitively Demanding’, ‘Boring’, ‘Others (please describe:)’. Finally, when a subject finishes all four sections of games, they will complete the last part of the questionnaire that assesses their acceptance of the competitive robot, and their subjective feelings towards the games. We modified the technology acceptance model (TAM) [47] and created the questions in Table. 2. We introduced two extra questions (*i.e.*, DE and IE) to understand the human subjects’ acceptance and desirability of a competitive robot companion in the future, and if a competitive robot can motivate them to engage in physical exercise more frequently. Other than the open-ended question, all TAM questions were measured on a 5-point scale where 1 = “Strongly Disagree,” 3 = “Neutral,” 5 = “Strongly Agree”. At the end of each gameplay section, a subject is also asked to consider and compare both the enjoyability and the difficulty of the games between the finished sections. One section can be equally, less or more enjoyable and difficult than another section. This will allow the participant to have two rankings of the four sections based on their perceived enjoyment and difficulty by the end of the experiment.

Experiment Procedure for User Study Two. This experiment contains two sections. In each section, a subject is asked to play 10 consecutive games with the same robot policy and rest for approximately 15 seconds between games. The order in which each subject plays against the baseline policy and characterized RL policy is randomized. After each section, a subject is asked to answer the modified TAM questions. After the final section, a subject is also asked to answer short questions 2, 3, and 4 in Table. 2.

Baseline Heuristic Policy. We aimed to design a strong baseline heuristic policy to create an intense human robot gameplay experience. Given an observation of the world, the robot orients its bat perpendicular to the human’s bat with random angular offsets drawn uniformly from -25 to 25 degrees on the x, y, and z axes. In order to ensure that the robot is always executing a competitive defense, the policy commands the robot to position the center of its bat in between the target area and the point on the human’s bat that is closest to the target area:

$$\bar{b_p} = \bar{tar} + (\bar{h_{close}} - \bar{tar}) \cdot uniform(0.5, 1)$$
$$\bar{h_{close}} = \bar{h_{low}} + ht \cdot (\bar{h_{up}} - \bar{h_{low}})$$
$$ht = \max\left(0, \min\left(1, (\bar{tar} - \bar{h_{low}}) \cdot (\bar{h_{up}} - \bar{h_{low}})/(2 \cdot L_{sword})\right)\right)$$

Where $\bar{b_p}$, $\bar{tar}$, $\bar{h_{up}}$ and $\bar{h_{low}}$ represent the position of the robot’s bat frame, the center of the target area, the upper end of human’s bat, and the lower end of human’s bat respectively. $\bar{h_{close}}$ indicates the point on the human’s bat that is closest to the center of the target area, and L_{sword} indicates the length of a bat. The function $uniform(0.5, 1)$ randomly determines how far apart the robot’s bat should be from the human’s bat. In addition, there is a 50% chance for the robot to execute the desired bat position calculated from the last time step instead of the latest desired pose. The added uncertainties introduce randomness to the robot’s behavior. This heuristic allows the robot to dominate the fencing game when it can move faster or as fast as the antagonist. However, human subjects are able to move slightly faster than our PR2 robot, which leaves room for human subjects to discover counter strategies.

A.4 Heart Rate

All heart rate data were recorded by a Polar OH1+ optical heart rate sensor. Fig. 4. b. compares the subjective descriptions between four groups of gameplay sections with different levels of average human heart rates. For each section in the user study, we first calculate the average peak heart rate over the corresponding five games in the section. A section’s heart rate level l is calculated by dividing the section’s average peak heart rate by the corresponding user’s walking baseline heart rate, which results in a percentage value describing how much more or less the average section heart rate is compared to the baseline. The low, medium, high, and ultra-high heart rate groups contain the sections that $l \leq 100\%$, $100\% < l \leq 120\%$, $120\% < l \leq 140\%$, and $140\% < l$ respectively.

	Questions
Perceived Usefulness (PU)	Having a competitive robot companion would improve the quality of my physical exercise.
Perceived Ease of Use (PEOU)	Learning to earn higher score (make progress) in the games with a competitive robot would be easy for me.
Attitude (ATT)	Using a competitive robot exercise partner to improve my exercise quality is a good idea.
Intention to Use (ITU)	Assuming I had access to a competitive robot for exercise, I would intend to use it.
Perceived Enjoyment (PENJ)	I would find competitive human-robot gameplays are entertaining.
Desirability (DE)	Based on your experience today, future physical exercises and games with a competitive robot will be desirable.
Increased Engagement (IE)	Having a competitive robot companion would make me more likely to engage in physical exercise.
Short Question 1 (used in study one)	Are there anything you would like to change to improve the interaction experience?
Short Question 2 (used in study two)	Which robot (Section 1, Section 2, or equally) do you think is more challenging/difficult to play against? and why?
Short Question 3 (used in study two)	Which robot (Section 1, Section 2, or equally) do you think is more enjoyable/fun to play against? and why?
Short Question 4 (used in study two)	Which robot (Section 1, Section 2, or equally) do you think is more intelligent? and why?

Table 2: Modified Technology Acceptance Model and Open-ended Questions

	Baseline Median	RL Median	t-statistic	p-value
PU	3.6	3.6	0.0	1.0
PEOU	3.9	3.6	0.85	0.41
ATT	4.1	3.9	0.41	0.69
ITU	3.7	3.6	0.2	0.84
PENJ	3.9	3.9	0.0	1.0
DE	3.7	3.9	-0.45	0.65
IE	3.0	3.5	-0.9	0.37

Table 3: TAM Response Comparison Between Baseline Policy and RL Policy in Pair-T Tests

A.5 Additional Experimental Results

Perceived Ease of Use. Interestingly, although we did not observe significant performance improvement within each section nor between sections, 62.5% of subjects perceived that it was easy for them to make progress when playing against the robot. Furthermore, some participants expressed a desire to beat the best score of previous participants. It is possible that some subjects focused only on attaining their perception of a “high” score for a small number of games rather than maintaining good performance in all 20 games. On the other hand, in a competitive game setting, we are still not sure exactly what it means for people to feel that getting a better score would be easy. In our experiment, it probably suggested that most participants considered the robot opponent to be surmountable because only a very small number of participants described their gameplay experience as “Frustrating”, “Intimidating”, or “Boring”. However, future work could study how the perceived ease of use affects the participants’ effort in competitive games.

Enjoyment and Difficulty. We found that perceived enjoyment and difficulty do not significantly vary as the amount of player experience increases, but the amount of data used to train the corresponding policy does have an affect. Fig. 6 shows the average ranking comparison for enjoyability and difficulty across both policies and experiment sections. Among the utilized policies, no significant variance was observed across the three characterized polices for both enjoyability($p = 0.40$)

and difficulty($p = 0.35$). Paired-T tests show that the warm-start policy is less enjoyable and difficult than characterized policies ($p < 0.01$), except for the enjoyability of the warm-start policy and policy 2 ($p = 0.13$). The result is similar in the time domain - no significant variance was found in the last three sections, in which characterized policies were randomly ordered (enjoyability: $p = 0.5$, difficulty: $p = 0.42$). Section one, which only uses the warm-start policy, is less enjoyable and difficult than other sections($p < 0.05$). Across all sections, the perceived difficulty is positively correlated to perceived enjoyment with a moderate coefficient of 0.6.

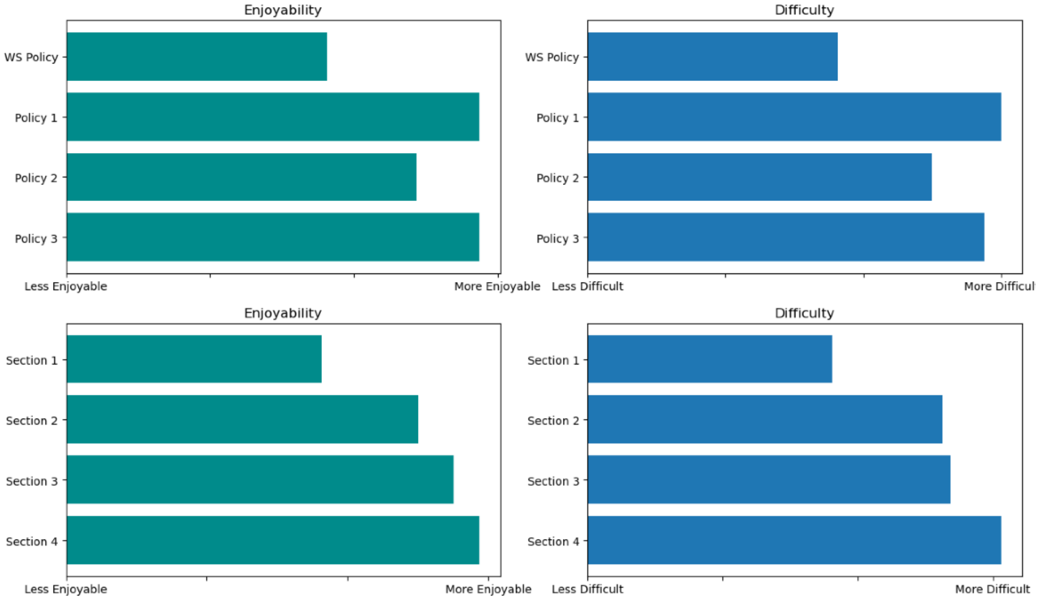

Figure 6: The average ranking comparison for enjoyability and difficulty across both policies and experiment sections.