Evaluating the Impact of Explainability on the Users' Mental Models of Robots over Time

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

***Abstract*— To evaluate how explanations affect the users' understanding of robots, researchers typically elicit the user's Mental Model (MM) of the robot and then compare it to the robot's actual decision-making and behaviour. However, the user's self-rating of their level of understanding, which we define as "user-perceived understanding", is generally not evaluated. Moreover, this evaluation is typically done only once, while robots are often expected to interact with the same users over long periods. In this work, we suggest a framework to analyse the evolution of the mental models over time across the dimensions of completeness and correctness. We argue that the goal of explainability should be two-fold. On one hand, it should help align the user's perceived understanding with the real one. On the other hand, explainability should enhance the completeness of the mental model to a target level, which varies depending on the user type, while also striving for maximum correctness.**

I. INTRODUCTION

One of the main goals of designing explainable robots is to improve the understanding of users about the robots' decisions and behaviours [12]. In turn, this will contribute to achieving other desired effects, such as raising users' satisfaction when interacting with these robots, who find them more usable, and eventually, who trust them more.

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 [5]. The evaluation of the effects of explainability on the user's understanding of the robot is often done in terms of building "better" mental models [6]. However, this evaluation is typically conducted after just a single interaction with the robot, which often fails to address the novelty effect adequately. Moreover, many robot use cases are designed for long-term interactions, involving multiple engagements over extended periods with the same user. Although the evolution of mental models over time has been studied from some perspectives (e.g. concerning robot-attributed anthropomorphism [8]), up to the authors' knowledge, it has not been addressed from an explainability point of view in robotics.

In this work, inspired by the works in [6], [13], we formalise the characterisation of mental models through two properties: completeness and correctness. The former refers to how many aspects of the robot are known, while the latter refers to the accuracy of those known facts. Moreover, we also consider the concept of user-perceived understanding, which is normally overlooked in previous works.

II. RELATED WORK

Regarding the evaluation of explainability, the review work in [6] divides eXplainable Artificial Intelligence (XAI) metrics into groups according to the measured aspect: the explanation goodness, the user satisfaction, the user's mental model, curiosity, trust, and human-AI performance. For the sub-field of eXplainable HRI (XHRI), a survey [13] defines similar and overlapping groups, which are the explanation content quality, faithfulness, effects (which include trust, mental models, and human-robot performance) and timing of explanations.

In many XHRI studies, the primary goal is to analyze how explainability measures impact user trust in robots [2], with several trust scales being developed [14], [11]. However, concerns regarding how trust is measured have been raised [3], [9]. In this work, we focus on the user's understanding of systems (through the evaluation of the user's mental model), which is considered a way to foster trust in the system [6].

According to [6], the elicitation of mental models is usually complex, but there is a consensus that mental models can be inferred from empirical evidence. It has been recommended to combine more than one method for eliciting a mental model [6], [13]. Several relevant properties have been identified to analyse them, including correctness and completeness [6], [13]. However, to the author's knowledge, analysing the elicited mental models has not yet been formalized in the XHRI field. Moreover, previous works on mental models' evaluation in XHRI do not specify how to consider their evolution over time.

Lastly, previous works have primarily focused on assessing mental models by evaluating "real" knowledge —what individuals actually understand about the world around them— to appraise explainability. However, research in social sciences evidences that people tend to have a wrong belief of their understanding [10], [7], but this phenomenon has not been well studied in the XAI field. We advocate that user-perceived understanding is part of the construct of mental models, considering that explainability should be a mechanism that steers the user-perceived levels of understanding closer to the real ones.

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

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

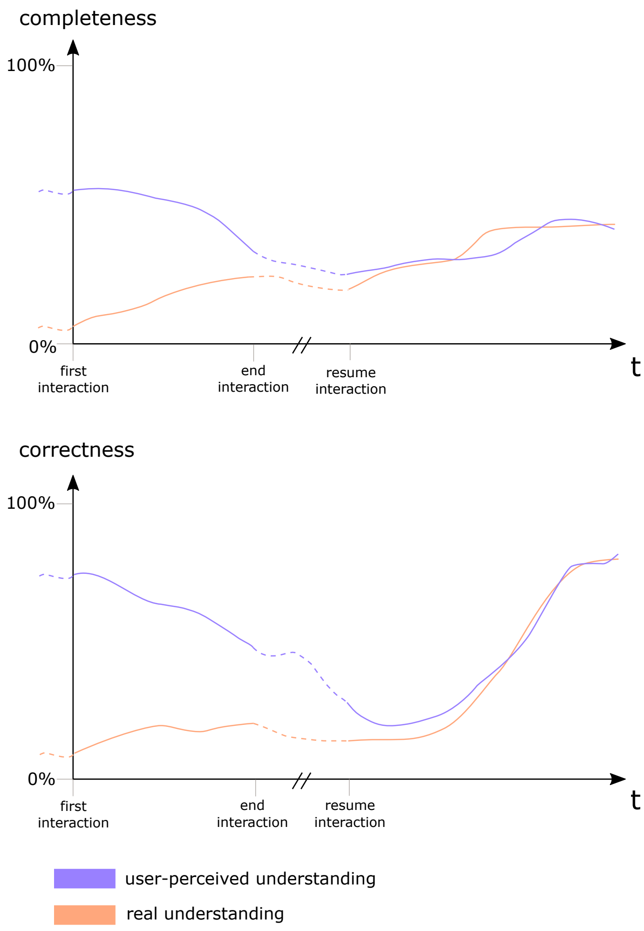

Fig. 1. Example curves for the evolution of a mental model's related understanding over time across the dimensions of completeness and correctness.

III. PROPOSED APPROACH

We propose to evaluate the evolution of the explainability impact on the individual's understanding of the robot by considering the perceived and real understanding across two dimensions: completeness and correctness. Completeness refers to the breadth of knowledge about the system's features, capabilities, and behaviour, whereas correctness is the accuracy of that knowledge.

We thus propose measuring and quantifying both dimensions for the real and the user-perceived understanding at different time points. This evaluation can be done by representing the curves for multiple users separately and then looking for patterns and profiles, or directly aggregating all the users' data to discover general trends.

Figure 1 provides an example of such a representation. Discontinuous lines represent periods when there are no interactions with the system. Here we include all previous experiences before the first interaction, which could correspond to preconceived knowledge, biases from different sources, or even experiences with similar systems. We consider the initial value of the user-perceived understanding as the *initial user expectations*, and the difference with the real understanding would correspond to the *expectations mismatch*.

We next describe the evolution of the curves through time based on explainability mechanisms and their interrelationships. We conclude by introducing practical considerations of the framework, including both methods for measuring the user-perceived and real understanding, and areas requiring further research.

A. Evolution of the MM curves shaped by explainability

We propose that the impact of explainability should be two-fold. First, explainability measures should adjust the user's beliefs and expectations to align the user-perceived understanding levels closer to the real ones, thereby reducing expectation mismatches. Second, the real understanding completeness should reach a target level, which varies depending on the user type, while aiming for high correctness.

With respect to the user-perceived understanding, explanations should support driving the curve closer to the real understanding. Previous studies have shown that, in general, users have the belief that they have a deeper understanding than the real one [10], [7]. Therefore, we expect user-perceived completeness and correctness curves to start with higher values than the real curves, and to decrease through time to finally track the real curves [10]. It has been argued that the anthropomorphism level of the robot's appearance is highly related to humans' expectations and perceptions [4], so we would expect that the initial gap between the real and user-perceived understanding is higher for more anthropomorphic robots.

Regarding the real understanding, a target completeness value should be defined, which represents the relevant features, skills, strategies, decision-making logic and behaviours of the systems to be known by a particular user type[1]. The robot should provide explainability measures to reach that target. The completeness target value should be carefully defined to ensure that the robot's aspects to be understood are going to be useful for the user, that is, that they will support effective interaction with the system, thus improving the usability and trust in it. Participatory design approaches can be used to define the aspects that are more useful and relevant for the users and define a completeness target that adapts better to the user's needs.

After setting a target completeness level, the robot should aim to raise the correctness to the highest possible values, as we consider that users should not have a wrong interpretation of the different aspects of the robot that they should know. Especially for low completeness values, when users need to achieve a general pragmatic understanding of the system, mechanisms for achieving very high correctness levels are a must. However, when the most complex aspects of a robot need to be understood by expert stakeholders, it might not be possible to achieve 100% correctness. For instance, this will be the case for systems that include black-box AI modules. We consider that the goal of the XAI generation of low-level features is precisely to provide auxiliary outputs that can lead

[1] We suggest to categorise the different users following the IEEE Standard for Transparency of Autonomous Systems P7001 [1] user type definition.

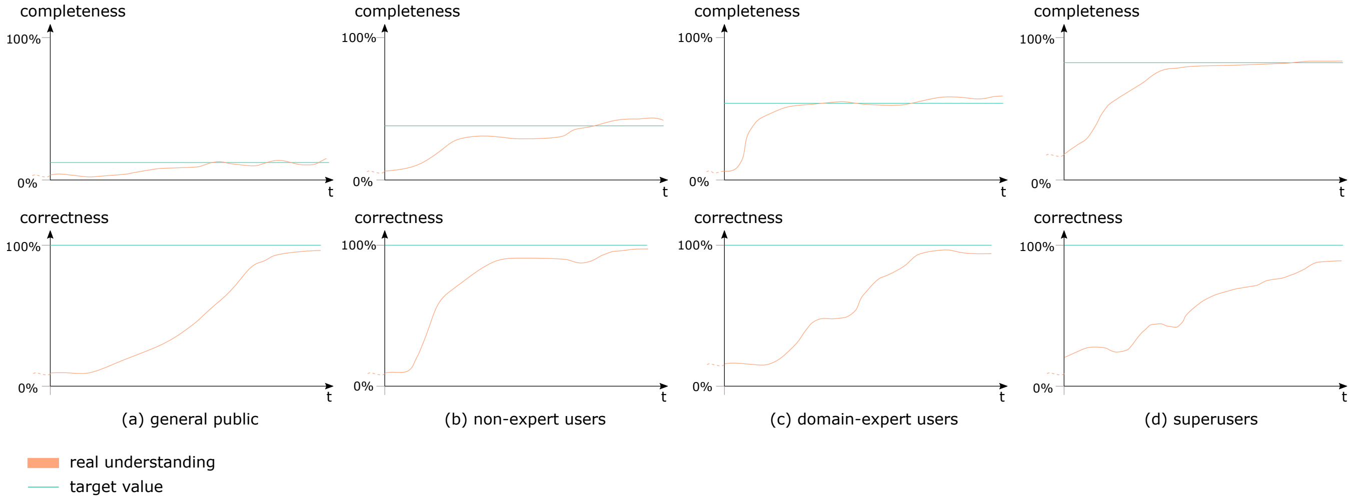

Fig. 2. Example real understanding curves for the (a) general public, (b) non-expert users, (c) domain-expert users and (d) superusers user types from the IEEE Standard P7001 [1]. They are example curves for a hypothetical security robot use case. For different use cases and users, the curve might follow different trends. Target values for completeness and correctness are included.

to higher levels of correctness for users with a high required completeness about complex systems.

As a hypothetical example, we consider a security robot that patrols a certain facility. Figure 2 depicts the MM curves for four types of users (as defined in [1]). The general public or bystanders, such as visitors of that facility, should have a low target completeness value, since they only need a rough idea of the general capabilities of the system, e.g., that the robot can move autonomously and that it will preserve their privacy. Non-expert users (e.g., general staff from the facility) completeness target should be significantly higher, as they should know how to sporadically interact with the robot, such as understanding some of the alarms that it can raise and performing a preliminary assessment. In contrast, domain-expert users, such as the security staff from that facility, should have a higher completeness target. They should know how to command the system, configure it, extensively interpret the alarms triggered by the robot, and assist in recovering from failures, among others. Finally, superusers (e.g., persons responsible for development, fault diagnosis, repair, maintenance and upgrade), would require close to maximum completeness targets. All user types should aim at a correctness of 100%, although for users with very high target completeness, it might be more difficult to reach such a goal.

B. Practical considerations

1) On the curves sampling: Curves in Figure 1 are continuous, but in practice, they will be composed of a set of discrete values. We recommend measuring the mental model metrics as often as possible to build an accurate representation of these, which will allow to apply effective strategies to achieve more complete and correct mental models and reduce expectations mismatch. They should be measured at least before starting the deployment of the system, at the end of the deployment, and before and after large interaction gaps.

2) On the curves assessments: Regarding the measurement of the real understanding, we suggest following the guidelines in [6], which provides a set of elicitation techniques, analysis guidelines and general recommendations.

To assess the user-perceived knowledge, we recommend creating a comprehensive set of questions covering various aspects of the robot, such as its capabilities, decision-making strategies, and behaviours. These questions should range from simple to complex, and include an option to answer ”not known,” indicating gaps in the user’s knowledge. The user-perceived completeness is calculated as the percentage of known aspects. For aspects the user claims to know, they should rate their understanding on a given scale. The user-perceived correctness is then determined as the average of these self-assessed ratings for all aspects the user claimed to know.

3) On the explainability mechanisms: We consider three different explainability mechanisms to impact on shaping those curves:

- *Previous information*, which cover training workshops and provided documentation that pursue building a higher real understanding before the first interaction, but also seek to reduce expectation mismatches by decreasing the user-perceived understanding.
- *Legible* behaviours are performed by the robot during normal usage and aim to improve the real understanding by acting intuitively, that is, in a way that matches what the user would expect. This means that *legibility* would shape the real understanding curve, rather than the user-perceived one.
- *Post-hoc* explainability mechanisms, which are actions triggered after a user request (e.g. “why did you do that?”), impact on increasing the real understanding and thus, reducing the mismatch with the user-perceived understanding. The number of requested *post-hoc* explanations will depend on the user’s curiosity about the system and satisfaction with previous requests.

4) On the influence of curiosity: High levels of curiosity and explanation satisfaction will lead the users to seek higher levels of completeness and correctness. In some cases, this can be crucial to reach the target values, as users who are not interested enough in the system will never reach the target levels of understanding. Moreover, high curiosity and explanation satisfaction will accelerate the improvement of the mental models, and might even lead to reaching higher values than the target ones. For example, a non-expert user might want to go beyond the necessary aspects for the basic usage of a system, seeking to know more about the inner workings of the robot's decisions and behaviours, thus advancing towards the domain-expert user's completeness target.

5) On the uncertainty: Finally, there will be in practice uncertainty associated with the correctness and completeness of the understanding curves. Future work could consider the inclusion of uncertainty by replacing the curves with distributions (e.g. normal distributions) that evolve over time. Evaluation techniques should consider and reduce the uncertainty coming from the mental model elicitation process. Moreover, explainability measures should attempt to reduce the uncertainty that users themselves report, e.g., some users might be able to provide an interval for the perceived completeness and correctness instead of a fixed value, as they would consciously know that there is some uncertainty in their belief.

IV. CONCLUSIONS

This work proposes a novel framework for evaluating the effects of explainability on users' mental models of robots over time. It distinguishes between real and user-perceived understanding, focusing on the dimensions of completeness and correctness.

Key contributions include introducing the dual-dimensional evaluation framework, emphasizing the often-overlooked user-perceived understanding, and suggesting practical methods and considerations for measuring the evolution of the mental models over time. Moreover, this work argues that the goal of explainability should be to improve real understanding while reducing the mismatch between real and user-perceived understanding.

Future work will further formalize the evaluation of real and user-perceived understanding, empirically validate the framework through user studies in diverse HRI contexts and explore methods to consider both the curiosity and uncertainty in the mental models.

ACKNOWLEDGMENT

This work has been partially supported by Horizon Europe Marie Skłodowska-Curie grant N. 101072488 (TRAIL).

REFERENCES

[1] Standard for transparency of autonomous systems. *IEEE Std 7001-2021*, pages 1–54, 2022.

[2] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*, pages 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[3] Meia Chita-Tegmark, Theresa Law, Nicholas Rabb, and Matthias Scheutz. Can you trust your trust measure? In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*, pages 92–100, 2021.

[4] Kerstin S Haring, Katsumi Watanabe, Mari Velonaki, Chad C Tossell, and Victor Finomore. Ffab—the form function attribution bias in human–robot interaction. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):843–851, 2018.

[5] Thomas Hellström and Suna Bensch. Understandable robots-what, why, and how. *Paladyn, Journal of Behavioral Robotics*, 9(1):110–123, 2018.

[6] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science*, 5:1096257, 2023.

[7] Justin Kruger and David 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*, 77(6):1121, 1999.

[8] Séverin Lemaignan, Julia Fink, Pierre Dillenbourg, and Claire Braboszcz. The cognitive correlates of anthropomorphism. In *2014 Human-Robot Interaction Conference, Workshop" HRI: a bridge between Robotics and Neuroscience"*, 2014.

[9] Tim Miller. Are we measuring trust correctly in explainability, interpretability, and transparency research? *arXiv preprint arXiv:2209.00651*, 2022.

[10] Leonid Rozenblit and Frank Keil. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5):521–562, 2002.

[11] Kristin E Schaefer. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In *Robust intelligence and trust in autonomous systems*, pages 191–218. Springer, 2016.

[12] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. A two-dimensional explanation framework to classify AI as incomprehensible, interpretable, or understandable. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 119–138. Springer, 2021.

[13] Lennart Wachowiak, Oya Celiktutan, Andrew Coles, and Gerard Canal. A survey of evaluation methods and metrics for explanations in human–robot interaction (HRI). In *ICRA2023 Workshop on Explainable Robotics*, 2023.

[14] Rosemarie E Yagoda and Douglas J Gillan. You want me to trust a robot? the development of a human–robot interaction trust scale. *International Journal of Social Robotics*, 4:235–248, 2012.