Exploring Embodiment and Dueling Bandit Learning for Preference Adaptation in Human-Robot Interaction

Sebastian Schneider[1] and Franz Kummert[1]

***Abstract*— Adaptation for social companions is a crucial requirement for future applications. Personalized interaction seems to be an important factor for long-term commitment to interact with a social robot. We present a study evaluating the feasibility of a dueling bandit learning approach for preference learning (PL) in Human-Robot Interaction (HRI). Furthermore, we explore whether the embodiment of the PL agent has an influence on the user's evaluation of the learner. We conducted a study (n=53) comparing a graphical user interface (GUI), a virtual robot and a real robot. We found no difference regarding the preference for the virtual or real robot. We used the obtained study data to compare the PL approach against a strategy that randomly selects preference rankings. The results show that that the dueling bandit PL approach can be used to learn a user's preference in HRI.**

I. INTRODUCTION

Robots have recently been introduced as tools that could assist user's during conventional rehabilitation, health care or learning programs (i.e. stroke-rehabilitation [1], dieting [2] or teaching [3]). The nature of these tasks requires a longer commitment of the user. Neither rehabilitation nor teaching or health care issues can be achieved during a single session. Hence, tools such as robots have to apply methods that engage users in long-term interaction. Furthermore, they will have to provide meaningful and personalized interaction because every person is an individual with a personal history that is represented in ones desires and preferences. While highly specialized physicians, therapists or coaches are trained to provide individualized personal interaction for each person, robots are still far from such capabilities. Hence, robotic tools can so far be used alongside trained personal. However, small steps are currently made to investigate the implementation of social robots in long-term use cases [4]. A review of different researches has concluded four major building blocks for robots to be able to engage users in long-term interaction: behavior, adaptation, empathy and design [4]. While all of these aspects are important for engaging users in long-term intervention, we are focusing on the aspects of adaptation in this work.

Several researchers have already worked on adaptation in Human-Robot Interaction (HRI) [1], [5], [6]. In these works, the robot adapts its verbal and non-verbal behavior to the user's preferences to increase the short-term acceptance. But what kind of adaptation is necessary for longer user commitment for socially assistive robots (SAR)? The main purpose of SAR is to assist users on a task. These tasks can have different difficulties, categories, duration or feedback types and so on. One possible adaptation is the adjustment of the task difficulty which has been presented in [1]. However, the preference for different tasks or task categories has not received a lot of attention. Since different kinds of tasks might lead to the same rehabilitation-, learning- or coaching goal, preference learning (PL) could be utilized to learn a user's task preferences over time.

In general, PL is already widespread in the domain of recommendation systems. Among these algorithms to optimize search results or provide customized advertisements bandit learning algorithms (e.g. exp3, ucb [13], [14]) are used. By showing the user different kinds of personalized advertisement or search results the algorithms learns the user's preferences. The way those algorithms learn is by the user's implicit feedback (i.e. clicking behavior) and thus personalize the user experience in the background. In this work, we want to know whether these kinds of algorithms can be used to learn the user's preferences in HRI for socially assistive tasks. Particularly, we study a special kind of bandit learning (i.e. dueling bandit learning [15]) for PL. In contrast to standard bandit learning techniques, this approach does not require a numerical reward function. This approach is specially suitable for learning tasks where the reward is dependent on the user's feedback, because humans are better in giving relational preference statements than quantitative preference statements [16]. Thus, those kind of PL algorithms seem to be more reliable.

A. Research Question

In our recent line of research, we focused on social assistance during exercising and sportive activities [17], [18]. Hence, our goal is to learn a user's exercise category preference. At the moment, we only consider a set of categories that are suitable for a robot to accompany or instruct a user in the near-future. These categories are strength, cardio, endurance, stretching and relaxation/meditation. Our first research question (**RQ1**) is to investigate whether dueling bandit learning is suitable for HRI and whether the algorithm can effectively learn the user's preferences? We target this research question by a) evaluating the suitability of one state-of-the-art dueling bandit learning algorithm in a human-computer-/human-robot-interaction study and by b) evaluating the learned user preference ranking of the learning algorithm against a simulated random ranking condition. Additionally, we are interested in the effects of the embodiment of the system. A recent literature survey on the

[1]Sebastian Schneider and Franz Kummert are with Faculty of Technology, Applied Informatics Group, CITEC, Bielefeld University, 33604 Bielefeld, Germany [sebschne,franz]@techfak.uni-bielefeld.de

TABLE I: Research in the field of adaptation and personalization in HRI.

work	method	variables	learning goals
Tapus *et al.* [1]	reinforcement learning	user personality traits nu. of performed exercises	interaction distances/proxemics, speed, and vocal content
Tsiakas *et al.* [7]	reinforcement learning	user performance, session state	adjust time of movement, move to next exercise, encourage user
Leite *et al.* [5]	multi-armed bandit learning	user's detected valence	choose appropriate emphatic behavior
Leyzberg *et al.* [3]	Bayesian net	puzzle state	provide personalized tutoring sessions
Lim *et al.* [8]	hybrid filtering	semantic knowledge, event episodic knowledge and emotion	enhance student's motivation to prevent negative emotions
Baraka *et al.* [9]	multi-armed bandit learning	numerical reward provided by user	robot's light animation
Mitsunaga *et al.* [6]	reinforcement learning	body signals	adjust interaction distance, gaze, motion speed and timing
Hemminghaus *et al.* [10]	reinforcement learning (Q-Learning)	gaze behavior, speech, game state	memory game assistance
Chan *et al.* [11]	hierarchical reinforcement learning	speech analysis, user state, activity state	giving instructions, empathy or help
Lee *et al.* [12]	Wizard of Oz	snack choices patterns, usage patterns, robot's prior behavior	personalized speech topics

effects of embodiment showed: "that a co-present, physical robot performed better than a virtual agent simulated using computer graphics. These studies found a co-present robot to be more persuasive, receive more attention and be perceived more positively than a virtual agent even when the behavior of the robot was identical to the behavior of the virtual agent and when both agents had similar appearance" [19]. Hence, in our second research question (**RQ2**) we want to know whether the embodiment of the learning agent influences the user's perceived likeliness, intelligence and persuasiveness during a PL task.

B. Hypothesis

We draw the hypothesis 1 (**H1**) that an embodied agent will increase the user's agreeableness with the learned preferences, the perceived intelligence and likeability compared to a virtual representation or no agent representation.

This paper is organized as follows: Section II gives an overview of related work in the field of PL and user personalization in HRI. Section III describes our PL framework. Section IV explains our study design. Section V presents our results which are then discussed in the last section.

II. RELATED WORK

What is going on in the world of personalization and adaptation in HRI? To summarize, there are basically two trends. One trend is to adapt the robot's behavior based on interactive machine learning. These approaches mostly utilize reinforcement learning with user feedback and sensor data (e.g. [1], [5]–[7], [9]). Other approaches create user models to adapt the robot's assistance and behavior ([3], [20]) or rely on techniques from recommendation systems like collaborative filtering [8]. One of the major applications of personalization in HRI is concerned with the adaptation of the robot's social behavior to match the user's personality or desires. In these cases, behavior adaptation is often based on personality matching to adjust interaction parameters like proxemics, speed, vocal content, robot's appearance or dialog topics ([1], [9], [12]). The goal of these adaptation techniques is to enhance the user's acceptance of the robot which is believed to increase the user's commitment to interact with the system in the long run. Other works include approaches like reinforcement learning, bandit learning or Bayesian nets to adjust session parameters or generate supportive and emphatic behaviors (e.g. [3], [5], [7], [10], [11]). In these scenarios personalization targets the user's learning gains, therapy success or enjoyment during games. Table I gives an overview of different research directions in the field of HRI.

Basically, all approaches show that an adapted robot behavior is preferred by the user and leads to better learning outcomes and a higher robot acceptance. However, most of the works include some kind of implicit direct feedback from the user (e.g. sensor data), require the user to fill out a questionnaire, or a wizard of oz to personalize the robot behavior. Furthermore, in many reinforcement learning approaches feedback needs to have numerical value in order to learn a user adapted policy. This approach can be a bottleneck of the implementation because a direct feedback is not available or it is based on the engineers understanding of how to represent the numerical feedback. In some applications it might be difficult to determine a numerical reward function or it might be challenging how to obtain the actual reward. Hence, this work extends the literature by evaluating how reinforcement learning, in our case bandit learning, can be used to personalize the human's HRI experience. We therefore draw from research that extended multi-armed bandit learning scenario to a dueling bandit learning scenario [15]. In those scenarios the agent learns the user's preference by presenting the user two items. The feedback is then represented by a qualitative preference feedback of the user. Based on this approach the agent can learn the user's preference of a given set of items without the need of having a numerical reward.

III. PREFERENCE LEARNING FRAMEWORK

To introduce the PL framework we describe the PL problem first.

A. Problem Statement

The classical multi-armed bandit (MAB) learning problem is motivated by the scenario of a gambler who has to decide which slot machine of a row of machines to play, how many times to play each machine and in which order. The agent has to simultaneously explore and exploit a set of choice

alternatives in a sequential decision process. Therefore, the agent needs some kind of real-valued reward. However, this is often not given and a numerical reward is not available [15]. For example, it would be more difficult for a human to associate an action with real-value reward than comparing two actions and choosing which one is better or which one they like more. This is because humans excel in giving relative preference statements in the form of qualitative comparisons between pairs of alternative [16]. Therefore, the MAB problem has been extended to an dueling bandit learning problem [21] which draws two (ore more) actions and receives a relative preference statement as reward. This procedure is more formally explained in the following paragraph.

The dueling bandit problem consists of $K(K \geq 2)$ arms, where at each time step $t > 0$ a pair of arms $(\alpha_t^{(1)}, \alpha_t^{(2)})$ is drawn and presented to a user. A noisy comparison result w_t is obtained, where $w_t = 1$ if a user prefers $\alpha_t^{(1)}$ to $\alpha_t^{(2)}$, and $w_t = 2$ otherwise. The distribution of the outcomes is presented by a preference matrix $P = [p_{ij}]_{KxK}$, where p_{ij} is the probability that a user prefers arm i over arm j (e.g. $p_{ij} = P\{i \succ j\}, i, j = 1, 2, .., K$).).

The goal of the PL task is, given a set of different actions (e.g. different sport categories), find the user's preference order for these categories by providing the user two α_i and α_j and update the user preferences based on the selection of the preference between $\alpha_i \succ \alpha_j$ or $\alpha_i \prec \alpha_j$.

Thus, the challenge is to find the user's preference by running an algorithm that balances the exploration (gaining new information) and the exploitation (utilizing the obtained information).

B. System Implementation

Figure 1 gives an overview of our learning framework. At each time step, the algorithm selects two candidates from the preference matrix (Step 1). In our implementation, we used the double Thompson sampling (DTS) algorithm as dueling bandit learning algorithm [22]. However, we neglected the exploitation phase, because in this work we are only interested in obtaining new information. Based on the selected categories, two specific exercises are selected randomly from an exercise database[1]. This database holds six different exercises for each sport category. Following, these exercises are presented as text on a display (Step 2). Subsequently, the user can give relative preference feedback by selecting the preferred exercise (Step 3). This feedback is then used to update the preference matrix accordingly (Step 4). After twenty iterations the system gives the user a ranking in relation to the learned preference matrix. The sport category which wins against most other categories is presented as first followed by the other categories in descending order by their number of wins. It takes approximately 10 minutes to execute all iterations.

[1]https://www.mongodb.com/,visited on 3/23/2017

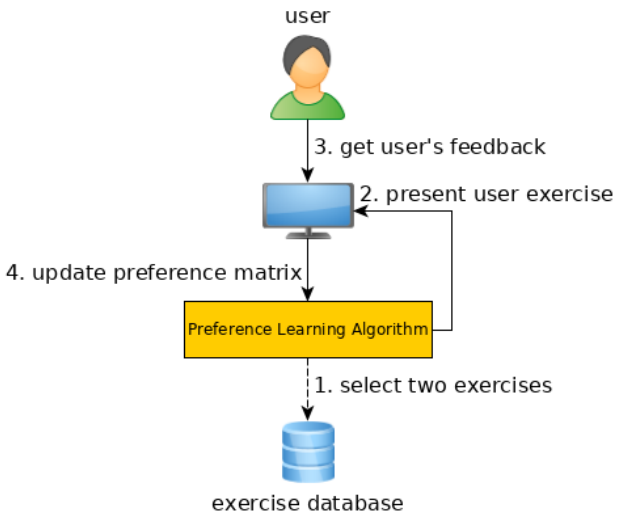


Fig. 1: Preference learning system interaction overview.

IV. STUDY DESIGN

In our paper we question two things: The feasibility of a PL algorithm for HRI (**RQ1**), and whether the embodiment of a robot has an effect on the perceived intelligence and likeability of the robot during a PL task (**RQ2**). We used the described PL framework to learn a user's exercising preference on which we can further evaluate the effectiveness of dueling bandit learning. In our study design we only tested one PL algorithm and all participants interacted with the same PL algorithm (i.e. DTS) running in background.

To answer **RQ1** we used the obtained preference rankings from this study, as a baseline, to compare the algorithm against one which randomly selects a preference ranking. We compared these two algorithms in a simulation by using preference ranking metrics. This evaluation requires that the learned preference ranking should not differ across the embodiment conditions. We will provide evidence in the result section that this is eligible.

To investigate **RQ2**, we manipulated the embodiment of the system (see Figure 2). Participants were randomly assigned to one of the following conditions: *computer* only, *virtual* Nao[2], *real* Nao. The computer condition only included a graphical user interface with buttons and a text area. The text area displayed an introduction text, exercise comparisons, explanations regarding the exercises and finally the learned preference ranking. The user can select their preferred exercise by pressing the according button. In the robot conditions, either a *virtual* Nao (displayed using Choregraphe) was presented on the computer display or a *real* Nao was standing next to the computer. Besides this manipulation, the system behavior was the same for all conditions. Both the real and the virtual Nao spoke the same text as was displayed on the computer. For speech synthesis and gesture generation

[2]https://www.ald.softbankrobotics.com/en/cool-robots/nao,visited on 3/23/2017

we used the ALAnimatedSpeech module of the NaoQi API.

A. Study Procedure

Each participant arrived individually at our lab and had to read and sign a consent form. The experimenter told the participant that s/he will interact with a system that will learn their exercise preferences by displaying different names of exercises and that s/he can select the one s/he is favoring. If the name of an exercise is unknown to the participant, s/he can get more information from the system regarding the category the exercise belongs to (i.e. "push-up is a strengthening exercises", "running belongs to endurance sports", and so on). After the instructions, the participant was guided to the experimental room and told that s/he should exit the lab after the interaction has finished. This is when the manipulation happened. In the room was either only the computer, the computer with a virtual Nao or a real Nao present. The experimenter did not explain anything else regarding the virtual or real robot. During the study, the system iterated through twenty exercise comparisons and in the end presented the learned exercise preference ranking. Afterwards, the participant left the room and answered a survey. Finally, the participant received a monetary compensation (4 Euro) and was debriefed.

B. Participants

We acquired 53 participants from our campus. They were equally distributed between the three conditions (computer: 18, virtual: 18, robot: 17). We had 18 male and 34 female participants. In each condition were 6 male participants. The average age was $M = 25.34$ with $SD = 5.47$.

C. Measurements

In the following, we describe the different measurements we used to investigate our research question. To evaluate the utility of PL (**RQ1**) we use quantitative ranking evaluations to analyses the efficiency of the PL algorithm compared to a randomly selected preference ranking. For the evaluation of the embodiment (**RQ2**) we use subjective user ratings of the system.

1) Personality: We used the Neo-FFI-30 personality scale to assess the participant's personality profile [23]. We used all five sub-scales Neuroticism (Cronbach's $\alpha = .79$), Extraversion (Cronbach's $\alpha = .62$), Openness (Cronbach's $\alpha = .72$), Agreeableness (Cronbach's $\alpha = .48$) and Conscientiousness (Cronbach's $\alpha = .76$).

2) Perception of the Agent: In order to assess different perception of the system between the conditions we asked the participant to rate the system based on the Godspeed questionnaire [24], a 5 point-based semantic differential scale with bipolar items. We used all subscales Animacy (Cronbach's $\alpha = .79$), Anthropomorphism (Cronbach's $\alpha = .83$), Likeability (Cronbach's $\alpha = .89$), Intelligence (Cronbach's $\alpha = .84$) and Perceived Safety (Cronbach's $\alpha = .67$).

3) System Usability Scale: We asked the participants to rate the system's usability on a ten item 5-point Likert scale (Cronbach's $\alpha = .85$) [25].

4) Perceived Information Quality and Openness: We assessed the participants perception of information (Cronbach's $\alpha = .76$) and comparison quality Cronbach's $\alpha = .78$) and the openness to influence (Cronbach's $\alpha = .88$) on a five-point Likert-scale [26].

5) Intrinsic Motivation and Interaction: To assess intrinsic motivation, we used a short German version of the Intrinsic Motivation Inventory (Cronbach's $\alpha = .84$) proposed by [27]. Furthermore, we asked the participants to rate the quality of the interaction on a 5-point Likert scale.

6) Learned Preference Quality: To gain insights on the perceived PL satisfaction, we assessed the participants satisfaction with the learned preference on a four-item 5-point likert scale (Cronbach's $\alpha = .9$). Additionally, if the participants were not satisfied with the learned preference, they could provide their own preference order which we will use later for our system evaluation.

7) Preference Ranking Error: To assess the quality of the obtained preference rankings, we use the two following ranking error functions: D_{PE} which is the position error distance and D_{DR} which is the discounted error. Given a set of items $X = x_1, ..., x_c$ to rank and r as the user's target preference ranking and $\hat{r}$ as the learned preference ranking. Both r and $\hat{r}$ are functions from $X \rightarrow \mathbb{N}$ which return the rank of an item x. The position error is defined as follows

$$D_{PE}(r, \hat{r}) = \hat{r}(argmin_{x \in X} r(x)) - 1 \tag{1}$$

The idea of this distance measure is that we want the target item (i.e. the highest ranked item from r) to appear as high as possible in the learned preference ranking $\hat{r}$. Thus, this distance gives the number of wrong items that are predicted before the target item. The discounted error is defined as follows

$$D_{DR}(r, \hat{r}) = \sum_{i=1}^{c} w_i \cdot d_{x_i}(\hat{r}, r) \tag{2}$$

where $w_i = \frac{1}{log(r(x_i)+1)}$. This distance measure gives higher ranked items from r a higher weight for the distance error d_{x_i} between the rankings. In other words, having a correct ordering of the high ranked values form r is more important than of the low ranked items of r.

V. RESULTS

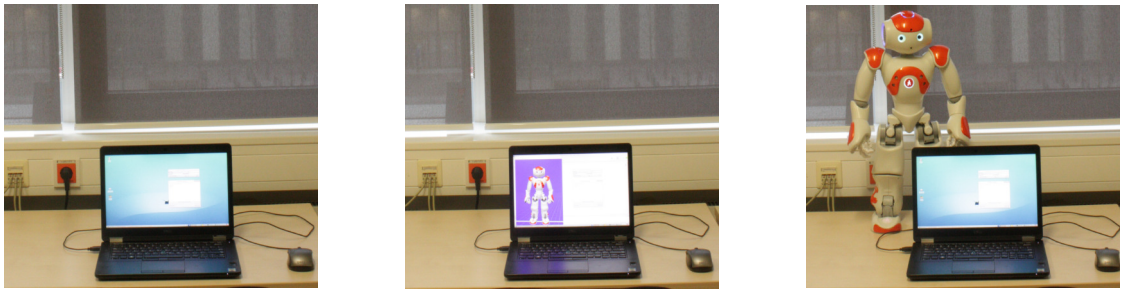
We analyzed the data with an analysis of variance (ANOVAs) when the assumptions[3] of an ANOVA where met. Otherwise, we used a Kruskal-Wallis Tests. To analyze frequency samples we used the Fisher's exact test. All computations have been done with R[4].

A. Manipulation Check

Using ANOVAs we did not find any difference for hours spent for sport per week ($P = .6$), age ($P = .63$). We also did not find any difference between the conditions based on their previous experience with interactive systems

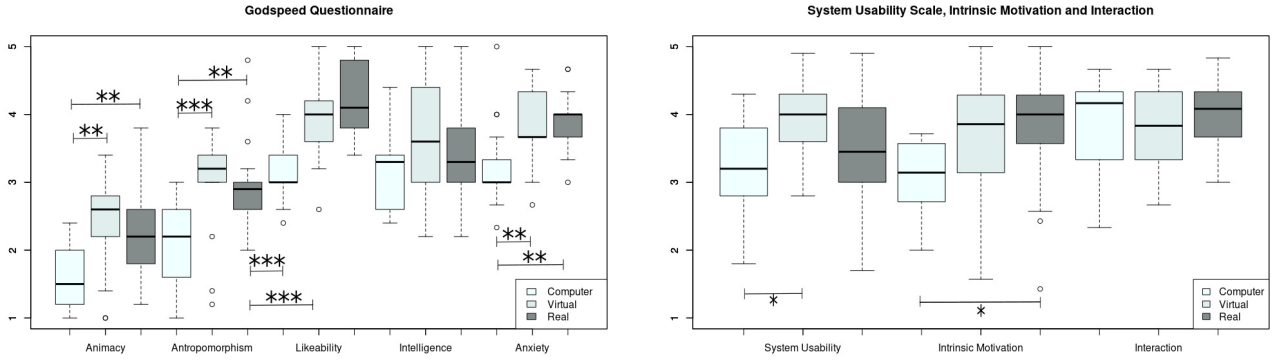
[3]We tested the data for homogeneity of variance using a Levene's Test and for normality using a Shapiro-Wilk Test.

[4]`https://www.r-project.org/`.



(a) Graphical user interface only (b) Virtual Nao (c) Real Nao

Fig. 2: The conditions from this study design.



(a) Godspeed Questionnaire ratings (b) System usability scale and intrinsic motivation scale ratings

Fig. 3: Subjective system evaluation (***:$p < .001$, **: $p < .01$, *: $p < .05$).

or robots ($P = .12$). Regarding the users personality, we did not find any difference for neuroticism ($P = .41$), openness ($P = .98$), agreeableness ($P = .27$), extroversion ($P = .48$) and conscientiousness ($P = .76$) between the different conditions. Thus, our randomization was successful.

B. Godspeed, System Usability and Intrinsic Motivation

We conducted ANOVAs and Kruskal-Wallis tests to measure differences in the Godspeed questionnaire rating between the three conditions. We found differences for the perceived animacy ($F_{2,50} = 9.27, p < .001$), anthropomorphism ($H(2) = 18.398, p < .001$), likability ($F_{2,50} = 21.04, p < .001$) and perceived safety ($H(2) = 14.64, p < .001$). However, we found no differences for perceived intelligence ($P = .3$). We conducted several pairwise comparisons using multiple comparion test after Kruskal-Wallis test or t-test with pooled SD and Bonferroni correction for the different items and conditions. Table II shows the p-values/observed differences for the pairwise comparisons. We found no significant differences between the *real* and *virtual* condition for animacy, antropomorphism, likeability and safety. Not surprisingly, we found significant different ratings between the *computer* condition and the other conditions for animacy, anthropomorphism, likeability and perceived safety. The computer was rated significantly less on all the Godspeed scales, except for intelligence (see Fig. 3a).

TABLE II: Results from post-hoc analysis.

Pairwise t-test with pooled SD

Godspeed item	conditions	p-value
Animacy	*real* vs. *computer*	$< .01$
	real vs. *virtual*	n.s.
	virtual vs. *computer*	$< .01$
Intelligence	*real* vs. *computer*	n.s.
	real vs. *virtual*	n.s.
	virtual vs. *computer*	n.s.
Likeability	*real* vs. *computer*	$< .0001$
	real vs. *virtual*	n.s.
	virtual vs. *computer*	$< .001$

Post hoc test for Kruskal-Wallis test

		obs.diff, p-value
Anthropomorphism	*robot vs. computer*	15.86, $< .05$
	real vs. *virtual*	5.44, n.s.
	virtual vs. *computer*	21.31, < 0.5
Perceived Safety	*real* vs. *computer*	17.33, $< .05$
	real vs. *virtual*	0.95, n.s.
	virtual vs. *computer*	16.37, $< .05$

An ANOVA for the system usability scale revealed significant difference across the conditions, $F_{2,50} = 4.59, p < .05$. Pairwise comparisons using t-tests with pooled SD and Bonferroni correction revealed significant differences between the *computer* and the *virtual* agent condition ($p < .05$). Using a Kruskal-Wallis test, we found that intrinsic motivation was significantly affected by the conditions, $H(2) = 8.66, p = .014$. A post-hoc test with focused comparisons of the mean ranks between conditions showed that intrin-

TABLE III: Frequency counts of Learned Sport Preferences on 1st and 2nd Rank

	Exercises				
Condition	Stretching	Cardio	Endurance	Strength	Relaxation
computer	7	4	11	10	4
virtual	8	3	9	7	7
robot	4	6	12	8	6

sic motivation were not significantly different in the *robot* condition ($difference = 1.06$) and the *computer* condition ($difference = 12.40$) compared to the *virtual* condition. However, the intrinsic motivation was significantly higher in the robot condition compared to the computer condition ($difference = 13.47$). Finally, the openness to influence was also not influenced by the embodiment ($P = .12$).

C. Preference Learning Evaluation

In all conditions (e.g. virtual, computer, robot) we used the same PL algorithm [22]. At this point we want to compare whether the embodiment of the system can influence the PL process. Additionally, we want to measure the effectiveness of the learning algorithm. Therefore, we used our collected preference rankings as evaluation criteria in a simulation to compare the learning algorithm against a random algorithm, which selects a random preference ranking.

1) Subjective Ratings: We conducted several ANOVAs to measure the user's perception of the effectiveness of the PL algorithm. An ANOVA for the perceived information quality ($P = .156$) and comparison quality ($P = .63$) showed no significant differences across the conditions. The perception of the learned preferences quality did not differ significantly across the conditions ($M_{computer} = 3.3$, $SD_{computer} = 1.0$, $M_{virtual} = 3.7$, $SD_{virtual} = .87$, $M_{robot} = 3.9$, $SD_{robot} = .75$, $P = .12$).

2) Preference Ranking Error: Frequencies for the learned sport preferences are summarized in Table III. A Fisher's exact test revealed no statistical significance ($p = .83$, FET). The ranking errors D_{PE} and D_{DE} are depicted in Figure 4. An ANOVA revealed no significant differences for D_{PE} ($P = .55$) and D_{DE} ($P = .32$) between the conditions. Hence, it seems plausible that the embodiment does not alter the learned preferences and that the obtained data can be used for an evaluation of the algorithm.

To measure the effectiveness of the PL algorithm we simulated a *random* condition where a ranking is randomly selected and compared this to the other conditions that used the DTS algorithm. We used the obtained ranking preferences as target criteria and computed the position and discounted error accordingly. Including this random condition in our ANOVA we receive a significant differences for D_{PE} ($F(3, 75) = 21.5, p < .001$) and D_{DE} ($F(3, 75) = 28.3, p < .001$). A pairwise comparison using t-tests with pooled SD revealed significant differences between the random and all other condition for the D_{DE} and D_{PE} (all $p < .001$).

VI. DISCUSSION

In this work investigated the suitability of a dueling bandit PL framework for personalization in HRI and the effects

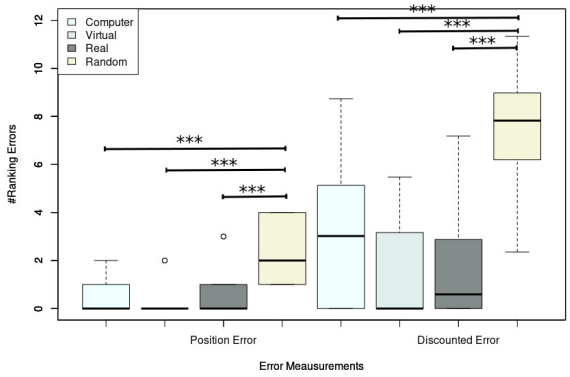


Fig. 4: Box plot showing the preference ranking errors for each condition based on the different error measurements (***:$p < .001$).

of the system's embodiment on the user's evaluation of the system. Recent work reported that embodied robots are found to be more persuasive, enjoyable and entertaining [19]. However, there also exist an ongoing debate on the effects of embodiment especially on the behavior effects a robot can have [28], [29]. With this work we contribute to this ongoing research. To answer our **RQ2**, we conducted a study to investigate how the user's perception of a PL system is influenced by the embodiment of the system. Our reported results from Section V do not support our hypothesis **H1**. The *real* robot and the *virtual* robot have been rated similar on all the sub-scales of the Godspeed questionnaire. Also the ratings for the system usability scale and the intrinsic motivation scale did not differ significantly between the *virtual* robot and the *real* robot. However, we found evidence that the embodiment of the system (both *virtual* and *real* robot) significantly increased the participants likeability of it compared to the computer only condition. Because the Godspeed questionnaire was designed to evalute robots, we do not discuss any differences on the animacy or anthropomorphism scales between the computer and the real/virtual robot. However, we assume that the likability of a computer system can be evaluated using the items of the likability subscale. Furthermore, the embodiment increased the perception of the system's usability between the *virtual* robot and the *computer* condition and increased the user's intrinsic motivation between the *computer* and the *real robot* condition. Regarding the user's preference ranking satisfaction, the embodiment also did not influence the subjective evaluation of the ranking quality. Hence, the participants in all conditions equally trusted the suggested preference ranking. There are several reason that could hinder an effect of the presence of the robot. The real and virtual robot were an additional interface to the graphical user interface. Hence, the real robot might not have been such a salient cue as expected compared to the virtual agent. Also the virtual robot should have been presented on an external monitor and not on the same display as the graphical interface. This aspects limit the generalization of this work and should be investigated in the future. However, other researchers

comparing the physical embodiment during task assistance also manipulated the presence and kept a graphical user interface alongside the robot (e.g. [30]). Since the user's did not evaluate the intelligence of the robot differently across the conditions, we could assume that the perceived intelligence is not influenced by its embodiment but by the underlying algorithms. Thus, more research on the influence of embodiment and algorithmic design on the perceived intelligence is needed.

In our **RQ1**, we wanted to investigate the effectiveness of the PL framework for HRI. First of all, the results indicate that the users were satisfied with the system's suggested preference ranking. Their agreeableness with the learned preferences is fairly high and the calculated ranking errors low. The comparison with a simulated random condition showed that dueling bandit learning reduces the ranking errors significantly (see Fig. 4). To the best of our knowledge this is the first work exploring the dueling bandit learning approach in real HRI. Hence, we propose that dueling bandit learning might be a suitabe framework for personalizing HRI experiences without cognitively overloading the user, needing a numerical reward function or taking a lot of time for the learning process.

VII. Conclusion

In this work we wanted to test whether the dueling banding learning paradigm works in a real HRI situation (**RQ1**) and whether the user experience is influenced by the systems embodiment (**RQ2**). We found support that the learning approach might be suitable for future applications. However, we could not find support for our hypothesis **H1**. The virtual agent and real agent where evaluated equally by the users. However, we can support that users prefer a real or virtual embodied agent over a non embodied system.

In our future work we will investigate the long-term effects of personalization in situations where the robot is actually doing the exercises together with the human or instruct them during the exercises and thus iteratively learns the user's preferences by exploring the exercises together. We assume that this framework can be used to adapt the exercise program of the user to his/her individual preferences. Furthermore, we will explore whether the personalization algorithm can be accelerated by a user model which predicts whether a user will like or dislike a certain exercise category based on their personality type.

ACKNOWLEDGMENT

This research was funded by grants from the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), Bielefeld University.

References

[1] A. Tapus *et al.*, "Hands-off therapist robot behavior adaptation to user personality for post-stroke rehabilitation therapy," in *Proceedings - IEEE International Conference on Robotics and Automation*, 2007, pp. 1547–1553. DOI: 10.1109/ROBOT.2007.363544.

[2] C. D. Kidd *et al.*, "Robots at home: Understanding long-term human-robot interaction," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 3230–3235. DOI: 10.1109/IROS.2008.4651113.

[3] D. Leyzberg *et al.*, "Personalizing robot tutors to individuals' learning differences," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, ser. HRI '14, Bielefeld, Germany: ACM, 2014, pp. 423–430. DOI: 10.1145/2559636.2559671.

[4] I. Leite *et al.*, "Social Robots for Long-Term Interaction: A Survey," *International Journal of Social Robotics*, vol. 5, no. 2, pp. 291–308, 2013. DOI: 10.1007/s12369-013-0178-y.

[5] I. Leite *et al.*, "Modelling empathy in social robotic companions," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7138 LNCS, 2012, pp. 135–147. DOI: 10.1007/978-3-642-28509-7_14.

[6] N. Mitsunaga *et al.*, "Robot behavior adaptation for human-robot interaction based on policy gradient reinforcement learning," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 218–225. DOI: 10.1109/IROS.2005.1545206.

[7] K. Tsiakas *et al.*, "An Interactive Learning and Adaptation Framework for Adaptive Robot Assisted Therapy," in *PETRA*, 2016. DOI: 10.1145/2910674.2935857.

[8] G. H. Lim *et al.*, "Robot recommender system using affection-based episode ontology for personalization," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 2013, pp. 155–160. DOI: 10.1109/ROMAN.2013.6628437.

[9] K. Baraka *et al.*, "Adaptive interaction of persistent robots to user temporal preferences," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9388, 2015, pp. 61–71. DOI: 10.1007/978-3-319-25554-5_7.

[10] J. Hemminghaus *et al.*, "Towards adaptive social behavior generation for assistive robots using reinforcement learning," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ACM, 2017, pp. 332–340.

[11] J. Chan *et al.*, "Social intelligence for a robot engaging people in cognitive training activities," *International Journal of Advanced Robotic Systems*, vol. 9, no. 4, p. 113, 2012.

[12] M. K. Lee *et al.*, "Personalization in hri: A longitudinal field experiment," in *7th ACM/IEEE International Conference on Human-Robot Interaction*, 2012.

[13] P. Auer *et al.*, "The nonstochastic multiarmed bandit problem," *SIAM journal on computing*, vol. 32, no. 1, pp. 48–77, 2002.

[14] P. Auer *et al.*, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[15] R. Busa-Fekete *et al.*, "A survey of preference-based online learning with bandit algorithms," in *International Conference on Algorithmic Learning Theory*, Springer, 2014, pp. 18–39.

[16] D. C. Kingsley, "Preference uncertainty, preference refinement and paired comparison choice experiments," *Dept. of Economics. University of Colorado*, 2006.

[17] I. B. Sebastian Schneider Luise Sssenbach *et al.*, "Long-term feedback mechanisms for robotic assisted indoor cycling training," in *Proceedings of the 3rd Conference on Human-Agent Interaction*, 2015.

[18] S. Schneider *et al.*, "Exercising with a humanoid companion is more effective than exercising alone," in *Proceedings of the IEEE-RAS Conference on Humanoid Robots*, 2016.

[19] J. Li, "The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents," *International Journal of Human-Computer Studies*, vol. 77, pp. 23–37, 2015.

[20] A. Sekmen *et al.*, "Assessment of adaptive human-robot interactions," *Knowledge-Based Systems*, vol. 42, pp. 49–59, 2013. DOI: 10.1016/j.knosys.2013.01.003.

[21] Y. Yue *et al.*, "The k-armed dueling bandits problem," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1538–1556, 2012.

[22] H. Wu *et al.*, "Double thompson sampling for dueling bandits," in *Advances in Neural Information Processing Systems*, 2016, pp. 649–657.

[23] A. Körner *et al.*, "Persönlichkeitsdiagnostik mit dem neo-fünf-faktoren-inventar: Die 30-item-kurzversion (neo-ffi-30)," *PPmP-Psychotherapie· Psychosomatik· Medizinische Psychologie*, vol. 58, no. 06, pp. 238–245, 2008.

[24] C. Bartneck *et al.*, "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.

[25] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[26] C. Nass *et al.*, "Can computers be teammates?" *International Journal of Human-Computer Studies*, vol. 45, no. 6, pp. 669–678, 1996.

[27] E. L. Deci *et al.*, *The intrinsic motivation inventory*, retrieved June 6, 2015, 2006.

[28] A. M. Rosenthal-von der Pütten *et al.*, "Robots or agents–neither helps you more or less during second language acquisition," in *International Conference on Intelligent Virtual Agents*, Springer, 2016, pp. 256–268.

[29] J. Kennedy *et al.*, "Comparing robot embodiments in a guided discovery learning interaction with children," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 293–308, 2015.

[30] D. Leyzberg *et al.*, "The physical presence of a robot tutor increases cognitive learning gains," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society*, 2012.