Co-designing Explainable Robots: A Participatory Design Approach for HRI

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

***Abstract*— Many research works currently focus on algorithms designed to generate explanations and then evaluate their effect on user trust and understanding of robots. Even though some projects attempt to design understandable interfaces, they usually serve as extra features for solutions that already exist. In this paper, we suggest a user-centric approach to design explainable robot systems from the very beginning. In particular, we provide a participatory design approach that places emphasis on the cooperative design of an understandable and intuitive interface between the user and the robot system. We suggest turning the attention to the robot's functionality and autonomous behaviours development after this interface has been established. We exemplify how to apply the proposed framework in a geriatric unit at an intermediate care centre.**

Fig. 1. Co-design workshop with the users during the proposed framework's implementation in a geriatric unit at an intermediate care centre.

I. INTRODUCTION

Explainability is often seen as a mechanism to make systems more understandable for humans [15], [25]. Many explainability works identify the need to focus more on the users to achieve understandable systems [7], [15], but few address this challenge [3]. Frameworks suggesting that understandable systems should grant an intuitive user experience (UX) have been proposed for general Artificial Intelligence (AI) systems [13], but the UX understandability perspective has not been studied in depth in the Human-Robot Interaction (HRI) field, where interactions are enhanced by the multimodal interface capabilities of robots, e.g. gestures, sound, speech, or displays to name a few.

Participatory Design (PD) approaches are a promising way to strongly 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 can shape robots that precisely adapt to their needs, which will in turn reshape the way users conceive robots, following a mutual shaping process [19]. Furthermore, PD would allow the inclusion of the explainability dimension from the start of the design process following a transparency-by-design approach [8], instead of being an add-on to already deployed systems.

Most works in eXplainable HRI (XHRI) focus on ways to generate explanations [27], and then evaluate their effect on the understanding and trust among users. In this work, we propose to approach XHRI from the opposite direction: through PD, we suggest first to co-design the interface with a strong focus on the user's explainability needs. We contend that the development of the robot's performant behaviour —which ought to offer the identified degrees of transparency and explainability— should only take place once users certify that interactions are intuitive and understandable.

We believe that transparency and explainability are mechanisms that can provide "understandable" robots, that is, robots whose behaviours are predictable by humans interacting with them. We argue that robots need to be "understandable" to be "intuitive", which along with being "performant" and effectively executing user-relevant tasks, will become "usable". Such conditions will drive humans to trust and convincingly use robots, and only then will robots truly assist them. Fig. 2 illustrates this chain of implications, which emphasises the necessity of integrating explainability into the fundamental design of robotic systems to benefit society.

The main contributions of this work are twofold. On the one hand, we present a PD framework for XHRI with a focus on the co-design of understandable robots, which in turn impacts the performance, intuitiveness and usability characteristics of robots. On the other hand, we introduce an example implementation of the framework for a robot in an intermediate care centre geriatric unit.

The paper is organized as follows. Section II provides a literature review. Section III details the proposed methodology. Section IV presents an example of how to apply the methodology in the design of a robot for a geriatric unit at a care centre (Fig. 1). Section V discusses insights from our experience, and Section VI concludes the paper.

II. RELATED WORK

In this section, we first review the terminology used in the eXplainable Artificial Intelligence (XAI) field, and we continue surveying the XHRI sub-field. We next explore lit-

[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

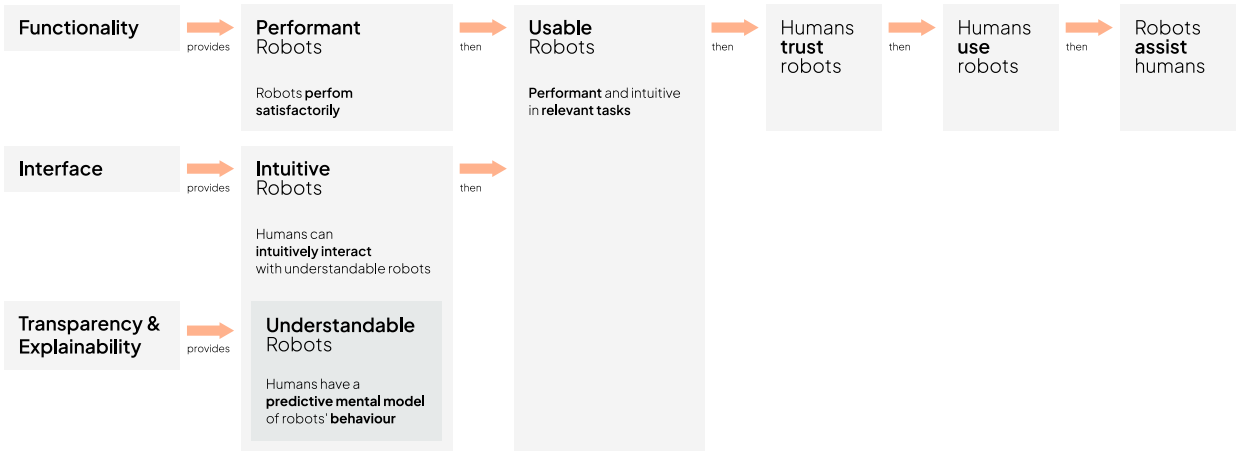

Fig. 2. The relation between usable, performant, intuitive and understandable robots, and the outcomes of humans trusting and using robots.

erature where PD approaches have been applied in robotics, and we conclude by reviewing 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 HRI as proposed in this work.

A. Transparency and Explainability

The field of XAI focuses on making AI systems more explainable. Many XAI works propose definitions of *explainability*, which has related terms such as *transparency*, *understandabilty* or *interpretability*. In [26], the authors analyse the terms used in the XHRI literature, being *explainability* and *transparency* the most common ones, with variant definitions across different publications. In [1] *transparency* is defined as "the transfer of information from an autonomous system [...] in a form meaningful to the stakeholder", while *explainability* is described as *transparency* addressed to non-expert users. In [25] 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 better *understand* a robot, considering that *transparency* discloses information and allows to answer "what" questions about the robot's decisions and behaviour, while *explainability* would further clarify the given information, and allow answering "how" and "why" questions about those decisions and behaviours [25].

B. Explainable Human-Robot Interaction

In the field of HRI, several frameworks have been presented to structure the design of explainable robots. Theory of mind (ToM) based frameworks have been proposed, where the robot performs communicative actions to reduce the mismatch between the robot's state-of-mind and the estimated human's model of the robot's state-of-mind [9]. A survey on explainable autonomous robots [20] uses the ToM approach to group the reviewed papers based on their main contribution: (1) to have an interpretable robot decision-making space, (2) to estimate the plan of the user, (3) to generate an explanation to reduce the mismatch between the robot's and user estimated interpretable plans, and (4) to encode the generated explanation into specific modalities.

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 various metrics [10]. One such example is the literature review in [26], which provides an overview of explainability's effects on trust, interaction robustness, and the mental model of the robot, finding that explainability almost always positively correlates with them.

C. Participatory Design relevance in HRI

Since the operationalisation of most studies covers specific tasks and domains, it has been argued that generalised 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 utilised to design explainable robotic systems while taking into account different user needs.

Participatory Design is a term often interchanged with co-design [21]. In PD, multiple stakeholders, such as end-users and domain experts [4] contribute to the design process as active co-designers. Non-roboticists can actively collaborate in the robot design [12] by playing a critical role [23] to iteratively construct the emerging design [24]. As such, PD approaches are a way to achieve mutual-shaping [19] of the interaction between robots and society.

Concretely, PD has been used in several HRI domains, such as the co-design of social robots to improve the mental health of teens [5] or older adults with depression [12]. These works employ a various set of tools such as card-sorting [17], role-playing [5] and prototyping [5], [17].

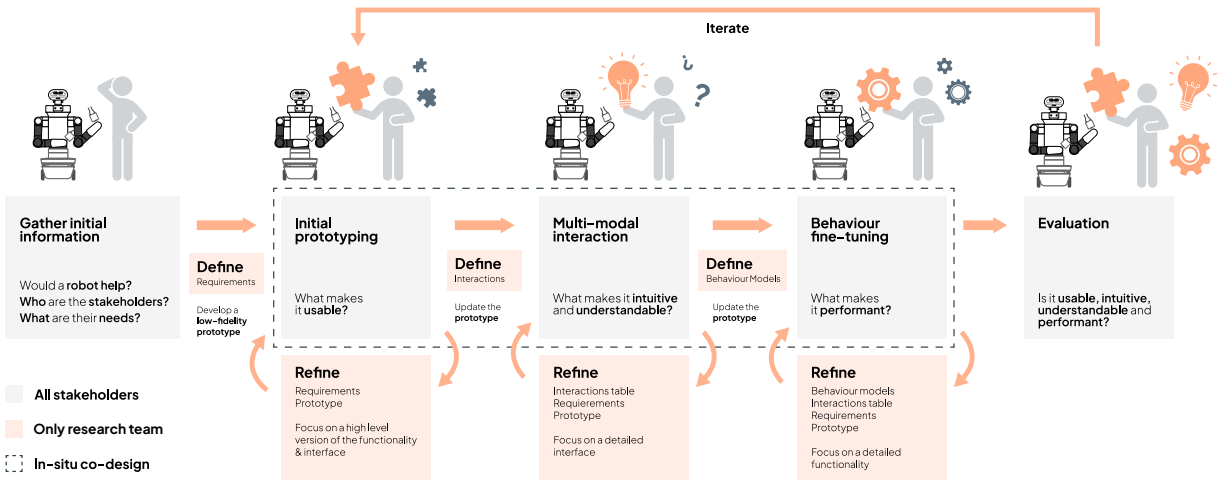

Fig. 3. Proposed participatory design process, where the core step is the in-situ co-design of an understandable multi-modal interaction.

D. Explainability through participatory design

Several works use PD methodologies in XAI. 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 [11] with a 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 [22], or a social media recommender system [14]. 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 existing interfaces.

In the surveyed literature, there is a gap in how PD can be used to provide explainability that improves robot systems understanding in the HRI field. The approach that we present in this work aims to bridge this gap.

III. PROPOSED APPROACH

In this work, we propose a 3-stage PD process (Fig. 3) to design explainable robot systems. First, a *gather initial information* stage is foreseen to identify the stakeholders' needs and evaluate if a robot is a potential solution. This is a full process based on a variety of techniques such as observation studies, interviews, or focus groups, that grant a complete use case understanding. By the end of this stage, a research team should be constituted, which will lead the PD. In this work, we do not delve into this first stage, but rather focus on the *in-situ co-design* stage, which is the core of the explainability design. This section details its 3 sub-steps, namely *initial prototyping*, *multi-modal interaction* and *behaviour fine-tuning*. The last stage is the system *evaluation* from all stakeholders' perspectives performed by people who have not been part of the process. It can be based on available XAI metrics for user satisfaction, mental models, curiosity or trust [10], and should be adapted to each particular use case objective.

A. Initial prototyping co-design

After the *gather initial information* stage, high-level requirements addressing the stakeholder needs should be drafted, with emphasis on end-users. We recommend splitting requirements between functional and interface requirements, which refer to the robot's behaviour and interactions respectively. Interface requirements might already incorporate explainability measures to make interactions understandable and intuitive. However, their specific definition will be later refined to widely cover the understandability perspective. Requirements should guide the development of a low-fidelity prototype.

We propose deploying the low-fidelity prototype and in-situ testing several potential tasks, that is, to simulate the complete interaction flows and robot behaviours in activities where the robot could assist the users. This way, a first rough idea of which technologically feasible tasks are user-relevant can be identified, as well as which interface modalities are more appropriate.

During this step, feedback should be used to refine the requirements and the prototype. We strongly recommend concluding the in-situ testing with a workshop where stakeholders clarify the design choices for functionalities and interfaces. At the end of this step, requirements should describe a usable robot, that is, a robot that is performant and intuitive in relevant tasks for the stakeholders (Fig. 2).

B. Multi-modal interaction co-design

This step focuses on how stakeholders, primarily the users, interact with the robot. We propose in-situ iterating the interface while keeping a low-fidelity prototype for the behaviour, which if needed can be teleoperated by the research team.

Situations where the robot's interface and behaviour are unclear should be identified in the *interaction table* (Fig. 4), which the research team should fill taking into account all the interaction modalities and later refine with the received feedback. This table structures the co-design of intuitive

Stakeholder	Interaction situation	Interaction situation probability	Interaction issues	Interaction issues severity	Interaction critical level	Explainability measures
type_1	*Description of the situation in which the stakeholder is interacting with the system*	*Value between 0 and 5 that indicates how often the situation is likely to happen*	*Questions from the stakeholder perspective about something not being understandable or intuitive e.g. "Why is the robot doing that? How can I command something?"*	*Value between 0 and 5 that indicates how severe is the fact that stakeholders do not understand or find intuitive the interaction situation*	*Value between 0 and 5, which is the mean of the situation probability and interaction issues severity*	Previous information: *Measures that provide information BEFORE the usage (e.g. handbook, FAQ, training...)*
						Legibility: *Measures that make the robot behave in a way that is more understandable DURING the usage*
						Post-hoc: *Measures that provide extra information AFTER A REQUEST*
type_1	...	...	...	...	...	...
[...]	*[...]*	*[...]*	*[...]*	*[...]*	*[...]*	*[...]*
type_n	...	...	...	...	...	...

Fig. 4. Interaction table with situations with potential intuitiveness or understandability issues, and explainability measures to mitigate those issues.

and understandable interactions. Moreover, it can help to establish the user-preferred balance for the performance-understandability trade-off, on the one hand, and on the other, it supports integration with the P7001 standard [1]. We next detail these three aspects.

1) The interaction table: It encapsulates potential interaction issues and explainability measures to address them. For each stakeholder defined in the *gather initial information* stage, the research team should add a row for each foreseen interaction, with the following fields:

- *Stakeholder* contains the stakeholder type.
- *Interaction situation* describes the interaction instance, that is, what the users and robotic system are specifically doing.
- *Interaction situation probability* is a value between 0 and 5 representing the probability of an interaction to take place. Higher values indicate more frequent situations.
- *Interaction issues* include aspects that stakeholders may not understand or may not find intuitive. We recommend formulating questions from the stakeholder's viewpoint to bring these ideas into tangible form.
- *Interaction issues severity* is a value between 0 and 5 representing the severity of a problem in an interaction. Higher values indicate a higher degree of non-understanding or non-intuitiveness.
- *Critical level* combines the probability and severity fields. We recommend computing it as the mean of probability and severity, but for certain applications, different functions might be more adequate, such as the maximum value.
- *Explainability measures* are actions that mitigate interaction issues grouped in the following categories: (1) 'previous information' are measures done before the robot's usage (e.g. documentation or training), (2) 'legibility' includes measures that implement understandability during the usage, and (3) 'post-hoc' are measures providing additional clarification responding a stakeholder's request.

We advocate again for in-situ testing to let stakeholders provide feedback on the different taken measures based on real interactions with the low-fidelity robot prototype. The research team should prepare the prototype with the necessary flexible interface capabilities, and then operate the robot to replicate the behaviours defined in the requirements, while simulating the interactions defined in the *interaction table*. Stakeholders should pinpoint interactions lacking intuitiveness, being difficult to understand, or proving overwhelming or redundant. Several iterations should be conducted to refine the *interaction table* and adapt the interaction issues, critical level, and explainability measures according to the received feedback. Moreover, additional unforeseen situations should be included.

In many implementations, the number of interactions will be large, and it might be unfeasible to have time to test them all. The *interaction table* has a critical level to be able to prioritize the in-situ tested interactions. Furthermore, we recommend focusing on the stakeholders that mostly interact with the system, which are the most frequent users.

During this step, the requirements and prototype should be refined to include interface improvements derived from the explainability measures. During the testing, stakeholders will provide design choices that also affect the behaviour, so the functional requirements should be updated accordingly. Because the behaviour is kept as a low-fidelity prototype, it should be fast to iterate on it, allowing to spend more time in updating an interface that by the end of the step should be relatively final.

2) Trade-off between performance and understandability: In many AI systems, the interface is mainly, if not only, a screen. However, robots extend their interfaces to conversational abilities, gestures, sounds and movements among others. The functionality and interaction dimensions are often intermixed, as they can share the same interface channel. Furthermore, there is typically a trade-off between system performance and system understandability: higher-performant algorithms sometimes provide lower explainability capabilities, e.g. an AI neural network vs a decision tree.

Fig. 5. In-situ teleoperating and testing the robot in the geriatric unit. From left to right: researcher teleoperating the robot; nurse interacting with the robot in a delivery task; robot detecting obstacles in its path (common in the given use case); robot delivering an item to a patient in a room; and prototype applications used in different devices to both, interact with and teleoperate the robot.

In these situations, stakeholders might prefer understanding unexpected behaviours over systems that fail less, or the other way around. Nevertheless, it has been shown that when the performance is too low, adding transparency might have a neutral or negative effect on trust [16], [18]. During the definition and validation of explainability mitigation strategies, this trade-off should be carefully discussed with the stakeholders, to ensure they can provide an informed choice. *Interaction table*'s row entries can be split to account for variant implementations. One variant might have better performance (i.e., a lower probability of having interaction issues), but fewer explainability options (i.e., fewer capabilities to mitigate a higher severity), while another variant would provide the opposite trade-off.

3) Integration with the IEEE P7001 standard: Our approach has been designed to fit under the IEEE Standard for Transparency of Autonomous Systems P7001 [1]. We propose the *interaction table* as a methodology to perform a P7001 System Transparency Specification (STS). According to P7001, "an STS is the process of defining the transparency requirements of an autonomous system, for each stakeholder group. An STS may be written at any time during a system's lifecycle, though the best and expected practice would be to specify transparency requirements prior to system design". On the one hand, we suggest mapping each stakeholder in the *interaction table* into a stakeholder category from P7001, such as *general public*, *users*, or the *end-users*, *domain-experts* and *superusers* subcategories. On the other hand, we propose to utilise each stakeholder category's mean critical level in the *interaction table* to determine P7001's transparency level.

C. Behaviour fine-tuning co-design

Before starting the *behaviour fine-tuning* step (the last step in the in-situ co-design stage), the research team should identify the user-relevant behaviour model variants and update the prototype to implement them. The prototype should also implement the automatic generation of the explainability measures from the previous step. Then, we suggest fine-tuning the behaviour models through an in-situ co-design with relevant stakeholders and assessing if its performance meets their expectations. In this stage, the focus should be directed towards the functionality performance. Nevertheless, the feedback might involve design changes in the interface and the *interaction table*. These should be updated and addressed accordingly.

After the in-situ iterative co-design, engineers should update the prototype into a final version ready for evaluation by users not involved in the co-design. In case results do not meet expectations, the interface and functionality can be refined in further iterations, where the requirements, *interaction table*, and behaviour models will be updated.

IV. EXAMPLE USE CASE

We next describe how we have applied the proposed framework in the SAFE-LY[1] project. In this paper, we only focus on the *multi-modal interaction* step (Fig. 3), which is the core contribution of this work, as it contains the methodology to include the explainability perspective into a PD process. The use case that we have selected is a robot that assists the staff in a geriatric unit of an intermediate healthcare centre (Fig. 5).

After running the *gather initial information* stage, where we discussed with the management team, head nurse, a geriatrician and the nursing staff, we proceeded to the *initial prototyping* step. We developed a low-fidelity prototype system composed of a robot and a mobile application. We teleoperated the low-fidelity prototype 4 hours a day for 7 days in 3 shifts (morning, afternoon, and night) to assess in which activities the robot could support the nursing staff. We tested 4 tasks: deliveries (between staff and patients, and between staff in the same or different units), staff-patient video calls, remote inspection of rooms, and autonomous patrolling to detect hazardous situations for the patients.

After the teleoperation period, we run a wrap-up session with the main stakeholders. We prepared a role-playing workshop to reproduce situations that we identified with a higher potential to assist the staff, given the insights from the in-situ testing. Then, we prepared a set of structured questions to discuss with them a first detailed version of the robot's interaction flows and behaviour design. It was decided that the patrolling task had a higher potential usage, and requirements were refined to include all the received feedback. It was agreed that the robot would continuously

[1]https://pal-robotics.com/collaborative-projects/safely/

Stakeholder	Interaction situation	Interaction situation probability	Interaction issues	Interaction issues severity	Interaction critical level	Explainability measures
nursing staff	*The robot triggers a "person fallen" alarm, and there is a person but seated (in a chair/ wheelchair).*	*3*	*What did the robot detect? Why did the robot think it was on the floor?*	*3*	*3*	Previous information: Documentation about detection capabilities, limitations, corner cases... Small training to set expectations on the system's performance
						Legibility: *Insistent sound in their phones. Once unlocked, display the robot-captured image together with the relevant region of the image, the reason for triggering the alarm (person on floor), and a simplified confidence level in a colour scale*
						Post-hoc: *Not required*

Fig. 6. Example row entry of the interaction table.

monitor the rooms configured by the staff, and it would trigger alarms after detecting a fallen patient on the floor, which the staff would receive and manage in a phone app. The robot would also trigger alarms for not-in-bed patients and closed room doors for a group of vulnerable patients.

A. Specifying the interaction table

Based on the observations and insights gathered through the *initial prototyping* step, we introduced 37 entries in the *interaction table* of the *multi-modal interaction* step (Fig. 3) for the nursing staff stakeholder type. Although other stakeholders such as the patients should be included in the table as well, in this example we focus on the main users.

In Fig. 6 we provide an entry example. It comprises interactions where the robot triggers a false alarm after the perception module fails to detect that a person has fallen on the floor, when the person is sitting on a chair instead. We only defined the fields 'previous information' and 'legibility' from the explainability measures, as we considered that in this case, the nursing staff would not be further concerned with 'post-hoc' explanations. For this same situation, other stakeholders might have different needs, so a row should be included from each stakeholder's perspective if needed.

B. Refining the interaction table

After filling the *interaction table* and selecting the most critical situations, the prototype interface was updated to test those interactions along with the mitigation strategies. The robot's behaviour remained teleoperated by engineers who simulated the autonomous functionalities. We tested the prototype with staff members of different ages, nursing seniority, and levels of previous explanations. Because of the in-situ testing, it was manageable for the users to comprehend the system's limitations, and assess how critical non-understandable situations were.

We confirmed that receiving in the phone an image with information on what had been detected helped to understand the robot's behaviour, and allowed them to identify the actual criticality of the situation to take action or dismiss the alarm. They also stressed the importance of clearly recognising the location where the alarm was triggered, which they suggested displaying in phone notifications, in addition to keeping the robot standing in front of that area with flashing LEDs. Finally, they also reinforced the need to use multi-modal signals and to employ different sounds and frequencies to effectively assess the distinct alerting situations. This way, it would be more intuitive to know what had been detected, even without the need to unlock the phone.

After iterating on the interface while implementing the modified explainability measures, we could define a final interaction flow that they approved to be intuitive and understandable. The next step, the *behaviour fine-tuning*, would continue from this prototype version. Because this contribution focuses on the explainability perspective, we do not include it in this section and it will be reported elsewhere.

V. DISCUSSION

We next present insights gained during the example use case implementation. The nursing staff are normally busy with daily tasks, so it was especially useful to sort the *interaction table* (Fig. 4) by critical level to prioritise the most critical interactions. It was also valuable to start focusing on the primary user stakeholder type. Given the iterative nature of the PD process, it is feasible to start focusing on the principal stakeholders, and gradually incorporate secondary stakeholders in the next iterations.

Stakeholders might have different interests and potential discrepancies concerning design choices, even within the same stakeholder type. In our use case, we noticed that regarding the performance-understandability trade-off, the management stakeholders were generally interested in having post-hoc explainability insights (e.g., to be able to track back incidents), while the nursing staff preferred higher performance. As part of the research team, we limited ourselves to informing about technological limitations, and we let the other stakeholders make the decisions based on their existing organizational models and hierarchies.

We confirmed the well-known mismatch between what engineers believe the users' needs are and what they actually need. Furthermore, we experienced that this mismatch is also present between secondary stakeholders, such as managers, and primary users. Secondary stakeholders can sometimes provide useful insights, but they are not always aware of the users' cognitive load and understandability needs. Even the users themselves might not be fully aware of them. We observed that what the staff thought would be helpful

was sometimes not usable after in-situ testing. For example, the staff realised only after experimenting with the robot's delivery of items that most patients could not reach the robot to pick up items, and that it took more time to specify the delivery locations and wait for the robot than actually bringing the items themselves.

Moreover, we also encountered that including a UX perspective in the co-design was necessary to improve the intuitiveness of the interactions, identifying modifications such as adjusting the app menu hierarchy to reduce the number of actions they needed to configure the patrolling.

In general, we experienced that building trust with the stakeholders was very important. Only after several days of collaboration would they open and share all their ideas. We corroborated that in-situ iterating the design allowed them to provide fluid, informal, and specific feedback about real situations, which was essential for designing a robot system that fits their needs, especially in terms of understanding.

VI. CONCLUSIONS

This paper presents a participatory design methodology to address explainability in HRI applications. We advocate that robots must be performant and intuitive in human-relevant tasks to truly assist people. Explainability becomes essential in this setting as a mechanism to achieve understandable and intuitive robots. The suggested method incorporates the explainability component into the co-design process, assisting in the identification of non-understandable circumstances and the testing of possible solutions in the multi-modal interaction with the system. This stands in contrast to merely adding explainability components into pre-existing systems. An example of the methodology's application was carried out in an intermediate healthcare facility, where we co-designed a robotic system to assist the staff in daily care routines.

ACKNOWLEDGMENT

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

ETHICAL APPROVAL

The Research Committee of the BSA hospital declared that this work does not require ethical approval. No user data has been gathered during this work.

REFERENCES

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

[2] J. Ahn, F. Campos, H. Nguyen, M. Hays, and J. Morrison. Co-designing for privacy, transparency, and trust in K-12 learning analytics. In *International Learning Analytics and Knowledge Conference*, pages 55–65, 2021.

[3] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling. Explainable agents and robots: Results from a systematic literature review. In *Int. Conf. on Autonomous Agents and Multiagent Systems*, pages 1078–1088, 2019.

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

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

[6] M. Eiband, H. Schneider, M. Bilandzic, J. Fazekas-Con, M. Haug, and H. Hussmann. Bringing transparency design into practice. In *Int. Conf on intelligent user interfaces*, pages 211–223, 2018.

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

[8] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, 26(6):3333–3361, 2020.

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

[10] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers in Computer Science*, 5:1096257, 2023.

[11] W. Jin, J. Fan, D. Gromala, P. Pasquier, and G. Hamarneh. Euca: The end-user-centered explainable ai framework. *arXiv preprint arXiv:2102.02437*, 2021.

[12] H. R. Lee, S. Šabanović, W.-L. Chang, S. Nagata, J. Piatt, C. Bennett, and D. Hakken. Steps toward participatory design of social robots: mutual learning with older adults with depression. In *Int. Conf. on human-robot interaction*, pages 244–253, 2017.

[13] Q. V. Liao, D. Gruen, and S. Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Conf. on Human factors in computing systems*, pages 1–15, 2020.

[14] M. Luria. Co-design perspectives on algorithm transparency reporting: Guidelines and prototypes. In *Conf. on Fairness, Accountability, and Transparency*, pages 1076–1087, 2023.

[15] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[16] B. Nesset, D. A. Robb, J. Lopes, and H. Hastie. Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of Int. Conf. on Human-Robot Interaction*, pages 313–317, 2021.

[17] A. K. Ostrowski, C. Breazeal, and H. W. Park. Long-term co-design guidelines: empowering older adults as co-designers of social robots. In *Int. Conf. on Robot & Human Interactive Communication*, pages 1165–1172, 2021.

[18] A. Papenmeier, D. Kern, G. Englebienne, and C. Seifert. It's complicated: The relationship between user trust, model accuracy and explanations in ai. *Transactions on Computer-Human Interaction*, 29(4):1–33, 2022.

[19] S. Šabanović. 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):439–450, 2010.

[20] T. Sakai and T. Nagai. Explainable autonomous robots: A survey and perspective. *Advanced Robotics*, 36(5-6):219–238, 2022.

[21] E. B.-N. Sanders and P. J. Stappers. Co-creation and the new landscapes of design. *Co-design*, 4(1):5–18, 2008.

[22] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. Van Den Bosch. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *Int. J of Human-Computer Studies*, 154:102684, 2021.

[23] D. Schuler and A. Namioka. *Participatory design: Principles and practices*. CRC Press, 1993.

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

[25] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman. A two-dimensional explanation framework to classify ai as incomprehensible, interpretable, or understandable. In *WS on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pages 119–138, 2021.

[26] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, and M. Chetouani. Explainable embodied agents through social cues: a review. *Transactions on Human-Robot Interaction*, 10(3):1–24, 2021.

[27] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood. Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. In *IEEE conference on computational intelligence and games*, pages 1–8, 2018.