International Journal of Social Robotics
https://doi.org/10.1007/s12369-024-01124-2

Personalizing Activity Selection in Assistive Social Robots from Explicit and Implicit User Feedback

Marcos Maroto-Gómez[1] · María Malfaz[1] · José Carlos Castillo[1] · Álvaro Castro-González[1] · Miguel Ángel Salichs[1]

Accepted: 11 March 2024

Abstract
Robots in multi-user environments require adaptation to produce personalized interactions. In these scenarios, the user's feedback leads the robots to learn from experiences and use this knowledge to generate adapted activities to the user's preferences. However, preferences are user-specific and may suffer variations, so learning is required to personalize the robot's actions to each user. Robots can obtain feedback in Human–Robot Interaction by asking users their opinion about the activity (explicit feedback) or estimating it from the interaction (implicit feedback). This paper presents a Reinforcement Learning framework for social robots to personalize activity selection using the preferences and feedback obtained from the users. This paper also studies the role of user feedback in learning, and it asks whether combining explicit and implicit user feedback produces better robot adaptive behavior than considering them separately. We evaluated the system with 24 participants in a long-term experiment where they were divided into three conditions: (i) adapting the activity selection using the explicit feedback that was obtained from asking the user how much they liked the activities; (ii) using the implicit feedback obtained from interaction metrics of each activity generated from the user's actions; and (iii) combining explicit and implicit feedback. As we hypothesized, the results show that combining both feedback produces better adaptive values when correlating initial and final activity scores, overcoming the use of individual explicit and implicit feedback. We also found that the kind of user feedback does not affect the user's engagement or the number of activities carried out during the experiment.

Keywords Social robots · Reinforcement Learning · Personalized Robots · Human–Robot Interaction · Multi-armed Bandits

1 Introduction

Human–Robot Interaction (HRI) explores how to facilitate the communication between humans and robots, improve their use, and personalize their behavior to each user [9]. Personalized robot behavior drives these machines to establish bonds with their users based on acceptance and trust [6]. In this context, Reinforcement Learning (RL) has gained attention in the last few years because it allows robots to learn from user feedback and explore the environment, thus producing adaptive and personalized behavior. RL methods have opened many new opportunities in social robotics, a field where HRI typically undergoes unforeseen changes and requires adaptation [2]. Nonetheless, HRI still faces many challenges, especially when the robot needs to interpret the user's feedback, preferences, and intentions [32]. Thereby, it is fundamental that the user feedback influences the robot's actions to correctly learn and succeed in the interaction [22].

This paper presents a RL framework for social robots to produce autonomous decision-making and drive entertainment or cognitive stimulation sessions using the user's preferences. The model, shown in Fig. 1, considers implicit

✉ Marcos Maroto-Gómez
marmarot@ing.uc3m.es

María Malfaz
mmalfaz@ing.uc3m.es

José Carlos Castillo
jocastil@ing.uc3m.es

Álvaro Castro-González
acgonzal@ing.uc3m.es

Miguel Ángel Salichs
salichs@ing.uc3m.es

[1] Systems Engineering and Automation Department, University Carlos III of Madrid, Butaque, 15, 28912 Leganés, Madrid, Spain

Published online: 09 April 2024

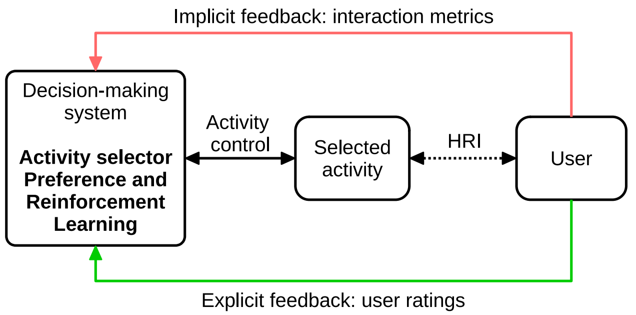
Implicit feedback: interaction metrics
Decision-making system
Activity selector Preference and Reinforcement Learning
Activity control
Selected activity
HRI
User
Explicit feedback: user ratings

Fig. 1 The Reinforcement Learning system for preference learning is used to personalize activity selection from the user's implicit and explicit feedback. The Decision-making system selects and controls the activities executed by the robot. During HRI, the robot obtains implicit feedback from how the user interacts and explicit feedback from the activity ratings. The model in the Decision-making system receives feedback to know the preferred activities and play them more often

and explicit user feedback to evaluate the system's performance and role in long-term HRI. The system can be applied to generate autonomous and personalized decision-making in any robot, dynamically generating adapted sessions.

In previous works [16, 18, 19], we developed decision-making architectures for autonomous social robots in cognitive stimulation and entertainment that make decisions by simulating an artificial biological state that drives motivated behavior. Nonetheless, these architectures barely consider the user's preferences in the decision-making process. More recently, we presented methods based on prediction [17] and conceptual models [20], but they needed to include the adaptive behavior and experiments included in this paper. The social robotics community agrees that social robots must be user-oriented to engage users and facilitate their use [12]. Consequently, this work concentrates on developing robot learning methods to personalize online activity selection during entertainment sessions by autonomously selecting the user's favorite activities more often.

The user's features are often unknown by the robot at the beginning of the interaction, so it does not have any information about how to personalize behavior. In these situations, user preferences must be obtained from the interaction experiences while adapting the robot's behavior step-by-step. Social robots have two ways to gain this experience from user feedback. On the one hand, they can ask the user to rate how much they liked the activity after executing it (*explicit feedback*). On the other hand, it can estimate the rating from the interaction metrics (*implicit feedback*).

The literature has previously employed explicit and implicit user feedback to improve HRI [4, 8, 13, 27, 35, 39]. However, we have failed to find a previous work that explores the impact of the user's feedback when learning the user's preferences to personalize the robot's behavior, and whether explicit and implicit feedback should be combined or used separately. Consequently, this work explores the impact of dynamically learning to personalize robot behavior from online user feedback, the implications of feedback on user engagement, and the influence on the number of activities during the sessions.

We evaluated the system through a long-term study with 24 participants (6 women and 18 men). Initially, the participants indicated their preferences from 0 to 5 points toward 15 multimedia activities using an online survey. Each participant performed at least 5 sessions of 20 min each (minimum interaction time of 100 min). The participants were equally divided into three conditions. These conditions investigated which feedback method produces the best robot adaptive behavior and whether the feedback method influences user engagement and the number of activities executed per session. The three conditions are:

- **Condition 1: Explicit feedback (C1).** The user preferences were updated using only explicit feedback. After executing an activity, the robot asked the user how much they liked it to obtain a rating on a 0 to 5-point scale.
- **Condition 2: Implicit feedback (C2).** The user preferences were updated using implicit feedback. The robot considered three interaction metrics to estimate a value from 0 to 5 points to indicate how much the user likes a particular activity.
- **Condition 3: Combined feedback (C3).** The user preferences were updated by combining explicit and implicit user feedback.

The definition of the experiment conducted to assess the role of user feedback led us to hypothesize:

(H1) C3 should be the best alternative in terms of preference adaptation, producing lower error, improving user engagement, and balancing activity exploration.

(H2) Combining the user's ratings (explicit feedback) with interaction metrics (implicit feedback) should lead the learned preferences to be more similar to the initial ratings of the users

(H3) Adapting the user's preferences using explicit feedback (C1) should produce accurate learning. However, asking the users for their responses after each activity will reduce the number of activities tested.

This manuscript is organized as follows. Section 2 reviews the current state of social robots, focusing on learning preferences from user feedback to state the gap that this paper addresses. Section 3 formalizes the RL problem focusing on Multi-armed Bandits applied to constant learning conditions. Section 4 describes the experiment to test the performance of the learning system. Section 5 presents the experiment results, comparing the learning system outcomes for the three conditions. Section 6 discusses the outcomes of this work and states its limitations. Finally, Sect. 7 provides the main conclusions.

2 Related Work

In the last few years, RL has been successfully applied to dynamic environments to provide social robots with adaptive behavior in applications such as social navigation [8, 15], education [7], and assistance [28]. Nonetheless, the literature contains only a few contributions that address adaptation from explicit and implicit user feedback in social robotics to personalize HRI sessions.

In this line of research, Baraka and Veloso [4] designed one of the first studies to generate personalized entertainment sessions. In particular, the authors studied the role of explicit and implicit user feedback in learning preferences. In their study, which is very close to the work presented in this manuscript, they classify users into different profiles (i.e., conservative, erratic, and consistent but fatigable) and study how to learn user preferences in simulation. However, their work was not tested with human users, and their definition of user feedback needs to be practically proposed. Similarly, Whitney et al. [39] presented a robot that reduces its errors in an object-fetching task by using explicit and implicit feedback from the user. In this case, the robot only asks the user for information when necessary, which avoids fatigue and increases the fluency of the interaction. The difference in this work was testing the system with just a few robot actions and short-term interactions without learning.

Hemminahaus and Kopp [14] investigated how a social robot can adapt its social behavior while interacting with humans to attain specific goals in unpredictable situations. The adaptive process uses RL and implicit user feedback during a memory game assistive task to improve the robot's decision-making. In a similar work, Moro et al. [21] proposed an RL-supported system that learns personalized behavior in daily assistive activities by considering the user's implicit feedback. Meanwhile, Ritschel and André [25] used RL to dynamically adapt their robot behavior to the human's personality profile to make the interaction more engaging. This setting employs implicit feedback because the robot does not explicitly ask the user about their preferences but instead uses social signals to estimate their level of introversion or extroversion. The primary difference between these works and ours is that the robot uses predefined short-term scenarios and only considers implicit feedback.

Later, Ritschel et al. [27] presented a robot that can adapt its communicative acts while performing activities such as information retrieval, reminders, communication, and entertainment and giving health-related recommendations. The adaptive process gathers explicit user feedback obtained during the interaction and uses an Upper Confidence Bound action selection method supported by RL to improve the process.

In the last few years, social robots have been used to drive HRI sessions with older adults and children. Wang et al. [38] studied the impact of service robots in interactions with older people by focusing on their interface preferences in multi-modal communication procedures. In education environments, RL has also been used in HRI by Park et al. [24] to help children improve their language skills. In this case, the robot gathers verbal and non-verbal feedback from the children to modulate their engagement and maximize their learning gain. Che et al. [8] presented a mobile robot that can produce efficient social navigation by combining explicit and implicit user feedback. This setting resembles our work because it conceptually uses explicit and implicit feedback to produce appropriate behaviors. However, the differences in our approach reside in applying preference learning in mobile instead of social robotics and the lack of real experiments to validate the impact of feedback on the robot's performance. In a similar scenario, Shi et al. [33] have shown the great potential of adaptive social assistive robots during long-term interventions for children with autism spectrum. However, user feedback plays a shallow role in this work because most session parameters are predefined, and only facial gestures are considered.

Focused on personalizing HRI, Tsiakas et al. [35] proposed an Interactive Reinforcement Learning framework combining explicit feedback from task performance and implicit feedback from user engagement. Their results show that combining explicit and implicit feedback drives real-time HRI personalization. The 69 participants interacted with the robot in a single session to evaluate if factors like exercise level or engagement were appropriately personalized. Olatunji et al. [23] studied the design of effective feedback strategies in person-following robots with older adults. Their results show how users preferred voice feedback over tone at a continuous rate to receive information about the robot's actions constantly. Akalin et al. [1] explored how different robot feedback (negative, positive, and flattering) influences the users' perception in cognitive training tasks. The results show how flattering and positive feedback was preferred. However, the study does not describe implicit or explicit feedback from the user. Besides, the evaluation was carried out in a single session. The main differences found in the previous works are in not evaluating the system in long-term experiments and analyzing the role of user feedback on the task.

Boggess et al. [5] developed a system to generate personalized explanations for path planning from user preferences. The robot answers the users' question using HRI and, using RL, define the best strategy for each situation. The main difference of this work is not using implicit feedback and evaluating the method using an online survey instead of real interactions. Recently, Asprino et al. [3] presented a software architecture considering adaptive behavior from user preferences. In this task, the architecture obtains explicit user feedback before interacting to store the user's favorite activities, which are later autonomously presented to the user. However, this paper does not provide an evaluation, and neither online adaptation nor user feedback are considered.

As Wirth et al. [40] have recently reviewed, numerous works have employed RL to produce a set of preferences toward a group of actions/activities. For this application, the authors present Multi-armed Bandits as an effective alternative to organize activities as a ranking based on the values of a model-free tabular RL algorithm. Many authors have deeply explored these methods [26, 30, 37] in the last years to personalize the interaction of social robots. These contributions agree in ordering a set of labels (in most cases, activities) to select those preferred by the user more often. Considering the positive results of these studies in HRI, we opted for using a Multi-armed Bandit action-based method in a non-stationary scenario with a constant learning rate.

Adaptive systems in HRI have a great potential to improve the interaction significantly. The analysis of the previous literature review provided in Table 1 highlights the lack of adaptive robots for long-lasting interactions that consider and analyze the role of explicit and implicit user feedback. The review shows some works [4, 8, 17, 35, 39] that personalize HRI from explicit and implicit user feedback. Besides, some works [5, 17, 21, 24, 26] include RL to improve further interactions from past experiences. However, none of these studies investigate the role of explicit and implicit feedback on learning preferences and whether user feedback influences the personalization and learning process in long-term experiments with onsite participants. Acknowledging this gap, this paper presents a framework that (i) generates online adaptive behavior during long-term sessions from RL, (ii) considers implicit feedback obtained from the user's actions during the interaction, (iii) considers explicit feedback by asking the users their preferences, and (iv) explores the influence of feedback in user engagement and activity execution.

3 Methods

This section formalizes the RL methods based on non-stationary Multi-armed Bandits [34, p. 32]. Action refers to the robot selecting and executing an activity.

3.1 Formalization

RL is a learning method that allows an agent to learn how to map situations to actions to maximize a numerical reward signal representing a goal [34, p. 1]. Initially, the agent does not know their action effects but can explore them by continuous interaction (i.e., trial and error) with the environment. When an action finishes, its effects are "perceived" by the agent and then converted to a numerical reward value that measures the action's quality in the agent's situation. Considering the previous idea, it is possible to infer that the reward function has to be predefined by the designer because it is specific to the learning scenario and dependent on the goal that the agent seeks to attain (see Sutton and Barto [34] for a review of reward function shaping).

Formally, RL problems are generally Markov Decision Processes (MDP) [36] that consider features of the agent's situation to develop a probabilistic model that is based on transition probabilities from one state (situation) to another. The transition occurs when the agent executes an action selected from a list of possibilities. Among RL methods, Monte Carlo, Temporal Difference, or Dynamic Programming algorithms consider a variant agent state learning in which action better suits each situation. Meanwhile, methods such as Multi-armed Bandits consider a constant agent state but focus on learning action values (i.e., optimal action execution) using environmental feedback. Although both streams have remarkable differences, TD and action-value methods have similarities in updating the values assigned to each state-action pair in MDPs or action-value methods. Both approaches consider the error between a target estimated value in opposition to previous agent estimations. Thus, an error appears when the current estimate differs from the previous knowledge, which drives its correction. The error correction is performed by moving a step toward the target. Equation (1) expresses this idea,

$$\text{Value} \leftarrow \text{Oldvalue} + \text{StepSize}\left[\underbrace{\text{Target} - \text{Oldvalue}}_{error}\right] \tag{1}$$

where the StepSize indicates the amplitude of the error correction towards the Target value using the old value and the new value.

This learning scenario requires the robot to learn a behavior policy (i.e., a sequence of actions) whose goal is to fulfill the goal defined by the reward function by maximizing the reward obtained after each action in a fixed agent state. Following Sutton and Barto's [34, p. 33] ideas, we opted for using Multi-armed Bandits since they are a simple and efficient solution to learn how well a specific action is executed in a static agent state. This method allows us to compare the most suitable action from a group of possibilities in preference selection. Equation 2 shows the original formulation for Multi-armed Bandits considering non-stationary rewards and variable step size.

Table 1 Related contributions to our method analyzing the similarities and differences for adaptive robot behavior for HRI sessions

Paper	Application	Similarities	Differences
Baraka and Veloso [4]	Preference Learning	Personalize entertainment sessions from explicit and implicit user feedback	Not tested with users, only tested in simulation
Whitney et al. [39]	Error reduction in HRI	Considers explicit and implicit user feedback	Short-term interactions. Limited robot action space
Hemminahaus and Kopp [14]	Assistive gaming	Adaptive behavior. Implicit feedback	Explicit feedback not used. Short-term interactions
Moro et al. [21]	Personalized sessions	Includes RL. Implicit feedback from social cues	Only implicit feedback. Predefined short sessions
Ritschel and André [25]	Robot personality adaptation	Includes RL. Implicit feedback from social cues	Only implicit feedback. Short-term evaluation
Ritschel et al. [27]	Adapt robot linguistic style	Adapted to user preferences. Includes RL	Only uses explicit feedback. Short-term interactions
Wang et al. [38]	HRI for older adults	Multi-modal adaptation. User preferences	Only implicit feedback. Short-term tests
Park et al. [24]	Children’s education	Includes RL	Only implicit feedback
Che et al. [8]	Social navigation	Both feedback types	Lack of real experiments
Shi et al. [33]	Autistic children	Multi-modal long-term adaptation	Only implicit feedback from facial cues
Maroto et al. [17]	Entertainment sessions	Preferences prediction	No adaptive behavior. No user feedback
Maroto et al. [20]	Entertainment sessions	Considers explicit and implicit feedback. Included RL learning	Conceptual model. No user validation
Asprino et al. [3]	Social robot architecture	User preferences. Explicit feedback	No implicit feedback. No user tests
Boggess et al. [5]	Path planning	Personalized explanations for path planning based on RL	Online user study. Only explicit feedback from questions.
Tsiakas et al. [35]	Cognitive training	Implicit and explicit feedback with RL	Unique session with the robot
Akalin et al. [1]	Robot-assisted training	Evaluates robot feedback and personalization in cognitive training	Robot feedback but not user feedback. One session for evaluation.
Olatunji et al. [23]	Person-following with older adults	Precise experiment definition about robot feedback	Implicit and explicit feedback not defined. No learning

$$Q(a) \leftarrow Q(a) + \frac{1}{N(a)} [r - Q(a)] \qquad (2)$$

In this equation, $Q(a)$ is a float value that represents how good it is to execute the action a, and the step size α is a function of the number of updates $N(a)$ expressed as $\frac{1}{N(a)}$. This change allows convergence because the error decreases with the action's number of updates $N(a)$.

3.2 Proposed Learning System

This paper's proposed learning system employs Multi-armed Bandits [34, p. 25] considering a constant learning rate. These methods can be applied to our problem due to two main properties:

- The learning process is continuous since the learning rate does not change, and learning occurs during the whole lifespan of the robot.
- In this set-up, the robot learns activity preferences instead of state transition suitability, as presented in Eq. (2). Thus, Eq. (2) deals with learning the best actions in a non-stationary scenario with a fixed agent state.

We opted for using Eq. (3) in our learning model. This equation is based on Eq. (2), setting a constant learning rate α.

$$Q(a) \leftarrow Q(a) + \alpha [r - Q(a)] \qquad (3)$$

In RL, the learning rate α often depends on the number of updates of the Q-value associated with the action $N(a)$. The original algorithm proposed in [34] and presented in the previous section considers this setting to converge to an optimal solution. However, converging on an optimal solution is not necessary in our application, where the user's preferences may vary in the long term. Instead, we propose that the learning system continuously adapts to the user's preferences. To find the best α value, we conducted a preliminary evaluation to choose between four empirically selected alternatives: 0.1, 0.25, 0.5, and 1. The preliminary evaluation consisted of simulating the learning system's dynamics using the four rates to update a single action during 20 iterations with different rewards. The results of this evaluation reported that $\alpha = 0.5$ was the best alternative because the learning rates of 0.1 and 0.25 produced prolonged adaptation, and the learning rate of 1 unit could not fit well to the initial user rating of the activity.

The Q-values representing the user preferences $Q(a)$ range from 0 to 5 points. Meanwhile, 0 indicates that the user does not like the activity, whereas 5 indicates that the user loves the activity. All action values $Q(a)$ start in 5 points to allow activity exploration at the beginning of the experiment and select preferred activities more often as the experiment progresses. The reward value can only be between 0 and 5 points to keep the values in the range. Recall that the user's feedback can be obtained from the activity ratings (explicit feedback) and interaction metrics (implicit feedback).

4 Experiment

This section describes the experimental setup of this work. It introduces the Mini social robot used to test the system's performance. Then, it describes the experimental setup and the session dynamics. Finally, we describe the robot's actions in the learning scenario.

4.1 Mini Social Robot

Mini [29] is a social desktop robot that assists older adults in cognitive stimulation therapies and entertainment sessions. Mini communicates with the users using a HRI manager [11] that manages the verbal and non-verbal interaction and obtains the user's feedback using perception to adapt its behavior to the situation it is experiencing. The user executes the robot's activities using a connected touch screen. The Decision-making system [16] manages activity selection, which employs the user's preferences to personalize the interaction.

4.2 Experimental Setup

The experiments were conducted to validate our approach, which consisted of comparing the learning system's performance under three conditions that define how the reward function is updated based on user feedback. In the experiment, we recruited 24 university students with little expertise in robotics (6 women and 18 men) aged from 20 to 30 years old ($\mu = 24.55$, $\sigma = 2.75$). These students had not previously interacted with the robot. They were equally and randomly divided into one of the three conditions to execute entertainment sessions that aim at completing activities related to watching photos, videos, and listening to music. This task was selected considering the application we want to give to the method: learning user preferences towards the robot's activities to personalize future activity selection and the versatility and repertoire that multimedia activities offer. Each session had a minimum allotted time of 20 min per session. The robot autonomously ended the session when the 20 min had elapsed, and the ongoing activity finished. Since the participants required four weeks to complete the experiment, the sample size was limited to 8 people per condition.

Mini is a desktop robot. Therefore, in the experiments, it was fixed to an office table where the participants individually interacted with Mini without the intervention of other people. The sessions were face-to-face interactions, as shown in Fig. 2. Each participant tested the robot's activities as described in

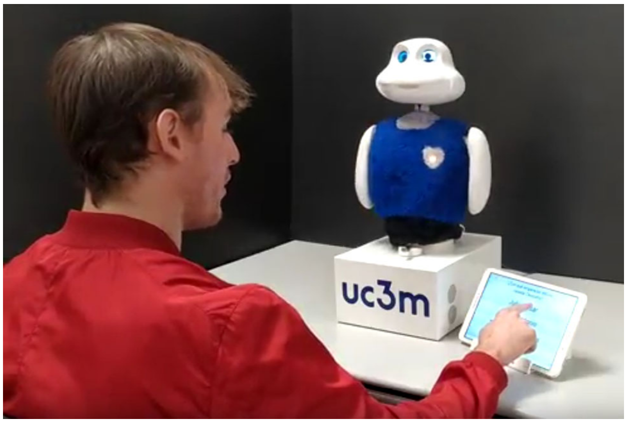

Fig. 2 Mini executing an activity with a user during the entertainment experiment

Sect. 4.3, participating only in one of the conditions (between subjects study). The participants decided when to interact with the robot and had 20 days (from Monday to Friday for four consecutive weeks) to complete a minimum of five sessions (total time of 100 min). The number of sessions (5) and the time per session (20 min) were set arbitrarily based on a previous user study we conducted with the robot [17]. However, all participants voluntarily completed more than the requested sessions, as indicated in Table 7. The conditions under evaluation were:

- **Condition 1: Explicit feedback (C1).** Only explicit user feedback was considered when updating the Q-values associated with each activity. Independently of the activity result, the user was requested to rate the activity once finished using a 0 to 5 point scale.
- **Condition 2: Implicit feedback (C2).** The users never rated an activity, but instead, implicit feedback was calculated using the interaction metrics to estimate how much the user liked the activity.
- **Condition 3: Combined feedback (C3).** This condition joins both of the previous approaches. In this scenario, the robot autonomously decides whether it asks the user to rate the activity after finishing it (the probability of asking is 50%). The reward value is predicted by considering the interaction metrics.

4.3 Session Dynamics

As shown in Fig. 3, the session dynamics followed the same course with subtle differences in the three conditions presented earlier. Before starting the experiment and testing the robot's activities, each participant completed an online survey to rate how much they liked different photos, videos, and music activities using a 0 to 5 point scale. These ratings were later used as a baseline to compare the initial and predicted preferences.

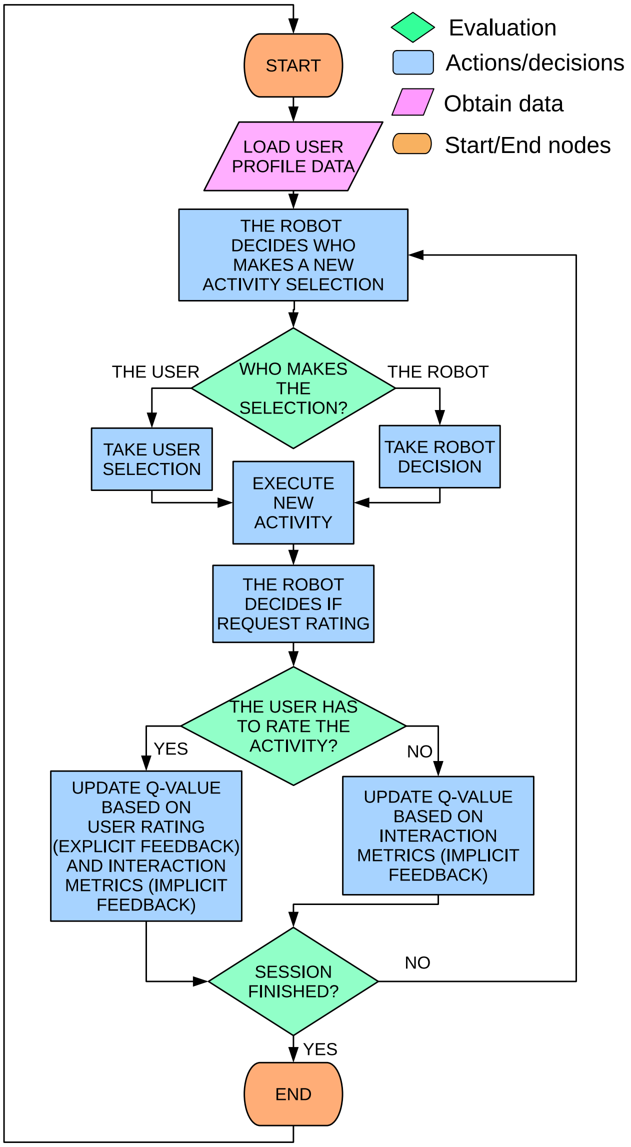
Evaluation
Actions/decisions
Obtain data
Start/End nodes
START
LOAD USER PROFILE DATA
THE ROBOT DECIDES WHO MAKES A NEW ACTIVITY SELECTION
WHO MAKES THE SELECTION?
THE USER
THE ROBOT
TAKE USER SELECTION
TAKE ROBOT DECISION
EXECUTE NEW ACTIVITY
THE ROBOT DECIDES IF REQUEST RATING
THE USER HAS TO RATE THE ACTIVITY?
YES
NO
UPDATE Q-VALUE BASED ON USER RATING (EXPLICIT FEEDBACK) AND INTERACTION METRICS (IMPLICIT FEEDBACK)
UPDATE Q-VALUE BASED ON INTERACTION METRICS (IMPLICIT FEEDBACK)
SESSION FINISHED?
NO
YES
END

Fig. 3 Flow diagram representing the experiment dynamics. The experiment starts by loading the user profile data, including preferences. Then, the robot decides if the user selects the next activity or itself. Once the activity finishes, the Q-value is updated using the implicit and explicit (if obtained) user feedback

At the beginning of the first session, the robot informed the users about the experiment's dynamics and the need to complete at least five sessions in four weeks while allowing them to execute more. At the beginning of each session, the participants had to press a *Start* button and select their name on the touch screen to load their information into the robot's memory. This profile contained basic personal information about the participants' features (e.g., age or name), which the designers previously included. The profile also included the participant's preferences towards the robot activities, with an initial value of 5 points adapted by the learning algorithms with the interaction. Thus, we ensured that all activities had the same probability of being selected at the beginning of the experiment.

The participants or the robot could select the activities in each iteration. The probability of the robot or the user selecting the activity was 50% in each case. Thus, depending on who made the decision, the activities could be selected in two ways:

- If the user selects the activity, a menu appears on the touch screen with the activities classified under the categories photos, videos, and music. The user can navigate these menus and select the robot's following activity.
- If the robot autonomously selected the activity, then the robot notified the next activity before starting. In this case, the robot employed the Boltzmann distribution [10] with a Temperature value set to 5 points to select the user's preferred activities more often and foster unexplored activities using the learned Q-values. We also foster less visited activities to be selected more often, increasing their probability.

The likelihood of preferred and less selected activities increased as the experiment progressed. Meanwhile, the selection probability of those activities often explored by participants with low ratings was substantially reduced. It is also important to remark that the user could stop the activity by touching the robot's right-hand shoulder. At that moment, the activity was paused, and the robot waited for the user's confirmation using the touch screen. After canceling an activity, the user or the robot could select a new activity if the session time was below 20 min.

When an activity finishes, three possible events occur based on the evaluation condition:

1. If the participants were in **Condition 1**, then they were requested to rate the activity using a 0 to 5 point scale with 100% chance.
2. If the participants were in **Condition 2**, then the user never rated the activity, and interaction metrics were used to update the activity Q-value.
3. If the participants were in **Condition 3**, then they were occasionally requested to rate the activity (50% chance). The interaction metrics were also used.

If the robot detected the user's inactivity (i.e., not answering the questions), then the session finished. This issue happened only once during the experiments since one user had to leave due to personal problems. This session was removed from the data and not considered in the analysis. The robot then returned to an idle state and waited for a new participant to press the *Start* button to begin a new session.

4.4 Obtaining User Feedback

Mini has two ways of obtaining user feedback and getting a numerical reward to update the user's preferences.

- *Explicit feedback* is obtained from the user ratings using the touch screen.
- *Implicit feedback*, estimated from the interaction using predefined metrics.

Both alternatives yield a numerical reward to update the previously executed Q-value associated with the activity. The reward value obtained after each activity ranges from 0 to 5 points, which limits the Q-value associated with each activity inside this range. On the one hand, the robot obtains explicit feedback by asking the user to rate how much they liked the activity from 0 to 5. When obtaining explicit feedback, the robot asked the user *How much you like to listen to/watch...?*, ending with the name of the activity just performed. Then, using the touch screen, the user can rate the activity from 0 to 5. Equation (4) shows how the numerical value associated with the explicit feedback is calculated.

$$r_{explicit} = \text{User rating in } \{0, 1, 2, 3, 4, 5\} \tag{4}$$

On the other hand, the robot obtains implicit feedback by estimating the numerical reward using customized parameters that define how good the interaction between both agents was while executing an activity. We defined three interaction metrics, which jointly represent the quality of the interaction process. These metrics are 0 or 1, depending on whether its associated definition is false or true. The three metrics that are used in this work and their related conditions are:

- User Selection (US): This metric represents if the user selected the activity (1) or if it was autonomously proposed by the robot (0).
- Activity Result (AR): These activities can have two possible outcomes: succeeded, in which case the value of this metric is 1; or aborted, which is provoked by the user when they voluntarily cancel the activity, in which case the value of the metric is 0.
- Execution Time (ET): This metric represents if the activity's execution time is similar to the execution times of other participants. To validate this condition, we calculate the mean execution time of the activity considering all users μ and its standard deviation σ. Then, we check if the current execution time is within the interval $[\mu - \sigma, \mu + \sigma]$. If it is, then the value of the metric is 1. Otherwise, the value is 0.

Equation (5) shows the reward value associated with the implicit feedback using the interaction metrics presented earlier. It is worth noting here that the interaction metrics *user selection* and *activity result* have a double influence on the reward when compared to *execution time* because we consider them to be more relevant and reliable in our scenario. However, in other scenarios, interaction metrics can be different.

$$r_{implicit} = 2 \cdot US + 2 \cdot AR + ET \tag{5}$$

Finally, if the reward value is calculated by combining the explicit and implicit feedback (C3), then it is the average value between the explicit ($r_{explicit}$) and implicit ($r_{implicit}$) feedback. Otherwise, if the explicit feedback is not obtained because the question is not asked the user, then the combined feedback is the implicit feedback, as Eq. (6) shows.

$$r_{combined} = \begin{cases} \frac{(r_{explicit}+r_{implicit})}{2} & \text{if question} \\ r_{implicit} & \text{if not question} \end{cases} \tag{6}$$

4.5 Activities

The learning system aims to obtain the user's preferences regarding the entertainment activities of the Mini robot. As mentioned earlier, the learning process adapts by obtaining explicit and implicit feedback from the user after executing each activity. Thus, learning can only succeed if the user executes each activity many times so that the robot can acknowledge how much the user likes each activity.

In the task that we designed to evaluate the learning system and the role of user feedback, we opted for the activities related to displaying multimedia content because they are easy to use and offer versatility and diversity in their execution since they have different options for each type of activity. The activities were classified into the categories *photos*, *music*, and *videos*.

Photos category have the activities of *animals*, *monuments*, *landscapes*, or *sad moments*. Music can be *Spanish pop*, *Spanish rock*, *English pop*, *English rock*, *Latin*, or *noise*. Finally, videos can be about *cooking recipes*, *funny moments*, *sport*, *film trailers*, and *comedy*.

The photos category displays eight photos for 5 s each. The total duration of the activities 40 s. The photos were downloaded from Google to create a database of around 1000. Each photo activity has a similar number of photos (around 250 each). The video category consists of displaying a single video for around 3 min. We downloaded the videos from YouTube to create a database of around 110 videos equally divided into the previous categories. Finally, the music category consists of playing a song selected from a database that stores around 90 songs equally sorted in activities. The duration of the music's activities is around 3 min. The item was selected inside each activity randomly, but remembering the last five items was selected to reduce the repetition chance.

The activities *sad moments* and *noise* inside the *photos* and *music* categories were included to have two activities that we believe the participants will dislike. Thus, we expect that they will give a negative evolution to the Q-values adaptation compared to the activities the participants typically like. Besides, using a big database and dividing activities into categories provide diversity so users have activities that they can like and dislike.

4.6 Evaluation

The evaluation of the learning performance exhibited by the robot was carried out in two stages.

1. First, we used the *Root Mean Squared Error (RMSE)* and the *Spearman correlation* to statistically report which condition produced the best adaptation to the initial preferences obtained from the online survey. This analysis was carried out for all activities in each category (i.e., Photos, Videos, and Music).
   The RMSE measures the absolute difference between observed and predicted values, which strongly indicates the deviation between two samples. Preference values range from 0 to 5 points, and RMSE is used to compare the three ways of learning user preferences. According to the range of our data, RMSE values below 0.5 units can be considered excellent, from 0.5 to 1 as moderately positive and above 1 as high.
   The Spearman correlation is a metric applied to non-normal distributions to obtain the degree of relationship between two variables. This metric ranges from -1 to 1, distinguishing between explicit and inverse relations. The lower RMSE and Spearman values close to 1 indicate a strong relationship [31] between the observed and predicted values.
   The final correlation scores are the average correlation values of the 8 users participating in each condition. The correlation score of each user is computed considering the initial preferences and learned values for all 15 activities, only the photos activities (4), only the video activities (5), and only the music activities (6).
2. Second, we statistically analyzed whether using different types of feedback led the users to interact more often with the robot and if this affected the number of updates of the activities. We employed the one-way ANOVA test to look for significant differences in the users' engagement between the three conditions, analyzing the impact of the user feedback.

3. We used the Kruskal-Wallis statistical test to determine whether the number of times each activity was updated was affected by how the robot obtained user feedback. This test is carried out for non-normally distributed small samples.

5 Results

This section presents the main results of the experiment described in the previous section. We statistically compare the three conditions used to adapt the Q-values, focusing on the error yielded by each approach and the correlation metrics. Additionally, we statistically analyze whether user feedback influences the interaction time as an indicator of increased engagement or if it is affected by the number of times the activities were updated. Table 2 shows a summary of the study participants and the main outcomes obtained in the study.

Table 2 Summary showing the details of the participants in this condition and the results of the experiment

Feature	C1	C2	C3	All
N participants	8	8	8	24
Age range	22 to 29	20 to 28	22 to 30	20 to 30
Mean age	24.88	23.5	26.13	24.55
Min sess. time	20.12	20.35	20.21	20.12
Max sess. time	24.32	23.68	23.79	24.32
Avg sessions	8.78	8.21	8.06	8.36
Avg HRI time (min)	175.63	164.38	161.38	167.13
Avg activ. user/sess.	5.90	6.93	9.51	7.45
Avg activ. user	88.50	103.95	142.75	111.75
RMSE	1.355	1.662	0.772	–
Spearman corr.	0.567	0.178	0.811	–

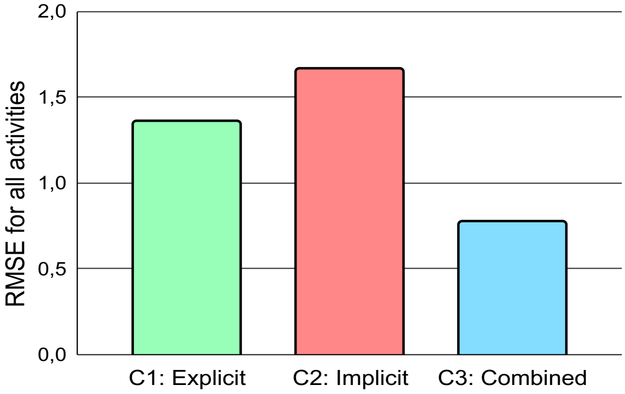

Fig. 4 RMSE when comparing the initial and predicted preference values for all the activities in each condition

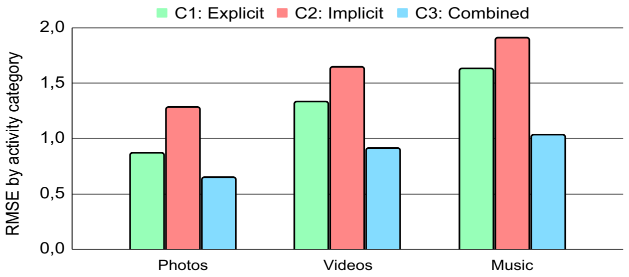

Fig. 5 RMSE value for each condition and type of activities when comparing the initial and predicted user preferences

5.1 RMSE and Spearman Correlation

The methodology presented in this manuscript was evaluated by comparing which kind of feedback produces better adaptive results when correlating the initial user preferences with the predicted preferences. The statistical analysis consisted of analyzing the RMSE and the Spearman correlation from data in Tables 4, 5, and 6.

Figure 4 shows the RMSE obtained for the three conditions evaluated in this work. As we initially hypothesized, the condition combining explicit and implicit feedback (C3) yields the lowest RMSE (0.772), which indicates that this alternative produces better predictions of the initial user preferences. Then, the condition using only explicit feedback (C1) reports a RMSE of 1.355 points, overcoming the use of implicit feedback (C2), which reports a RMSE of 1.662 points.

Moving deeper into our analysis, we also explored for which category (photos, videos, or music) the RMSE was lower. As Fig. 5 shows, the system produces the best results for the activities showing photos, followed by videos and music. Considering explicit feedback C1, the photos category reports a RMSE of 0.86 points, videos a value of 1.32 points, and music a value of 1.63 points. Regarding implicit feedback C2, photos obtains a score of 1.28, videos 1.64, and music 1.91. Finally, combined feedback C3 obtains the best RMSE value with 0.64 points for photos, 0.90 for videos, and 1.03 for music.

From this analysis, it is possible to perceive two interesting results. First, it is possible to observe the same tendency for all the categories. Independently of the type of activity, C3 produces lower RMSE than C1 and C2. This result supports our hypothesis about the benefits of combining explicit and implicit feedback rather than using them separately. Second, it seems the system fits the preferences better for shorter activities, producing better scores for photos (40 s) than videos and music (duration around 3 min) since the data used to train the RL algorithm can be obtained more often improving the learning speed.

From the previous outcomes, we wanted to analyze whether the participants' mean duration of the activities in each condition affects the RMSE value and, therefore, the learned preferences. We conducted a second statistical analysis using the Spearman correlation metric to find similarities between the mean duration of the activities and the RMSE of each user participating in the three conditions. The correlation results obtained were 0.23 points for Condition 1 (implicit feedback), 0.28 for Condition 2 (explicit feedback), and (0.16) for Condition 3 (combined feedback). These values are considered low Spearman correlations, so the activities' duration in this study does not affect the RMSE values.

Considering the previous RMSE scores, we can conclude that when computing the RMSE for all activities, the RMSE value when combining explicit and implicit feedback (C3) indicates that the model produces good learning. However, these results are not positive when using explicit (C1) or implicit (C2) feedback, as the RMSE values are high. The analysis of the RMSE values considering each condition and each category (photos, videos, and music) reports significant differences. For example, the RMSE value is positive for the combined feedback condition (C3) for the photos and videos categories but not that positive for music. Considering conditions C1 and C2, only the learning values for the photos category using explicit feedback can be considered positive. The other cases are not positive since RMSE values are above 1 unit. The comparison of the 3 conditions evaluated using the RMSE value shows that the learning values produced when combining explicit and implicit feedback are much better than when individually considering explicit or implicit feedback.

After analyzing the RMSE, we used the Spearman correlation metric for non-normal distributions to determine the correlation between the initial user ratings obtained from the online survey and the predicted Q-values. Table 3 shows the correlation metrics obtained for each condition considering all the activities and sorting them by categories. As mentioned in this manuscript, stronger correlations are represented by Spearman coefficients close to 1, as the study [31] states. This correlation can only be interpreted if the analysis reports statistical significance ($p - value > 0.05$).

The results we obtained regarding the Spearman correlation for all activities separating the three conditions under evaluation show that the observed and predicted values are strongly correlated (0.811) when combining explicit and implicit feedback (C3). Similarly, using only explicit feedback (C1) also reported a moderate correlation (0.567), which suggests that the user's ratings should be included in the loop. However, as occurred with the RMSE, the implicit feedback alone (C2) did not report a significant correlation, and therefore, the Spearman value is not worth interpreting.

Table 3 Spearman correlation values when comparing the initial user preferences and the model predictions

Condition	C1	C2	C3
All activities	0.567**	0.178	0.811**
Photos	0.826**	0.462**	0.850**
Videos	0.542**	0.123	0.779**
Music	0.594**	-0.028	0.817**

The analysis was conducted considering all of the activities and by categories. All significant statistical differences in this study are on level $p - value \leq 0.01$ (bilateral), indicated with $**$

The analysis by categories supports the outcomes produced by computing the RMSE. As shown in Table 3, combining explicit and implicit feedback (C3) produces a strong correlation for all categories, with values of 0.850 for photos, 0.779 for videos, and 0.817 for music. Similarly, the use of explicit feedback (C1) leads to positive outcomes, with a moderate Spearman correlation for videos (0.594) and music (0.542) and a strong correlation for photos (0.826). However, the initial and predicted preferences, when only considering implicit feedback (C2), reports a moderate correlation for photos with 0.462 points but not for videos (0.123) and music (-0.028).

5.2 Engagement and Number of Updates

The last statistical analysis we conducted aimed to assess whether how feedback was obtained affected user engagement and the number of times each activity was executed. We used the results in Table 7 for this analysis.

We carried out the one-way ANOVA test for the engagement analysis because the data was normally distributed. In this case, the one-way ANOVA did not report a significant statistical difference between the groups of the condition $F(2, 21) = .119$, $p - value = .889$.

We then conducted the Kruskal–Wallis non-parametric test to determine if the number of times each activity was performed was influenced by how the robot obtained user feedback. The Kruskal-Wallis test was conducted because the data was not normally distributed in this case. As in the previous case, the results did not provide any significant statistical difference ($p - value > 0.05$). More specifically, the Kruskal–Wallis test reported $H(2) = 2.949$, $p - value = .229$ for photos activities, $H(2) = 2.573$, $p - value = .276$ for video activities, and $H(2) = 3.822$, $p - value = .148$ for music activities. From the previous results discussed in the following section, we cannot assure that the type of feedback used to update the user preferences affects the number of updates or the user engagement. This is because, during the experiments, the number of times each user executes the activities changes, and this can not be accurately evaluated.

6 Discussion

Generating adaptive behavior in HRI is a process that continuously looks for user preferences. Most RL algorithms seek to optimize a problem finding to attain a specific goal. However, in this application, the learning system's goal is not to find a solution but to learn the users' preferences towards a group of activities. This fact implies that Q-values will not converge to a final value unless the rating is repeated for updating the reward function. If the user's rating is always the same, the learning system converges to the same value that the user set as their initial preferences. It is worth mentioning here that if a different updating value is obtained, then the adaptive process will modify the previous value to correct the error and cancel the convergence process.

Our experiment required each participant to interact with the robot for at least 100 min in four weeks to obtain accurate learning dynamics. This experiment differs from most of the HRI studies found in the literature, which typically design short-term experiments with only one or two activity sessions. However, we are aware that more participants and experiments are required to precisely define the influence of user feedback on HRI and which factors play a role in this process, especially on implicit feedback, due to the numerous metrics involved. This study serves as an initial step to evaluate the impact of explicit and implicit user feedback on preference learning and adaptive behavior in long-term scenarios.

The experiment results show that users were free to interact with the robot, so the amount of time spent with Mini differed across users. However, they all carried out more sessions than expected, which may indicate their will to interact with Mini in this task. Our results also show that combining explicit and implicit feedback provides the most similar learning values to the users' initial preferences, highlighting this approach's potential benefit. These results suggest that designing efficient methods to obtain implicit feedback is important to HRI and supports the information explicitly provided by the users.

The results suggest that implicit feedback alone is not a good alternative because the correlation is not significant for the categories of videos and music and is weak for photos. A possible reason for the worse results produced by only considering implicit feedback might reside in the interaction metrics we selected to compute it. Unlike explicit feedback, which obtains the real user preferences from their ratings, implicit feedback in our model is computed from the interaction time, activity result, and user selection. However, other metrics like for example the difficulty of the activity, the user experience or knowledge level, the number of times the activity has been repeated or the errors that might appear during the execution due to the activity programming/design.

The statistical analysis conducted on the data indicates that the kind of feedback used to retrieve the user preferences does not affect user engagement since the interaction time with the robot does not decay with time. This analysis also reports that the number of updates of each activity is not affected by how the robot obtains user feedback. This suggests that more invasive methods, like continuously asking the user to rate the activity, are not perceived as negative. The statistical analyses conducted in this study ignore the relevance per subject and user since we treat all activities and users equally. We are aware that possibly there are differences in the activities and subjects, but these differences are subjective and can not be easily analyzed from the current data.

Based on the previous discussion and the three hypotheses we enumerated at the beginning of the manuscript, it is possible to provide the following statements. The first hypothesis (H1) is partially accepted. The results show that condition C3 yields better adaptive results than C1 and C2, but no significant differences could be obtained for better user engagement and activity exploration. The second hypothesis (H2) can be validated since the results prove that combining implicit and explicit feedback in C3 is the best way to learn user preferences to personalize HRI. Finally, the third hypothesis (H3) is partially accepted. C1 provides good learning results since it uses real user feedback but does not impact activity execution.

From the previous results, we also statistically analyzed if gender, age, and the other demographic factors obtained from the demographic survey impacted the results of our experiment. However, we did not observe any effects on these factors. For the specific case of gender, we think that the kind of activity might impact the feedback provided. However, the low number of women in the sample makes obtaining results in line with this hypothesis difficult. A similar issue occurs with age. Since the recruited participants are all university students of similar ages, we do not have clear indications that this factor influences how the robot obtains user feedback to update the activity preferences.

6.1 Limitations

Design factors and the tasks chosen affect the learning system described in this contribution. Next, we enumerate the limitations of this work:

1. In tabular RL methods, such as the one used in this work, the number of actions greatly impacts learning speed and tractability. This means the learning process is slower as the number of actions increases. Thus, we propose to use function approximation methods to speed up the learning process in further tests. In our approach, the method used will not affect final values because feedback will not change.

2. Adaptation rate is controlled by α, which is a practical value that regulates how fast the error is corrected. High alpha values can drive the system not to fit the Q-value properly, while shallow values affect the speed of the adaptive mechanisms. This work sets the α parameter to 0.5, which limits the error correction per time step. Although this and other values modify the learning performance, the designer can tune them to make the system work as they expect.
3. Activity exploration balances the update process of all of the actions equally. In this study, we apply Boltzmann's distribution using a low Temperature value of 5 units to explore all activities at the beginning of the experiment but promote selecting preferred activities as the experiment progresses. This method contains randomness in the activity selection since it is based on probabilities generated from the user preferences (Q-values). Consequently, as mentioned earlier, the randomness in action selection and user preferences may lead to unexplored activities. This issue might subtly affect the Spearman metric and the RMSE.
4. The reward function design is the key to learning correctly. In our method, the interaction metrics define the implicit feedback reward. We know that the interaction metrics may affect the reward value; therefore, we expect to explore the impact of the interaction metrics used in future studies. Besides, in this scenario, we give more importance to a couple of metrics (user selection and activity result) since we believe they are more important than the execution time. However, the designer can change these weights or the interaction metrics depending on their application.

7 Conclusion

The results of this study show that combining the users' explicit and implicit feedback to learn the users' preferences toward a group of activities to personalize the interaction improves the results of individually using explicit or implicit feedback. This fact indicates that the robot can obtain more information from the interaction to improve HRI. The results also show that the kind of feedback does not affect activity exploration and user engagement, suggesting that including explicit questions does not influence the execution of activities. From these results, it is possible to state that only hypothesis 2 (H2) can be entirely accepted. The other hypotheses (H1 and H3) can only be partially accepted since not all the assumptions are confirmed based on the data analyses.

Considering the positive results provided by this study, we would like to investigate these methods to produce adaptive sessions in assistive robotics, extend the activity repertoire of the robot, and explore the role of user feedback in other areas and applications. We would also like to combine our system with a preference predictor that we developed in a previous work [17] to enable the robot to start the personalization of activities from a user-oriented prediction and not from scratch to evaluate which factors influence user feedback in HRI. Finally, exploring Continuous and Active Learning methods that might be a good alternative to solving preference learning from user feedback would be interesting.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. The research leading to these results has received funding from the grants PID2021-123941OA-I00, funded by MCIN/AEI/10.13039/501100011033 and by "ERDF A way of making Europe"; TED2021-132079B-I00 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR; Mejora del nivel de madurez tecnologica del robot Mini (MeNiR) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR; Portable Social Robot with high level of Engagement (PoSoRo) PID2022-140345OB-I00 funded by MCIN/AEI/10.13039/501100011033 an by "ERDF A way of making Europe". Universidad Carlos III de Madrid (Agreement CRUE-Madroño 2024).

Data availability The data used for our study is in the Appendix. No additional data further than these were used

Declarations

Conflict of interests The authors declare that they have no conflict of interest

Human Participants Disclosure The University Carlos III of Madrid approved the Human–Robot Interaction experiment, and all participants gave their consent and approved their participation

Appendix

See Tables 4, 5, 6 and 7.

Table 4 Data collected from the experiment regarding the initially indicated and predicted Q—values for the activities related to the Photos category

Id	Cond	Photos Animals Real	Photos Animals Pred	Photos Landscapes Real	Photos Landscapes Pred	Photos Monuments Real	Photos Monuments Pred	Photos Sad Real	Photos Sad Pred
1	1	3	4,04	4	3,56	2	3,48	1	1,4
2	1	4	3,66	4	3,96	4	3,75	1	1,23
3	1	4	3	4	3,31	3	3,47	1	1,17
10	1	5	4,45	4	4,35	4	3,71	1	1,6
11	1	5	5	5	4,65	5	4,81	4	3,87
17	1	4	4,28	5	4,91	5	4,9	1	4,33
19	1	5	4,64	5	4,59	4	4,5	2	1,6
24	1	4	4,41	5	4,47	3	4,31	1	2,64
4	2	5	4,29	5	4,45	3	3,15	2	1,59
7	2	5	4,19	5	4,55	3	3,85	1	1,71
8	2	4	4,25	5	3,76	3	3,61	2	1,27
12	2	4	4,67	4	5	2	4,16	1	5
13	2	5	4,92	5	4,81	4	4,33	1	2,11
18	2	5	4,51	4	4,85	2	3,8	1	1,24
20	2	3	5	5	4,88	3	5	2	2,76
23	2	3	5	4	5	4	5	2	3,31
5	3	4	4,05	5	4,13	3	3,95	1	1,27
6	3	4	3,97	5	4,69	4	3,49	1	1,33
9	3	4	3,98	4	4,23	3	3,7	1	0,97
14	3	4	4,67	4	4,77	4	4,84	2	2,54
15	3	5	4,78	5	5	4	4,75	2	1,76
16	3	4	4,08	5	5	3	2,87	2	4,6
21	3	5	5	4	4,9	4	4,05	2	2,06
22	3	4	4,01	3	3,21	3	3,14	1	1,09

Table 5 Data collected from the experiment regarding the initially indicated and predicted Q—values for the activities related to the Videos category

Id	C	Videos Cooking Real	Videos Cooking Pred	Videos Film Trailers Real	Videos Film Trailers Pred	Videos Funny Real	Videos Funny Pred	Videos Comedy Real	Videos Comedy Pred	Videos Sport Real	Videos Sport Pred
1	1	4	3,77	4	3,61	5	4,2	3	4,25	1	2,9
2	1	4	3,98	3	3,46	3	3,57	1	3,63	2	3,47
3	1	3	3,82	4	4,52	5	3,91	3	3,93	4	4,39
10	1	1	3,6	2	3,8	4	4,81	5	5	5	4,53
11	1	3	4,54	5	4,74	5	4,62	3	4,56	4	5
17	1	2	4,7	2	5	5	4,9	5	5	5	4,9
19	1	4	4,59	4	4,65	3	4,7	1	1,56	2	4,38
24	1	3	3,45	5	4,91	3	4,55	4	4,61	2	4,37
4	2	5	4,37	4	4,06	5	3,93	3	3,35	2	1,15
7	2	5	3,68	3	2,37	5	4,54	2	3,23	1	2,39
8	2	4	3,63	4	4,05	4	4,07	4	3,99	1	1,39
12	2	2	4,73	3	4,28	3	4,73	1	3,11	1	4,8
13	2	4	4,34	3	2,45	5	4,86	5	5	3	3,89
18	2	2	4,7	4	4,18	5	4,33	2	4,9	1	4,8
20	2	2	3,89	1	4,62	4	4,38	1	4,9	3	4,62
23	2	4	4,62	5	4,72	5	4,43	5	4,48	5	4,4
5	3	4	4,14	5	4,44	3	3,06	3	4,07	3	3,05
6	3	4	4,18	3	4,1	3	3,95	3	3,84	2	2,86
9	3	3	4,03	2	2,81	5	4,88	2	3,12	5	4,78
14	3	3	2,9	3	4,57	4	4,81	5	4,48	4	4,45
15	3	4	4,73	4	4,46	2	2,57	1	1,48	1	2,42
16	3	1	1,54	5	4,77	3	4,52	2	4,9	3	3,28
21	3	4	4,25	4	4,73	4	4,73	4	4,73	1	2,8
22	3	4	4,16	3	2,82	5	4,63	3	2,85	1	2,21

Table 6 Data collected from the experiment regarding the initially indicated and predicted Q—values for the activities related to the Music category

Id	C	Music Spanish pop Real	Music Spanish pop Pred	Music English pop Real	Music English pop Pred	Music Latin pop Real	Music Latin pop Pred	Music Spanish rock Real	Music Spanish rock Pred	Music English rock Real	Music English rock Pred	Music Noise Real	Music Noise Pred
1	1	3	4,34	4	4,79	2	3,99	4	4,56	4	4,56	1	2,69
2	1	4	3,95	4	3,79	4	3,81	1	4,32	1	3,05	1	1,9
3	1	3	2,81	4	4,06	5	4,96	3	3,94	3	4,12	1	3,15
10	1	5	5	4	4,51	4	4,75	2	4,91	2	3,5	1	4,6
11	1	2	3,97	3	4,81	4	5	4	5	4	5	2	5
17	1	5	5	5	5	4	5	4	5	4	5	1	4,6
19	1	2	4,29	2	4,68	1	4,76	5	4,65	5	4,81	2	2,8
24	1	4	4,74	3	4,72	2	2,94	5	5	4	4,81	1	2,43
4	2	4	3,58	4	4,41	2	2,08	4	3,86	5	3,4	1	4,25
7	2	5	3,89	5	3,61	5	3,43	4	3,59	3	3,65	1	1,26
8	2	4	3,09	4	4,41	2	3,6	5	2,19	5	4,31	1	3,69
12	2	2	4,81	4	4,73	5	4,69	1	4,9	1	4,9	1	4,62
13	2	2	2,98	2	1,72	4	4,24	3	3,53	1	2,33	1	1,6
18	2	3	4,16	3	4,72	2	4,81	2	4,41	2	4,7	1	4,61
20	2	3	5	4	4,59	3	4,9	4	4,81	4	4,81	1	1,97
23	2	4	4,59	2	4,49	3	5	4	4,55	2	4,72	1	4,62
5	3	4	3,78	4	2,83	3	3,98	3	2,5	3	3,42	1	0,91
6	3	3	3,41	3	3,78	2	1,16	4	4,11	4	4,55	1	0,87
9	3	4	3,55	5	4,27	5	4,93	4	3,72	4	4,46	1	1,23
14	3	4	4,81	4	4,62	4	4,62	3	3,35	3	4,62	1	2,23
15	3	3	4,9	4	4,59	1	1,71	1	2,65	1	1,9	1	1,71
16	3	4	4,23	4	4,91	4	4,81	4	4,15	4	3,56	2	1,76
21	3	4	4,08	3	3,51	3	4,1	5	4,55	5	4,73	1	1,12
22	3	5	4,9	3	3,76	4	4,37	2	1,96	1	2,19	1	1,07

Table 7 Data collected from the experiment regarding the interaction time of each participant and the number of times that each activity was executed

Id	C	Time (min)	Photos Animals N	Photos Landscapes N	Photos Monuments N	Photos Sad N	Videos Cooking N	Videos Film Trailers N	Videos Funny N	Videos Comedy N	Videos Sport N	Music Spanish pop N	Music English pop N	Music Latin N	Music Spanish rock N	Music English rock N	Music Noise N
1	1	247	13	8	7	8	8	8	14	4	5	6	14	5	2	11	3
2	1	315	20	21	13	10	20	9	9	5	5	28	12	21	4	4	6
3	1	206	14	8	5	11	7	14	10	16	12	6	15	7	2	6	4
10	1	117	5	6	5	4	4	3	2	5	7	4	4	3	2	5	1
11	1	132	6	4	4	4	3	4	2	3	3	4	2	0	1	1	0
17	1	124	6	5	4	2	3	4	5	1	5	4	4	3	2	3	1
19	1	138	3	6	4	4	5	4	1	7	3	3	3	2	4	2	6
24	1	126	5	7	4	5	4	4	7	3	3	3	2	3	2	2	4
4	2	278	26	25	9	8	21	17	16	7	23	15	15	8	10	26	2
7	2	230	22	15	8	6	13	6	9	8	7	11	8	10	15	11	7
8	2	206	14	20	10	9	10	13	16	16	6	6	3	1	12	8	2
12	2	130	7	4	5	0	3	13	3	6	1	2	3	5	1	1	2
13	2	96	4	6	5	4	7	9	6	6	3	5	8	4	8	5	6
18	2	135	4	4	3	7	2	5	4	1	3	13	2	2	4	1	2
20	2	123	5	7	5	3	3	2	9	1	2	0	5	1	2	2	7
23	2	117	4	2	2	3	2	2	5	4	8	5	4	1	3	2	2
5	3	214	22	24	12	12	29	30	20	14	17	11	22	10	13	17	13
6	3	221	19	19	30	21	24	26	21	12	8	11	9	11	15	44	12
9	3	189	8	12	6	7	11	7	13	5	14	7	10	9	6	7	7
14	3	144	12	7	7	6	6	4	14	10	9	2	2	6	6	2	4
15	3	119	5	7	6	6	3	6	4	7	5	1	6	7	5	6	5
16	3	112	9	7	8	1	6	6	3	1	6	4	2	4	8	6	9
21	3	125	8	6	9	6	4	3	3	3	7	9	4	3	7	4	8
22	3	167	8	8	8	7	5	9	12	11	4	14	8	8	10	6	7

References

1. Akalin N, Kristoffersson A, Loutfi A (2019) The influence of feedback type in robot-assisted training. Multimodal Technol Interact 3(4):67
2. Akalin N, Loutfi A (2021) Reinforcement learning approaches in social robotics. Sensors 21(4):1292
3. Asprino L, Ciancarini P, Nuzzolese AG, Presutti V, Russo A (2022) A reference architecture for social robots. J Web Semant 72:100683
4. Baraka K, Veloso M (2015) Adaptive interaction of persistent robots to user temporal preferences. In: International conference on social robotics. Springer, pp 61–71
5. Boggess K, Chen S, Feng L, (2020) Towards personalized explanation of robot path planning via user feedback. arXiv:2011.00524
6. Caleb-Solly P, Dogramadzi S, Huijnen CA, Heuvel HVD (2018) Exploiting ability for human adaptation to facilitate improved human-robot interaction and acceptance. Inf Soc 34(3):153–165
7. Ceha J, Law E, Kulić D, Oudeyer P-Y, Roy D (2022) Identifying functions and behaviours of social robots for in-class learning activities: Teachers' perspective. Int J Soc Robot 14(3):747–761
8. Che Y, Okamura AM, Sadigh D (2020) Efficient and trustworthy social navigation via explicit and implicit robot–human communication. IEEE Trans Robot 36(3):692–707
9. Cross ES, Hortensius R, Wykowska A (2019) From social brains to social robots: applying neurocognitive insights to human–robot interaction
10. Cruz F, Wüppen P, Fazrie A, Weber C, Wermter S (2018) Action selection methods in a robotic reinforcement learning scenario. In: 2018 IEEE Latin American conference on computational intelligence (LA-CCI). IEEE, pp 1–6
11. Fernández-Rodicio E, Castro-González Á, Alonso-Martín F, Maroto-Gómez M, Salichs MÁ (2020) Modelling multimodal dialogues for social robots using communicative acts. Sensors 20(12):3440
12. Fox J, Gambino A (2021) Relationship development with humanoid social robots: applying interpersonal theories to human–robot interaction. Cyberpsychol Behav Soc Network 24(5):294–299
13. Haas Md, Baxter P, deJong C, Krahmer E, Vogt P (2017) Exploring different types of feedback in preschooler and robot interaction. In: Proceedings of the companion of the 2017 ACM/IEEE international conference on human–robot interaction, pp 127–128
14. Hemminahaus J, Kopp S (2017) Towards adaptive social behavior generation for assistive robots using reinforcement learning. In: 2017 12th ACM/IEEE international conference on human–robot interaction (HRI). IEEE, pp 332–340
15. Holtz J, Biswas J (2022) Socialgym: a framework for benchmarking social robot navigation. In: 2022 IEEE/RSJ international conference on intelligent robots and systems (IROS). IEEE, pp 11246–11252
16. Maroto-Gómez M, Castro-González Á, Castillo JC, Malfaz M, Salichs MA (2018) A bio-inspired motivational decision making system for social robots based on the perception of the user. Sensors 18(8):2691
17. Maroto-Gómez M, Castro-González Á, Castillo JC, Malfaz M, Salichs MÁ (2022) An adaptive decision-making system supported on user preference predictions for human-robot interactive communication. User Model User-Adapted Interact 33(2):359–403
18. Maroto-Gómez M, Castro-González Á, Malfaz M, Salichs MÁ (2023) A biologically inspired decision-making system for the autonomous adaptive behavior of social robots. Complex Intell Syst 9(6):6661–6679
19. Maroto-Gómez M, Malfaz M, Castro-González Á, Salichs MÁ (2023) A motivational model based on artificial biological functions for the intelligent decision-making of social robots. Memetic Comput 15(2):237–257
20. Maroto-Gómez M, Villarroya SM, Malfaz M, Castro-González Á, Castillo JC, Salichs MÁ (2022) A preference learning system for the autonomous selection and personalization of entertainment activities during human–robot interaction. In: 2022 IEEE international conference on development and learning (ICDL). IEEE, pp 343–348
21. Moro C, Nejat G, Mihailidis A (2018) Learning and personalizing socially assistive robot behaviors to aid with activities of daily living. ACM Trans Hum Robot Interact (THRI) 7(2):1–25
22. Nasir J, Bruno B, Chetouani M, Dillenbourg P (2022) What if social robots look for productive engagement? Int J Soc Robot 14(1):55–71
23. Olatunji S, Oron-Gilad T, Sarne-Fleischmann V, Edan Y (2020) User-centered feedback design in person-following robots for older adults. Paladyn J Behav Robot 11(1):86–103
24. Park HW, Grover I, Spaulding S, Gomez L, Breazeal C (2019) A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. Proc AAAI Conf Artif Intell 33:687–694
25. Ritschel H, André E (2017) Real-time robot personality adaptation based on reinforcement learning and social signals. In: Proceedings of the companion of the 2017 ACM/IEEE international conference on human–robot interaction. pp 265–266
26. Ritschel H, Baur T, André E (2017) Adapting a robot's linguistic style based on socially-aware reinforcement learning. In: 2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE, pp 378–384
27. Ritschel H, Seiderer A, Janowski K, Wagner S, André E (2019) Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback. In: Proceedings of the 12th ACM international conference on PErvasive technologies related to assistive environments, pp 247–255
28. Salhi I, Qbadou M, Gouraguine S, Mansouri K, Lytridis C, Kaburlasos V (2022) Towards robot-assisted therapy for children with autism—the ontological knowledge models and reinforcement learning-based algorithms. Front Robot AI 9:713964
29. Salichs MA, Castro-González Á, Salichs E, Fernández-Rodicio E, Maroto-Gómez M, Gamboa-Montero JJ, Marques-Villarroya S, Castillo JC, Alonso-Martín F, Malfaz M (2020) Mini: a new social robot for the elderly. Int J Soc Robot 12:1231–1249
30. Schneider S, Kummert F (2017) Exploring embodiment and dueling bandit learning for preference adaptation in human–robot interaction. In: 2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN). IEEE, pp 1325–1331
31. Schober P, Boer C, Schwarte LA (2018) Correlation coefficients: appropriate use and interpretation. Anesth Analg 126(5):1763–1768
32. Sheridan TB (2016) Human-robot interaction: status and challenges. Hum Factors 58(4):525–532
33. Shi Z, Groechel TR, Jain S, Chima K, Rudovic O, Matarić MJ (2022) Toward personalized affect-aware socially assistive robot tutors for long-term interventions with children with autism. ACM Trans Hum Robot Interact (THRI) 11(4):1–28
34. Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. MIT Press, Cambridge
35. Tsiakas K, Abujelala M, Makedon F (2018) Task engagement as personalization feedback for socially-assistive robots and cognitive training. Technologies 6(2):49
36. Van Otterlo M, Wiering M (2012) Reinforcement learning and Markov decision processes. In: Reinforcement learning. Springer, pp 3–42
37. Wakayama S, Ahmed N (2023) Active inference for autonomous decision-making with contextual multi-armed bandits. In: 2023

IEEE international conference on robotics and automation (ICRA). IEEE, pp 7916–7922

38. Wang N, Di Nuovo A, Cangelosi A, Jones R (2019) Temporal patterns in multi-modal social interaction between elderly users and service robot. Interact Stud 20(1):4–24
39. Whitney D, Rosen E, MacGlashan J, Wong LL, Tellex S (2017) Reducing errors in object-fetching interactions through social feedback. In: 2017 IEEE international conference on robotics and automation (ICRA). IEEE, pp 1006–1013
40. Wirth C, Akrour R, Neumann G, Fürnkranz J et al (2017) A survey of preference-based reinforcement learning methods. J Mach Learn Res 18(136):1–46

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Marcos Maroto-Gómez is an Assistant Professor at Carlos III of Madrid University. He obtained his B.Sc. degree in Industrial Electronics and Automation Engineering from the University of Castilla-La Mancha, Toledo, Spain, in 2015 and the M.Sc. and Ph.D. in Robotics and Automation from the Carlos III University of Madrid, Madrid, Spain, in 2022. He is member of the Robotics Lab Research Group nd works on human-robot interaction, decision-making, adaptation, autonomy, and machine learning applied to social robots.

María Malfaz is an Associate Professor at the Carlos III University of Madrid. María Malfaz received her degree in Physics Science at La Laguna University in 1999. In October 2001, she became a M.Sc. in Control Systems at Imperial College of London. She received the Ph.D. degree in Industrial Engineering in 2007 and the topic was "Decision Making System for Autonomous Social Agents Based on Emotions and Self-learning". Her research area follows the line carried out in her thesis and, more recently, she has been working on multimodal human-robot interaction systems. She belongs to several international scientific associations: IEEE-RAS (IEEE Robotics and Automation Society), IFAC (International Association of Automatic Control), and CEA (Comité Espanol de Automática).

José Carlos Castillo holds an M.Sc. degree in Advanced Computer Technologies (2008) and a PhD degree in Computer Science (2012) from Castilla-La Mancha University, Spain. From 2006 to 2012, he worked at the natural and artificial Interaction Systems (n&aIS) group at the Albacete Research Institute of Informatics, Spain, focusing on computer vision techniques to detect human activities and frameworks for intelligent monitoring and activity interpretation. From October 2012 to September 2013, he worked as a post-doctoral researcher at the Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST) of the Technical University of Lisbon (UTL), where he focused on networked robot systems, robotics and computer vision and intelligent control systems. In September 2013 he moved to the RoboticsLab of the Universidad Carlos III de Madrid, where he is an Associate Professor working on social robotics and computer vision techniques for Human- Robot Interaction.

Álvaro Castro-González received the B.Sc. degree in Computer Engineering from the University of León, León, Spain, in 2005, and the M.Sc. and Ph.D. degrees in Robotics and Automation from the Carlos III University of Madrid, Madrid, Spain, in 2008 and 2012, respectively. He is a Member of the RoboticsLab Research Group and Associate Professor at the Department of Systems Engineering and Automation of the Carlos III University of Madrid, Madrid, Spain. He has been involved in several national, European, and corporate sponsored research projects. His research lines are human-robot interaction, social robots, expressiveness in robots, decision making, and artificial emotions.

Miguel Ángel Salichs received the degree in electrical engineering and the Ph.D. degree from the Polytechnic University of Madrid. He is Full Professor at the Systems Engineering and Automation Department, Carlos III University of Madrid. His research interests include autonomous social robots, multi-modal human-robot interaction, mind models, and cognitive architectures. He was a member of the Policy Committee of the International Federation of Automatic Control (IFAC), the Chairman of the Technical Committee on Intelligent Autonomous Vehicles of IFAC, a responsible of the Spanish National Research Program on Industrial Design, a Production Member of the Spanish Society on Automation and Control (CEA), and the Spanish Representative with the European Robotics Research Network (EURON).