Dynamics of Mental Models: Objective vs. Subjective User Understanding of a Robot in the Wild

Ferran Gebellí[1], Anaís Garell[2], Séverin Lemaignan[1] and Raquel Ros[3]

***Abstract*—In Human-Robot Interaction research, assessing how humans understand the robots they interact with is crucial, particularly when studying the impact of explainability and transparency. Some studies evaluate *objective understanding* by analysing the accuracy of users' mental models, while others rely on perceived, self-reported levels of *subjective understanding*. We hypothesise that both dimensions of understanding may diverge, thus being complementary methods to assess the effects of explainability on users. In our study, we track the weekly progression of the users' understanding of an autonomous robot operating in a healthcare centre over five weeks. Our results reveal a notable mismatch between objective and subjective understanding. In areas where participants lacked sufficient information, the perception of understanding, i.e. subjective understanding, raised with increased contact with the system while their actual understanding, objective understanding, did not. We attribute these results to inaccurate mental models that persist due to limited feedback from the system. Future research should clarify how both objective and subjective dimensions of understanding can be influenced by explainability measures, and how these two dimensions of understanding affect other desiderata such as trust or usability.**

***Index Terms*—Social HRI, Long term Interaction, Human-Centered Robotics**

I. Introduction

ONE of the main goals of designing explainable robots is to improve the understanding of users about the robots' decisions and behaviours. In turn, this will contribute to achieving other desiderata, such as raising users' satisfaction, usability or trust when interacting with these robots [1], [2].

In the Human-Robot Interaction (HRI) field, the Theory of Mind (ToM) approach assumes that users build an internal Mental Model (MM) about the robot, which helps them to predict the robot's decisions and behaviour [3]. The evaluation of the user's understanding of the robot is often done by analysing those mental models [4]. The literature review on human-centred eXplainable AI (XAI) [5] identifies two types of understanding that can be evaluated. *Objective understanding* is the actual comprehension of the system, usually measured as the accuracy of the user's mental model of the system in a proxy task. *Subjective understanding* is the user-perceived and self-rated level of understanding, considered as the confidence that users have about their objective understanding, and is usually measured through questionnaires. Most of the previous XAI works assess either objective or subjective understanding, but not both [5], sometimes using subjective understanding as a replacement for objective understanding [2]. However, previous literature has not yet explored the relationship between the two types of understanding in HRI, which is a gap that we address in this work. We believe that both metrics should be analysed separately, since other desiderata, such as trust, usability or performance, might be affected differently by each type of understanding.

Fig. 1. Evaluation of understanding with part of the nursing staff right after the tutorial session and before any usage of the system.

Analysing the evolution through time of objective and subjective understanding requires multiple engagements over extended periods with the same users [6]. However, the evaluation of understanding is typically conducted after very short-term interactions with the robot, which often overlooks the novelty effect adequately [7]. Those study settings are not realistic, being often in-lab experiments [8]. Although the dynamics of mental models over time have been studied in XAI, e.g. with recommender systems [9], up to the authors' knowledge, long-term studies of user understanding in the wild have not been addressed in HRI.

In this work, we conduct a user study (Fig. 1) where a robot is deployed in the wild for 5 weeks in the geriatric unit of an intermediate care centre. The robot assists the nursing staff in identifying potentially hazardous situations for patients. We address two research questions:

Manuscript received: January, 28, 2025; Revised March, 30, 2025; Accepted May, 30, 2025. This paper was recommended for publication by Editor Ki-Uk Kyung upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Horizon Europe grant N. 101072488 (TRAIL).

[1] Ferran Gebellí and Séverin Lemaignan belong to PAL robotics (Barcelona, Spain) ferran.gebelli@pal-robotics.com, severin.lemaignan@pal-robotics.com

[2]Anaís Garrell belongs to the Institut de Robòtica i Informàtica Industrial (CSIC-UPC) (Barcelona, Spain) anais.garrell@upc.edu

[2]Raquel Ros belongs to the Institut d'Investigació en Intel·ligència Artificial (CSIC), (Bellaterra, Spain) raquel.ros@iiia.csic.es

Digital Object Identifier (DOI): see top of this page.

- **RQ1**: What is the relationship between *objective* and *subjective* understanding?
- **RQ2**: How does the user's understanding of an autonomous robot in the wild evolve with increased exposure and interaction?

This work is organised as follows. Sec. 2 reviews the related work. Sec. 3 details the study design, and Sec. 4 presents the results. A discussion of the outcomes is presented in Sec. 5. Finally, Sec. 6 concludes the paper.

II. Related Work

In this section, we will first review why understanding is a key aspect of measuring the effects of explainability. Then, we survey objective and subjective understanding assessments, and the evaluation of understanding's dynamics over time.

A. Explainability and understanding

The review work on XAI metrics [4] proposes three stages in XAI evaluation: (1) an assessment of the explanation goodness and user satisfaction, (2) a test of comprehension (or understanding) which measures the mental model and (3) the measuremnt of other metrics like performance or trust which are affected by understanding. Later, the work in [1] further generalises this process, stating that the primary goal of explainability is facilitating understanding, which in turn will affect other desiderata such as trust, performance, usability or satisfaction.

The review works in XAI [2], [5], and in eXplainable HRI (XHRI) [10] also acknowledge the major role of understanding. In the next subsection, we explore how understanding has been measured, together with current limitations.

B. Objective understanding

Mental models, which have their origins in psychology research, are internal representations that people use to understand, explain and predict the real world [11], [12]. Mental models have been used to model shared understanding between users [13], [14], to design intelligent systems [9], [15], and to design [3], [16] and evaluate [2], [4], [10] XHRI.

The obtained mental models are usually analysed by comparing them with the true decision-making and behaviour of the system [4], [10]. By assessing the accuracy of the users' mental models, objective understanding can be measured. A key aspect in evaluating objective understanding is choosing an appropriate proxy task. According to [17], the selected task should "maintain the essence of the target application". One of the most widely recognised proxy tasks is forward simulation [17], [18], which involves requiring participants to simulate or predict what the system would do.

C. Subjective understanding

In other cases, the perceived, subjective understanding, has been used as a way to estimate objective understanding. However, it has been argued that subjective understanding might not reflect objective understanding, since participants might have unjustified beliefs [2]. Research in social sciences evidences that people tend to have a wrong perception of their understanding [19]. Some user studies have supported the idea that the subjective level of understanding is initially high but gradually declines as time progresses [20], [21], while other works [9] report the opposite, i.e. that subjective understanding begins at a lower level and then increases over time.

D. Objective vs subjective understanding

Some works have measured both objective and subjective understanding, but they have not included its comparison as a major research question. For example, [22] measures them jointly with 7 other metrics, such as cognitive load, trust, or explanation preference. Despite the noteworthy differences between the two understanding metrics, such distinction is not properly addressed, and the metrics are merely used as complementary ways to measure the significance under test conditions.

This pattern is present in many other works [23], [9], [24], [25]. Only [25] presents a short comparison of both understanding metrics, but it is solely a sentence stating that "participants tend to overestimate their understandability [...] the relationship between subjective and objective understandability is an interesting topic for future work".

Other works have compared objective and subjective metrics which are related to explainability, but do not explicitly measure understanding. For example, in terms of subjective metrics, in [26] the participants rate how satisfied users are with explanations, while [27] defines a set of questions to measure not only understandability but "explainability" as a whole, including the "simulatability, transparency, and usability" of the explanations. Then, [28] includes the same "explainability" subjective metric from [27] while introducing an explicit self-reported subjective understanding as a secondary outcome, which confirms that the subjective metric from [27] does not exclusively map to understanding. Nevertheless, results are mixed: in [27], there is a positive correlation between the two types of metrics, while in [28], the relation is, on average, negative, with different results across user groups with different expertise.

In this work, we explicitly focus on the relation between objective and subjective understanding while investigating its dynamics, as we review in the next subsection.

E. Dynamics of understanding over time

Some works have studied the progression of understanding over time for users of XAI non-embodied intelligent systems [9], [15], [29].

It has been argued that mental models are relatively persistent over time [9], [15], that is, people tend to adhere to their initial beliefs and explainability measures cannot always modify them. However, these works concede that the information provided was probably too short and basic, and that other types of explainability measures might lead to other results. The work in [9] reports that receiving extra information correlates with a higher understanding, but in [15], [30] this

claim does not hold. This contradiction further confirms that the type and nature of the provided explanations play a major role in the evolution of understanding with time. Moreover, these works do include the time dimension when measuring understanding, but do not explicitly consider the comparison of objective and subjective understanding.

Although some theoretical works treat the evolution of mental models, e.g. concerning robot anthropomorphism [31], up to the authors' knowledge, there is a lack of user studies in HRI that evaluate the dynamics of understanding (objective and subjective) in the long run. In this work, we aim to cover such a research gap.

III. Study design

A. Environment and task

The study took place in the rehabilitation unit of an intermediate healthcare facility, which primarily serves elderly patients undergoing medium- to long-term rehabilitation, typically lasting from weeks to months. The robot's functionality was defined through a 7-month co-design process with the staff [32]. The robot was tasked with patrolling patients' rooms and alerting the nursing staff in case of potential hazards (Fig. 2). These alerts included detecting a person lying on the floor (indicating a fall), a person standing alone in the room, or a room door being closed when it was expected to remain open.

The main users of the system are the nursing staff, who are in charge of configuring the robot patrolling routines and addressing the alerts raised by the system via a mobile app available on each of the staff member's phones. Moreover, the robot has a screen on the chest and LEDs on the base to display its status (e.g., patrolling, idle, low battery, etc.).

Users can schedule patrolling rounds indicating the rooms to monitor at specific times. They can also specify the rooms where the standing patient and closed-door alarms should be active. Users can always view, edit or delete the active and scheduled routines.

When the robot triggers an alert, all phones start vibrating and emitting a sound that depends on the alert's severity. The screen of the phone will show the location and type of alert, as well as an image of the scene taken by the robot's camera, as shown in Fig. 2 (right). At that point, any user can press a button to stop the alert on all other phones. The user who pressed the button should address the incident and specify if it was a true or false alert. Besides the alerts corresponding to risk situations for the patients, alerts related to robot failures are also present: when the robot gets lost, when its path is fully blocked, a motor overheats, or the emergency button has been pressed. When any of these alerts are triggered, the user receives instructions on how to fix the problem. When any of those alerts are active, an optional "explain me more" button will appear to receive additional explanations regarding the event. Explanations are fetched in a dictionary-based approach, indexed by the alert type. They are delivered on-demand for two reasons: on the one hand, to avoid overwhelming users with non-priority information, and on the other hand, to explicitly measure when an explanation has actually been requested.

Fig. 2. The robot autonomously returns to its charging station after completing a patrol round (left) and user interface displaying a fall alert actioned from the robot sensors (right).

Some examples of daily interactions between the users and the robot are: (a) the robot raises a standing person alarm, a nurse goes to that room to assist the patient; (b) the robot cannot move because it is obstructed, a staff member assists the robot by clearing the space around it; (c) a staff member enters a room and negotiates the shared space with the robot.

B. Apparatus

We used a TIAGo LITE robot from PAL Robotics, a mobile platform with a touch screen on the chest. It runs Ubuntu 20.04 LTS with ROS Noetic middleware on an Intel Core i7-10700 CPU @ 2.90GHz. It uses the ROS navigation stack to navigate autonomously using a LIDAR and RGBD camera. It is also equipped with a thermal camera, which, along with the RGBD camera, is used to identify the potential risk situations described above.

Each participant had a phone with a dedicated app to interact with the robot. The app automatically logs out the current user in every shift change, forcing the next shift to log in. By doing so, we can separately track the interactions of each user with the system through the app.

C. Participants

Participants were recruited from the nursing staff of the rehabilitation unit. A total of $N = 31$ participants took part in the study, comprising 27 females and 4 males, including 9 nurses and 22 nursing assistants. Of the participants in the study, only two had been actively involved in the participatory design process, while the rest were unfamiliar with the system prior to the study. The nursing staff operates in four shifts: morning, afternoon, and two alternating night shifts, with 7-8, 4-5, and 3-4 members working simultaneously during each respective shift. Over the course of the study, we recruited a total of 12 morning-shift, 5 afternoon-shift, and 14 night-shift staff members. All participants provided written consent for their voluntary participation after receiving a detailed briefing about the study.

D. Procedure

After participants had signed the consent form[1], they filled in the understanding questionnaire described in the next section, prior to any contact with the system. Next, we conducted

[1]The ethical committee from the hospital where the study was performed (BSA) allowed the study's execution (March 2024).

a hands-on tutorial session for the whole shift to explain the usage and main features of the system. The robot was present in the session and each participant experimented with the system. Right after the tutorial, participants answered the same understanding questionnaire, and then the deployment of the robot started. The procedure was repeated for every shift.

From this point on, the system worked autonomously 24 hours, 7/7 days throughout the study period, and the researchers went to the facility only to fix an autonomous navigation issue during the first weeks and to collect the weekly questionnaires until the end of the pilot.

An additional tutorial session took place in the middle of the study, addressed to those participants who joined a shift to cover the holiday breaks of the permanent staff. The timeline's summary is shown in Fig. 3.

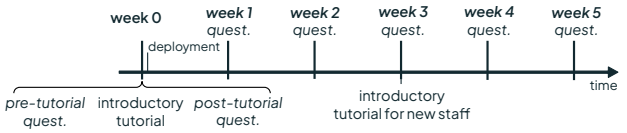

Fig. 3. Timeline of the tutorial, questionnaires (as in Fig. 4) and deployment.

E. Dependent variables

In this section, we specify the understanding questionnaires and metrics. The questionnaire (Fig. 4) included 7 question blocks, Q0 to Q6, and was always answered individually in the presence of a researcher in a "think-aloud" manner to avoid question misunderstandings and cover missing response options. Q0 provided a 1-to-7 Likert scale to rate the general subjective understanding, similar to [20]. The following 6 question blocks are composed of three sub-questions:

1) The question itself, with 5 single-choice answer options. This question was used to evaluate the predictive capabilities and, therefore, the objective understanding.
2) A 1 to 7 Likert scale to rate the level of understanding of each specific topic, similar to what has been done when analysing understandability in [24] (in Section IV we validate that we are effectively measuring subjective understanding by correlating the results to Q0).
3) A multi-choice question to specify the reasoning why participants believe they know the answer.

To compute the objective understanding score, each of the single-choice answers received a mark according to its correctness (0 or 1). Similarly, for the subjective understanding, the response was normalised between 0 and 1. These values can be averaged for all questions or across a particular subset.

This normalisation of both the objective and subjective understanding to 0 and 1 is inspired by previous works [33], [22]. However, a direct comparison between them should be carefully considered because they are obtained differently. However, since both of them are assessed based on the same robot features, we consider that comparing their trends (e.g. an increase in only one of them) is a fair comparison.

F. Independent variables

The following independent variables have been considered:

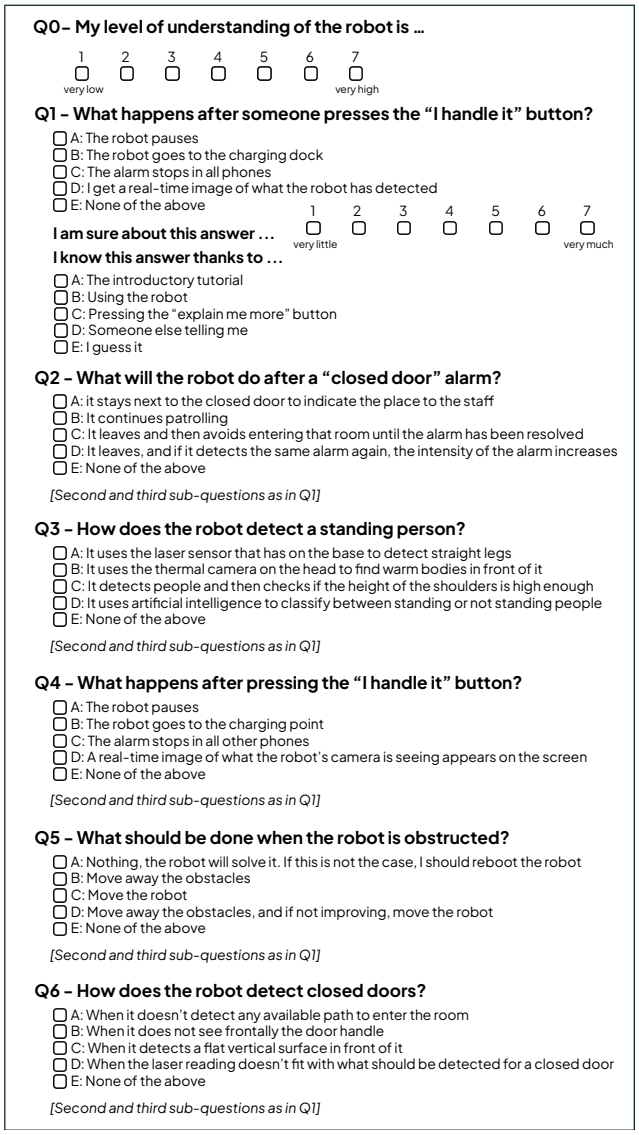

Fig. 4. Distributed questionnaire with question blocks Q0 to Q6.

a) Tutorial attendance: The questionnaire was distributed right before and after attending the introductory tutorial. Thus, a variable to consider is the pre-tutorial against the post-tutorial condition.

b) Worked weeks: We compute the accumulated number of days each participant has worked in contact with the robot at each measurement time-point. We divide this number by the weekly mean workdays of a staff member (5 for the morning and afternoon shifts, 3.5 for the night shift) to obtain the equivalent worked weeks, after rounding to the higher integer. This variable is meant to estimate how understanding is affected by merely sharing the space and witnessing the robot in action. Because not all the participants started on the same date, and the extent of holidays was distinct for each of them, the equivalent worked weeks varied throughout the pilot differently for participants.

c) Usage hours: The *worked weeks* variable does not necessarily mean that participants interact with the system,

as they are free to interact as much as they want. On the contrary, the *usage hours* variable measures the actual level of interaction. We are aware that in this pilot it is challenging to obtain an accurate metric for this variable, particularly given that the robot operates autonomously in the field for 5 weeks, 24 hours a day, without the research team's presence. We base this metric on the accumulated logged-in hours in the system.

d) Learnable vs unlearnable features: Based on the participant's interactions with the system, only specific features can be learned, that is, the information was accessible in some way and at some point. When assessing the *tutorial attendance*, the *learnable features* include only questions Q1 and Q4, since these questions address aspects explicitly explained in the tutorial and that can be further reinforced through usage. When evaluating the *worked weeks* and *usage hours*, questions Q2 and Q4 are also included in the *learnable features* for participants who have been working alongside the system, as their answers can be acquired through observation of the robot's behaviour. Finally, questions Q3 or Q6 are included in the *learnable features* only for participants who received a related explanation to the specific question after pressing the "explain me more" button. For example, the explanation that is considered related to Q6 is "The robot assumed the door was closed because it did not find any available path to go into the room.". A summary of the distribution of questions among *learnable* and *unlearnable features* can be seen in Table I.

TABLE I
CATEGORISATION BETWEEN LEARNABLE AND UNLEARNABLE FEATURES.

		learnable	**unlearnable**
Has the participant attended the introductory tutorial?	yes	Q1, Q4	-
	no	-	Q1, Q4
Has the participant been working while the robot was operating?	yes	Q2, Q5	-
	no	-	Q2, Q5
Has the participant requested an explanation related to the question?	yes	Q3, Q6	-
	no	-	Q3, Q6

G. Hypotheses

Given the above-mentioned dependent and independent variables, we define the following hypotheses:

- **H1.1**: The *tutorial attendance* leads to an increase of the subjective understanding for both the *learnable features* and *unlearnable features*.
- **H1.2**: The *tutorial attendance* leads to an increase of the objective understanding only for the *learnable features* category.
- **H2.1**: The subjective understanding increases with the *worked weeks* for both the *learnable features* and *unlearnable features*.
- **H2.2**: The objective understanding increases with the *worked weeks* only for the *learnable features* category.
- **H3.1**: The subjective understanding increases with the *usage hours* for both the *learnable features* and *unlearnable features*.
- **H3.2**: The objective understanding increases with the *usage hours* only for the *learnable features* category.

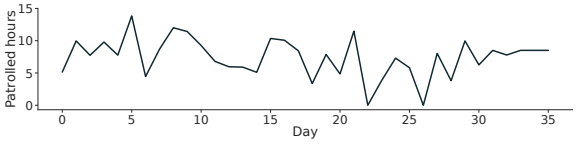

Fig. 5. The daily patrolled hours during the five weeks demonstrate the continued usage of the system with no relevant decay over time.

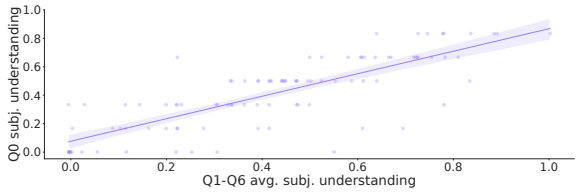

Fig. 6. Correlation between the average score of the subjective understanding of questions Q1 to Q6 and the subjective understanding of question Q0.

IV. RESULTS

In this section, we report the results of the study, with a significance level of $\alpha = 0.05$. In each of the below evaluations, we report the number of samples n, which will vary across different tests. A total of $M = 105$ understanding questionnaires were completed, with an average of 15 participants per questionnaire round. Due to noteworthy holiday breaks and irregular shifts, not all participants were present throughout all the study, with varying participation.

In a study in the wild prolonged in time, the system must be useful and robust to avoid drawing wrong conclusions caused by low relevance in real-world tasks. The iterative co-designing process, which we carried out for 7 months before the deployment, ensured a relevant and functional system. Observing the patrolled hours' evolution in Fig. 5, we verify the robot was practically continuously used for the deployed weeks with no significant continuous drop. During the deployment, the robot travelled more than 78km, entering a total of 7656 rooms.

To verify that the Q1-Q6 Likert questions are well-framed, we correlate the values of Q0 with those of Q1-Q6. As depicted in Fig. 6, we confirm they are strongly correlated (Pearson correlation, $r = 0.80$, $p < 0.001$, $n = 105$). First, this assessment validates that the Q1-Q6 Likert statements effectively reflect subjective understanding as in Q0. Moreover, the correlation indicates that the perceived complexity of the individual Q1-Q6 Likert questions is adequate, that is, it is neither excessively high (otherwise, the general subjective level of understanding in Q0 would be high while the mean score across questions would be low) nor overly simplistic (which would lead to a high score across Q1-Q6 Likert questions, while the general subjective understanding in Q0 would remain low).

Next, we confirm **H1.1** and **H1.2** from the results in Fig. 7. With a sample of $n = 19$, which is the number of participants that did both the pre and post-tutorial questionnaires, their subjective understanding increases for both the *learnable features* (Wilcoxon signed-rank test, mean pre-tutorial $\mu_0 = 0.14$,

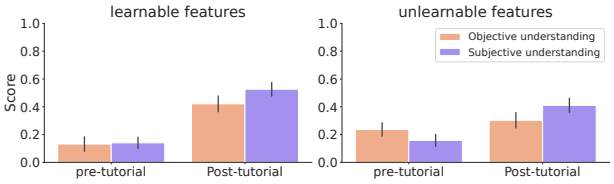

Fig. 7. Effect of attending the tutorial, comparing understanding right before and after the tutorial and separating by features that were explained or not in the tutorial. Vertical lines indicate the standard error.

mean post tutorial $\mu_1 = 0.53$, effect size Cohen's $d = 1.95$, $p < 0.001$) and the *unlearnable features* (Wilcoxon signed-rank test, $\mu_0 = 0.16$, $\mu_1 = 0.41$, $d = 1.21$, $p < 0.001$) conditions. We also observe an increase in the objective understanding in the *learnable features* condition (Wilcoxon signed-rank test, $\mu_0 = 0.13$, $\mu_1 = 0.42$, $d = 1.21$, $p < 0.001$), but not for the *unlearnable features* one (Wilcoxon signed-rank test, $\mu_0 = 0.24$, $\mu_1 = 0.30$, $d = 0.29$, $p = 0.37$). We selected the Wilcoxon test due to discrete data on paired individuals which did not pass a Shapiro-Wilk normality test.

To evaluate the remaining hypotheses, we shift Q2 and Q4 to the *learnable features* category, and also Q3 and Q6 for participants who pressed the "explain me more" button, as explained in Sec. *III.F*.

Concerning the hypotheses **H2.1** and **H2.2**, we do find trends according to the hypotheses, but the results do not provide enough significance to validate them. We compare the levels of understanding of the participants who have worked up to one week ($n = 23$) with the ones who have worked for an extended period, which we consider to be at least one month ($n = 28$). We exclude the tutorial questionnaires' data to validate the hypotheses only based on permanent usage. The full evolution is represented in Fig. 8. More specifically, we have obtained the following results: (1) subjective understanding and *learnable features* (Mann–Whitney U test, $\mu_0 = 0.36$, $\mu_1 = 0.58$, $d = 0.83$, $p = 0.10$); (2) subjective understanding and *unlearnable features* (Mann–Whitney U test, $\mu_0 = 0.27$, $\mu_1 = 0.45$, $d = 0.77$, $p = 0.12$); (3) objective understanding and *learnable features* (Mann–Whitney U test, $\mu_0 = 0.30$, $\mu_1 = 0.37$, $d = 0.35$, $p = 0.48$); and (4) objective understanding and *unlearnable features*(Mann–Whitney U test, $\mu_0 = 0.20$, $\mu_1 = 0.17$, $d = 0.11$, $p = 0.83$). We employed the Mann–Whitney U test to compare two distributions assumed to be independent, where the data was discrete and did not pass a Shapiro-Wilk normality test.

We confirm **H3.1** and partially confirm **H3.2** by correlating the logged-in hours on the system with the understanding scores, as shown in Fig. 9. We exclude the tutorial questionnaires, resulting in $n = 67$ remaining data points. We verify **H3.1**, since the subjective understanding positively correlates for both the *learnable features* (Pearson correlation, $r = 0.62$, $p < 0.001$) and the *unlearnable features* (Pearson correlation, $r = 0.30$, $p = 0.004$) conditions. Regarding **H3.2**, we find a non-significant positive trend for the *learnable features* (Pearson correlation, $r = 0.10$, $p = 0.42$) while for the *unlearnable features* condition there is a significant negative

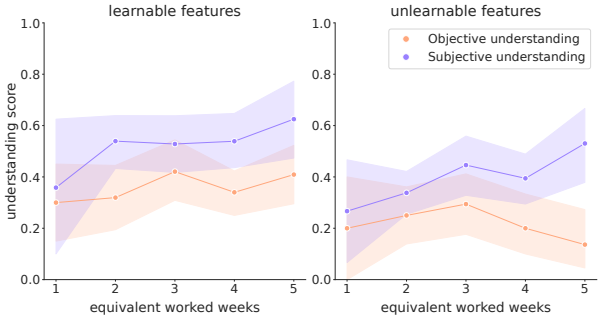

Fig. 8. Evolution of the objective and subjective understanding over the worked weeks, separately for the features that users can and cannot learn. Vertical bands indicate 95% confidence intervals.

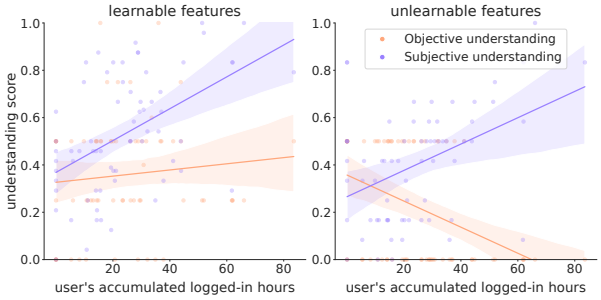

Fig. 9. Correlation between the number of accumulated logged-in hours in the app, which is a measure of the usage of the system, with the levels of understanding. The left and right plots include questions related to features that users can and cannot learn given their usage, respectively. Vertical bands indicate 95% confidence intervals.

correlation (Pearson correlation, $r = -0.40$, $p < 0.001$).

As a complementary analysis, we also evaluated the justifications behind the participant's answers, i.e. the source of their knowledge gathered in the understanding questionnaire "I know this answer thanks to..." from Fig. 4. The results are summarised in Fig. 10 segmented by answers where the reported subjective understanding was either low, middle or high (1-2, 3-5 and 6-7 in the Likert, respectively).

V. Discussion

A. Objective vs subjective understanding

The main finding from the above-reported results is that the subjective understanding of a robot can increase while the objective understanding does not, meaning that the two types of understanding can be decoupled. This supports the need to measure both metrics to correctly track both what humans think they know and what they actually know.

When assessing the *unlearnable features* category, our results reveal that when the information to acquire knowledge about specific features was not available, the subjective understanding significantly increased (for the *tutorial attendance* and *usage hours*), while the objective understanding did not.

The results for the *worked weeks* variable point in the same direction, although with not enough statistical significance. This fact could imply that using the system (*usage hours*) has a stronger impact compared to merely working more hours in the same space as the robot (*worked weeks*).

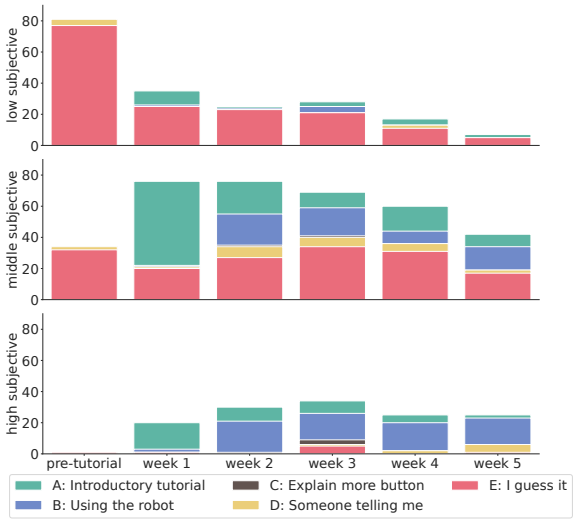

Fig. 10. Count of the reasoning behind the choices that participants reported to guide their answers along the weeks, and separated by low (top), medium (middle) and high (bottom) reported levels of subjective understanding. This question was multi-choice, so the figure reports the total count of each reasoning across all the questionnaires.

We corroborated that the objective understanding does not increase for the *unlearnable features*, but we did not foresee that it could decrease, as we observe in the results. We attribute this fact to robot behaviours that are not aligned with what the participants intuitively expect. In those cases, the user's beliefs are never corrected as they refer to information about features that have never been disclosed, either because the user has not experienced them or asked for information about them. In any case, we exclusively aimed to verify there is not an increase in the objective understanding for the *unlearnable features* category, which our results do support.

Regarding the *learnable features* category, we aimed to verify that both types of understanding do increase when information is available. We have statistically verified that for the *tutorial attendance*, and for the *usage hours* only for the subjective understanding, while finding positive trends in the rest of the cases. Moreover, the fact that for the *usage hours* and the *learnable features* category the subjective understanding significantly increases while the objective understanding shows only a positive trend could be a further sign of decoupling of the two types of understanding. One implication would be that, for aspects that have been explained, humans believe they understand them better than they actually do. This trend is also observable for the *worked weeks* evolution.

B. Participants' reported reasoning

Results from Fig. 10 indicate that before the tutorial, the main reasoning behind their answers is "I guess it", while right after the tutorial, the tutorial itself becomes the main reasoning, except for the low subjective understanding category. Over the weeks, we can observe how participants associate with "using the robot" aspects that were before attributed to the training, especially for the middle and high subjective categories. This trend is not present in the low subjective understanding category, where the "I guess it" reasoning remains the main choice. However, we did not expect the constant and relatively high share of the "I guess it" reasoning for the middle category. This implies that many participants are fairly sure that they know something just because they are envisioning it, but they still rate to have a middle subjective understanding level. This finding could explain why subjective understanding increases even for aspects that cannot be learned, as longer contact with a system might lead to an increase in the subjective understanding of aspects that humans know out of their imagination. A possible reason would be that increased confidence in the usage causes a generalised baseline increase in the perception of understanding, that will remain high unless proper system feedback makes users aware of their real level of understanding, adjusting their illusion.

C. Limitations

On the one hand, we reckon that the level of direct interaction with the robot system was limited, as a result of the high level of the system's autonomy, where intervention from the user was partially required to fulfil the tasks. Although participants shared their workspace with an autonomous robot for many hours each day, the robot rapidly became integrated into their daily routines. On the other hand, we could not monitor a constant set of participants for the whole study due to changes in shifts and holidays, and we might have missed some significant effects due to the relatively reduced average number of participants per questionnaire round.

However, an autonomous system in a real-world context also adds greater value to the results, as they are based on participants' genuine interactions and understanding of a system they were free to use, and that was co-designed according to their needs. This is not common in the HRI community, and we would like to stress that part of the contribution of this work is the outcome of a controlled experiment of an autonomous robot deployed in the wild for an extended period.

D. Future research directions

First, our findings demand a more detailed study of the impact that objective and subjective understanding have on other desiderata, such as trust, usability and performance. We expect that other desiderata are going to be unevenly affected by the two dimensions of understanding. For example, trust might be more influenced by subjective, self-perceived understanding, while human-robot performance in a certain task would be more impacted by objective understanding.

Apart from investigating how understanding impacts other desired effects on users, research should illustrate how both dimensions of understanding can be shaped. On the one hand, sufficient explainability should be provided to leverage the objective understanding according to application, user and context-specific targets. On the other hand, future research should clarify how subjective understanding can be influenced not only by the objective understanding itself but also by feedback from the robot. This feedback would tune the perception of understanding, but could also provoke changes in

expectations, which would modify the internal scale that users maintain to assess their degree of understanding.

Nevertheless, although actively manipulating subjective understanding may provide benefits, we foresee issues if significant understanding illusions suddenly vanish, which could result in frustration, mistrust and, ultimately, abandonment of the system. Therefore, we anticipate that robots should aim at maintaining a low mismatch between objective and subjective understanding. The other extreme, where the objective understanding is significantly higher than the subjective one, could imply that the benefits of a higher subjective understanding, such as satisfaction or confidence, are missed.

VI. Conclusions

This work sought to enhance the comprehension of the dynamics of user understanding in HRI. To this end, we conducted a user study with a robot deployed in a real-world healthcare facility. Our evaluation examined changes in objective and subjective understanding after a tutorial session, considering the time spent working in the shared environment, and for different levels of interaction with the system. The findings reveal that objective and subjective understanding can evolve independently. Notably, subjective understanding may increase even when objective understanding remains unchanged. We advise exploring sufficiently how to shape both the objective and subjective understanding with explainability measures and investigating how they influence other desiderata such as trust, usability or performance.

VII. Acknowledgements

This work has been supported by Horizon Europe Marie Skłodowska-Curie grant N. 101072488 (TRAIL) and Horizon 2020 grant N. 857188 (SAFE-LY-PHARAON).

References

[1] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum, "What do we want from explainable artificial intelligence (xai)?," *Artificial intelligence*, vol. 296, p. 103473, 2021.

[2] T. Speith and M. Langer, "A new perspective on evaluation methods for explainable artificial intelligence (xai)," in *Int. Requirements Engineering Conference Workshops*, pp. 325–331, 2023.

[3] T. Hellström and S. Bensch, "Understandable robots-what, why, and how," *Journal of Behavioral Robotics*, vol. 9, no. 1, pp. 110–123, 2018.

[4] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[5] Y. Rong, T. Leemann, T.-T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, and E. Kasneci, "Towards human-centered explainable ai: A survey of user studies for model explanations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 4, pp. 2104–2122, 2023.

[6] M. Reimann, J. van de Graaf, N. van Gulik, S. Van De Sanden, T. Verhagen, and K. Hindriks, "Social robots in the wild and the novelty effect," in *Int. Conf. on Social Robotics*, pp. 38–48, 2023.

[7] C. V. Smedegaard, "Reframing the role of novelty within social HRI: from noise to information," in *Int. Conf. on Human-Robot Interaction*, pp. 411–420, 2019.

[8] M. Jung and P. Hinds, "Robots in the wild: A time for more robust theories of human-robot interaction," *Transactions on Human-Robot Interaction*, vol. 7, no. 1, pp. 1–5, 2018.

[9] T. Kulesza, S. Stumpf, M. Burnett, and I. Kwan, "Tell me more? The effects of mental model soundness on personalizing an intelligent agent," in *Proc. Conf. on human factors in computing systems*, pp. 1–10, 2012.

[10] L. Wachowiak, O. Celiktutan, A. Coles, and G. Canal, "A Survey of Evaluation Methods and Metrics for Explanations in Human-Robot Interaction," in *ICRA Workshop on Explainable Robotics*, 2023.

[11] P. N. Johnson-Laird, *Mental models: Towards a cognitive science of language, inference, and consciousness*. No. 6, Harvard University Press, 1983.

[12] D. A. Norman, "Some observations on mental models," in *Mental models*, pp. 15–22, Psychology Press, 2014.

[13] P. Dillenbourg, S. Lemaignan, M. Sangin, N. Nova, and G. Molinari, "The symmetry of partner modelling," *Int J of Computer-Supported Collaborative Learning*, vol. 11, p. 227, 2016.

[14] J. E. Mathieu, T. S. Heffner, G. F. Goodwin, E. Salas, and J. A. Cannon-Bowers, "The influence of shared mental models on team process and performance," *J of applied psychology*, vol. 85, no. 2, p. 273, 2000.

[15] J. Tullio, A. K. Dey, J. Chalecki, and J. Fogarty, "How it works: a field study of non-technical users interacting with an intelligent system," in *Proc Conf on Human Factors in Computing Systems*, pp. 31–40, 2007.

[16] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond explanation as soliloquy," *arXiv preprint arXiv:1701.08317*, 2017.

[17] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[18] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[19] J. D. Trout, "The psychology of scientific explanation," *Philosophy Compass*, vol. 2, no. 3, pp. 564–591, 2007.

[20] L. Rozenblit and F. Keil, "The misunderstood limits of folk science: An illusion of explanatory depth," *Cognitive science*, vol. 26, no. 5, pp. 521–562, 2002.

[21] J. Kruger and D. Dunning, "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments," *Journal of personality and social psychology*, vol. 77, no. 6, p. 1121, 1999.

[22] D. Kontogiorgos and J. Shah, "Questioning the robot: Using human non-verbal cues to estimate the need for explanations," in *Proc Int Conf on Human-Robot Interaction*, pp. 717–728, 2025.

[23] M. Faria, F. S. Melo, and A. Paiva, "Understanding robots: Making robots more legible in multi-party interactions," in *Int Conf on Robot&Human Interactive Communication*, pp. 1031–1036, 2021.

[24] G. LeMasurier, A. Gautam, Z. Han, J. W. Crandall, and H. A. Yanco, "Reactive or proactive? How robots should explain failures," in *Proc Int Conf on Human-Robot Interaction*, pp. 413–422, 2024.

[25] Y. Mualla, I. Tchappi, T. Kampik, A. Najjar, D. Calvaresi, A. Abbas-Turki, S. Galland, and C. Nicolle, "The quest of parsimonious xai: A human-agent architecture for explanation formulation," *Artificial intelligence*, vol. 302, p. 103573, 2022.

[26] L. Yuan, X. Gao, Z. Zheng, M. Edmonds, Y. N. Wu, F. Rossano, H. Lu, Y. Zhu, and S.-C. Zhu, "In situ bidirectional human-robot value alignment," *Science robotics*, vol. 7, no. 68, p. eabm4183, 2022.

[27] A. Silva, M. Schrum, E. Hedlund-Botti, N. Gopalan, and M. Gombolay, "Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction," *Int J of Human-Computer Interaction*, vol. 39, no. 7, pp. 1390–1404, 2023.

[28] G. Y. Gombolay, A. Silva, M. Schrum, N. Gopalan, J. Hallman-Cooper, M. Dutt, and M. Gombolay, "Effects of explainable artificial intelligence in neurology decision support," *Annals of clinical and translational neurology*, vol. 11, no. 5, pp. 1224–1235, 2024.

[29] R. Yang and M. W. Newman, "Learning from a learning thermostat: lessons for intelligent systems for the home," in *Int Joint Conf on Pervasive and Ubiquitous Computing*, pp. 93–102, 2013.

[30] A. Bunt, M. Lount, and C. Lauzon, "Are explanations always important? A study of deployed, low-cost intelligent interactive systems," in *Proc Int Conf on Intelligent User Interfaces*, pp. 169–178, 2012.

[31] S. Lemaignan, J. Fink, P. Dillenbourg, and C. Braboszcz, "The cognitive correlates of anthropomorphism," in *WS in the Human-Robot Interaction Conference*, 2014.

[32] F. Gebellí, R. Ros, S. Lemaignan, and A. Garrell, "Co-designing Explainable Robots: A Participatory Design Approach for HRI," in *Int Conf on Robot&Human Interactive Comm*, pp. 1564–1570, 2024.

[33] X. Wang and M. Yin, "Effects of explanations in ai-assisted decision making: Principles and comparisons," *Tran on Interactive Intelligent Systems*, vol. 12, no. 4, pp. 1–36, 2022.