Participatory design for explainable robots

Ferran Gebellí Guinjoan
ferran.gebelli@pal-robotics.com
Pal Robotics
Barcelona, Spain

Raquel Ros
raquel.ros@pal-robotics.com
Pal Robotics
Barcelona, Spain

Anaís Garrell
anais.garrell@upc.edu
Universitat Politècnica de Catalunya
Barcelona, Spain

ABSTRACT

Currently, many works focus on algorithms that generate explanations, and evaluate the impact on user trust and understanding over robots afterwards. We suggest a user-centric approach for explainability from the very beginning. Concretely, we propose a participatory design methodology with three main steps: (1) to find together with users what makes an application usable, (2) to co-design the interface to make sure it is intuitive and understandable, and only then (3) to redefine and develop the robot's functionality and autonomous behaviours. We perform the proposed framework's first steps together with different stakeholders in a geriatric unit at an intermediate care centre.

CCS CONCEPTS

• **Computer systems organization** → **Robotics**; • **Human-centered computing** → **Participatory design**.

KEYWORDS

Participatory Design, HRI, Explainability, Transparency

ACM Reference Format:
Ferran Gebellí Guinjoan, Raquel Ros, and Anaís Garrell. 2024. Participatory design for explainable robots. In *Explainability for Human-Robot Collaboration Workshop '24, March 11, 2024, Boulder, Colorado.* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

In many Human-Robot Interaction (HRI) studies, the main target is to analyse the user trust on robots [3]. However, the concept of trusting a robot is not well generalized, i.e. should we trust robots the same way we trust bridges, or as we trust people? That is, in terms of performance or beyond? Moreover, many works often forget that trust is a path to another goal in robotics, which is that humans become eager users of robots. If people do not find robots usable, they will discontinue using them.

If a robot is not understandable, it will not be usable. We consider that for robots to be understandable, humans should be able to predict the robot's behaviour. Explainability is a mechanism to make systems more understandable, and although there are many explainability works that point out the need of focusing more on the users to achieve understandable systems [7, 14], few address it [3].

X-HRI Workshop, March 11, 2024, Boulder, Colorado

Figure 1: In-situ usability design in a the geriatric unit.

Understandable robots should grant an intuitive user experience (UX), employing the concept as broad as the interface capabilities of robots, e.g. gestures, speech, displays. The UX explainability perspective has not been well studied in HRI, though there are frameworks for non-robotic systems [12].

Participatory Design (PD) approaches are a promising way to put a strong focus on users, actively involving them in the co-design process as integral members of the design team from the very beginning. This way, users will shape robots that better adapt to their needs, which will in turn re-shape the way users conceive robots, following a mutual shaping process [19]. Furthermore, PD allows to include the explainability dimension from the start of the design process, following a a transparency-by-design approach [8].

Most works in explainable HRI focus on ways to generate explanations [26], and then evaluate their effect on improving the understanding and trust among users. In this work, we propose to approach explainability in HRI from the opposite direction: through participatory design, we suggest to first discover what makes a robot usable, and co-design the interface with a strong focus on the user's explainability needs. We argue that only when users approve that interactions are intuitive and understandable, should the focus be shifted to the robot performance, which should be able to provide the required transparency and explainability.

This work is structured as follows. Section 2 provides a literature review. Section 3 details the proposed methodology. Section 4 presents some insights and lessons learned from an on-going PD process in a geriatric unit at an intermediate care centre (Figure 1), and Section 5 concludes the paper.

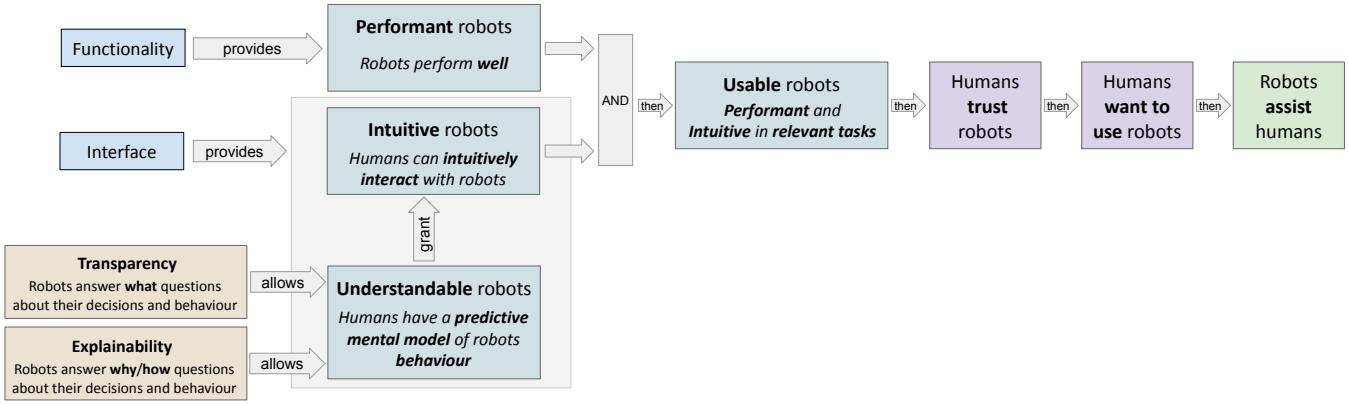

Figure 2: The relation between Usable, Performant, Intuitive and Understandable robots.

2 RELATED WORK

In this section, we review the terminology used in the eXplainable Artificial Intelligenc (XAI) field, and we continue surveying eXplainable HRI (XHRI) sub-field. Then, we explore publications where PD is used in robotics, and we finish looking into works that use PD for XAI in non-robotic applications. To the extent of the authors' knowledge, there are no prior works where PD is used for XAI in robotics.

2.1 Transparency and Explainability

The field of XAI focuses on making AI systems more explainable. Many XAI works try to define *explainability*, which has related terms such as *transparency*, *understandabilty* or *interpretability*. A review publication [25] analyses the terms used in the XHRI literature, being *explainability* and *transparency* the most used ones, with variant definitions across different publications. The IEEE Standard for Transparency of Autonomous Systems P7001 [1] defines *transparency* as "the transfer of information from an autonomous system [...] in a form meaningful to the stakeholder", while *explainability* is defined as *transparency* addressed to non-expert users. In [24] a two-stage framework is proposed, where *transparency* discloses information about a system to make it *interpretable*, and *explainability* clarifies the information to provide *understandability*.

In this work, we consider that both *transparency* and *explainability* are mechanisms to *understand* a robot better, considering that *transparency* allows to answer "what" questions about the robot decisions and behaviour, while *explainability* would allow to answer "how" and "why" questions [24].

2.2 Explainable Human-Robot-Interaction

A literature review on explainable robots [3] reveals that a considerable portion of the reviewed papers propose conceptual studies without evaluations. Nevertheless, several works do focus on user studies where explainable robots are evaluated with metrics [9]. Another literature review [25] provides an overview of the effects of explainability on trust, interaction robustness and mental model of the robot, and shared tasks efficiency, finding that explainability almost always correlates with higher trust and robustness, while efficiency has a mix of positive correlation and non-significance.

However, the operationalization of those studies makes hard to generalize practices and explanation modalities. It has been argued that generalized design recommendations for explainability and transparency cannot be defined, and that they should be adapted to the user types [7]. In light of such views, participatory design is a promising approach that can be used to design explainable robotic systems while taking into account different user needs.

2.3 Participatory design approaches

Participatory design is a term often interchanged with co-design [20]. In PD, multiple stakeholders, such as end-users and domain experts [4] contribute to the design process as active co-designers. In PD, non-roboticists can actively collaborate in the robot design [11] by playing a critical role [22] to iteratively construct the emerging design [23].

Concretely, in HRI, PD has been used in several domains, such as the co-design of social robots interacting with teens [5], cognitively impaired citizens [18] and older adults [11]. These works use a various set of tools such as workshops [11, 16], card-sorting [16], role-playing [5] and prototyping [5, 16]. In [4] a generic framework is presented with PD tools that can be applied to a wide range of HRI applications.

2.4 Explainability through participatory design

Several works use PD methodologies to improve the transparency or explainability of AI systems. In [6] a stage-based PD framework to improve transparency of interfaces is defined, which is validated on a fitness coach app. Another generic framework [10] with focus on UX design uses four AI-assisted decision-making tasks to evaluate the methodology. Other works do not propose generic frameworks, but detail the execution of PD processes for various applications such as a learning analytics tool [2], a clinical decision support system [21], or a social media recommender system [13]. However, these works do not follow the transparency-by-design principle [8], as they seek to add explainability or transparency to already deployed applications by improving or creating new interfaces.

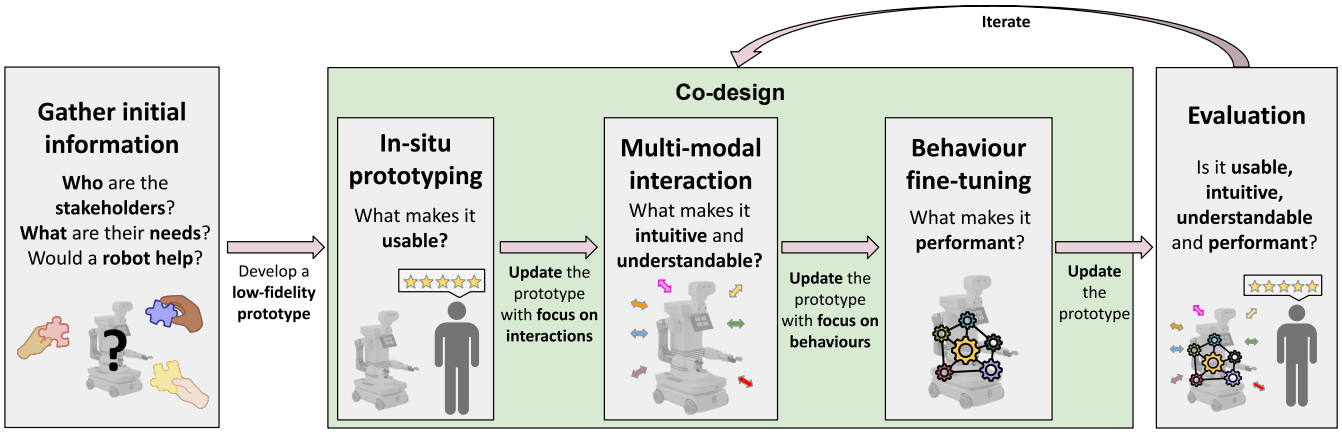

Figure 3: Overall process of the proposed methodology.

3 PROPOSED APPROACH

The goal of social robots is to assist humans. To do so, it is essential that humans are eager to use robots, which will only happen if they trust them. In order to be trusted, robots should be usable: they should perform well in relevant tasks and should come with intuitive interactions, thus allowing users to understand the robot's behaviour. The functionality itself provides performant robots, while intuitively interactive robots demand an interface that exposes the required transparency and explainability. Figure 2 illustrates this chain of implications.

In Figure 3, we define a participatory design process. At the core of this process there are three principal steps; first, we define what makes the system usable, and then co-design a multi-modal interaction that assures that the robot is intuitive and understandable. Only at this point we focus on the robot performance and development of the behaviour.

3.1 Gather initial information

An important fist step is to assess if a robot is indeed a potential solution: maybe it is an overkill, or the technology is not mature enough to meet the general expectations. Actually, if not already available, we recommend to perform a previous exploratory analysis to reveal if a robot is a potential solution before spending too much time on the next steps.

In this first step, it is important to list all the involved stakeholders, meet most of them, and visit the place where the robot would be used. Then, high-level requirements addressing the main user needs should be drafted, to develop a low-fidelity prototype to be teleoperated. This prototype can provide a basic interface towards the users, and should be flexible enough to have engineers teleoperating the robot for a wide range of possible tasks.

3.2 In-situ prototyping co-design

In many systems the interface is only a screen, while for robots the embodiment extends it with conversational abilities, gestures and movements, among others. Therefore, the functionality and interface are intermixed, as the same channel could be used both for a better performance and interaction. Usable robots should take into account this effect and provide a balanced trade-off between performance and understandability: better performance algorithms can sometimes provide lower transparency (and understandability), while in some situations users might prefer systems that barely fail over understanding unexpected behaviours [15, 17].

It is important to be in-situ with the robot in order to identify the tasks that users normally do, and to try to incorporate the robot in as many as possible. It is crucial to mimic real interactions to evaluate if the robot would effectively help and if users would truly use it. On the one hand, it should be verified that interactions are intuitive enough and require an admissible cognitive load from the users side, resulting in a positive balance between the effort to use the robot and its benefit. On the other hand, it should be validated that the robot can effectively perform the tasks.

The in-situ testing should be finished with a refinement workshop to define in detail the tasks that have been identified as more useful. Feedback should be used to develop a refined prototype, which should have a complete interface that is open to changes, and could still have some teleoperated autonomous capabilities.

3.3 Multi-modal interaction co-design

In this step it is important to focus on how the users interact with the robot, with a strong UX perspective. Interactions should be intuitive and fluid, and the users should help to define the interaction flows, devices, and modalities.

In this step, there should be an effort to identify situations where the robot's decision-making and behaviours remain unclear for the users, and therefore there is a need for more explanations or transparency. We propose to have a specific session with the users to treat this issue and find ways to bridge gaps between the user's mental model of the robot and the actual decision-making.

We advocate again for in-situ testing to let the users provide ideas based on real interactions with the robot. This will help to shape the interface according to their needs, and clarify the requirements.

At the end of this step, engineers will have to implement the resulting design. This is the version that will be tested in the next step, incorporating also a prototype of the functionality, which should no longer be teleoperated, thus fully autonomous. The interface

Figure 4: Workshop with the hospital staff.

and understandability requirements will constrain the development of such autonomous behaviours.

3.4 Behaviour fine-tuning co-design

Since users have been long involved in the design process, this step should require less involvement from their side. We recommend again an in-situ testing to engage the users in finding bugs, tuning the desired behaviour, and assessing if the performance meets their expectations.

Focus should be put on the functionality performance, but minor improvements of the interface can also be done. This step should conclude with the engineers refining the whole application with the collected ideas and feedback from all users.

3.5 Evaluation

The last step consists in evaluating the whole system from all the stakeholders perspectives. Evaluating that the system is really intuitive and performant should be done by an extensive on-site deployment and testing. In case that some caveats are found, the interface and functionality can be refined and improved in further iterations.

4 FIRST EXPERIENCE & LESSONS LEARNED

We have started validating the proposed framework in a geriatric unit of an intermediate healthcare centre, where the principal identified task is to provide support to the staff in monitoring hazardous situations for patients. It is an ongoing work, and so far we have executed the steps of requirements gathering and in-situ prototyping co-design. We have teleoperated the robot in the healthcare centre during a whole week, to assess the activities where the robot could support the staff in different shifts (morning, afternoon, night). We tried different tasks such as deliveries, videocalls or patrolling. We wrapped up with a workshop (Figure 4) with some staff members to clarify the functions and ways of interacting they would find more usable.

We next list describe few lessons learnt so far:

- While it is well-known that there is a mismatch between what engineers believe are the end-users needs and what they really need, it is often the case that such mismatch is also present between secondary stakeholders' perspective, such as managers, and the primary users.
- While some stakeholders can provide insight on useful tasks, they are not always aware of the cognitive load that end-users could dedicate to the robot in each situation.
- It is important to try in-situ real situations, as the users themselves might not be 100% aware of their needs. We saw in some cases that what the staff thought would be helpful was not usable in practice.
- It is important to build trust with the users in a participatory design. Only after several days collaborating with the staff would they really open to share all their ideas.

5 CONCLUSION

This paper presents a participatory design methodology aimed at addressing explainability in Human-Robot Interaction (HRI) applications. We begin with the premise that for robots to genuinely benefit humans, they need to be both 'performant' and 'intuitive'. Only then can the adoption of such systems in real-world scenarios be considered. It is in this context that explainability becomes crucial. The proposed approach integrates the dimension of explainability throughout the entire design process, advocating for a comprehensive development of functionality and transparency. This stands in contrast to merely incorporating explainability components into pre-defined systems. The initial steps of this methodology were implemented in an intermediate healthcare center, where we engaged in co-designing a robotic system to provide support to the staff in their daily tasks.

ACKNOWLEDGMENTS

This work has been partially supported by Horizon 2020 grant agreement No 857188 (SAFE-LY - PHARAON), Horizon Europe grant agreement No 101070254 (CoreSense), and Horizon Europe Marie Skłodowska-Curie grant agreement No 101072488 (TRAIL).

REFERENCES

[1] 2022. Standard for Transparency of Autonomous Systems. *IEEE Std 7001-2021* (2022), 1–54. https://doi.org/10.1109/IEEESTD.2022.9726144

[2] June Ahn, Fabio Campos, Ha Nguyen, Maria Hays, and Jan Morrison. 2021. Co-designing for privacy, transparency, and trust in K-12 learning analytics. In *LAK21: 11th international learning analytics and knowledge conference*. 55–65.

[3] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. 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*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.

[4] Minja Axelsson, Raquel Oliveira, Mattia Racca, and Ville Kyrki. 2021. Social robot co-design canvases: A participatory design framework. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 1 (2021), 1–39.

[5] Elin A Björling and Emma Rose. 2019. Participatory research principles in human-centered design: engaging teens in the co-design of a social robot. *Multimodal Technologies and Interaction* 3, 1 (2019), 8.

[6] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*. 211–223.

[7] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamo-Larrieux. 2019. Robots and transparency: The multiple dimensions of transparency in the context of robot technologies. *IEEE Robotics & Automation Magazine* 26, 2 (2019), 71–78.

[8] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. 2020. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics* 26, 6 (2020), 3333–3361.

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

[10] Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier, and Ghassan Hamarneh. 2021. EUCA: The end-user-centered explainable AI framework. *arXiv preprint arXiv:2102.02437* (2021).

[11] Hee Rin Lee, Selma Šabanović, Wan-Ling Chang, Shinichi Nagata, Jennifer Piatt, Casey Bennett, and David Hakken. 2017. Steps toward participatory design of social robots: mutual learning with older adults with depression. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. 244–253.

[12] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–15.

[13] Michal Luria. 2023. Co-Design Perspectives on Algorithm Transparency Reporting: Guidelines and Prototypes. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1076–1087.

[14] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[15] Birthe Nesset, David A Robb, José Lopes, and Helen Hastie. 2021. Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 313–317.

[16] Anastasia K Ostrowski, Cynthia Breazeal, and Hae Won Park. 2021. Long-term co-design guidelines: empowering older adults as co-designers of social robots. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 1165–1172.

[17] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's complicated: The relationship between user trust, model accuracy and explanations in ai. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–33.

[18] Kasper Rodil, Matthias Rehm, and Antonia Lina Krummheuer. 2018. Co-designing social robots with cognitively impaired citizens. In *Proceedings of the 10th Nordic conference on human-computer interaction*. 686–690.

[19] Selma Šabanović. 2010. Robots in society, society in robots: Mutual shaping of society and technology as a framework for social robot design. *International Journal of Social Robotics* 2, 4 (2010), 439–450.

[20] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.

[21] Tjeerd AJ Schoonderwoerd, Wiard Jorritsma, Mark A Neerincx, and Karel Van Den Bosch. 2021. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human-Computer Studies* 154 (2021), 102684.

[22] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

[23] Clay Spinuzzi. 2005. The methodology of participatory design. *Technical communication* 52, 2 (2005), 163–174.

[24] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. 2021. 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*. Springer, 119–138.

[25] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable embodied agents through social cues: a review. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 3 (2021), 1–24.

[26] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE conference on computational intelligence and games (CIG)*. IEEE, 1–8.