Adapting explanations' level of detail in a longitudinal in-the-wild office delivery robot: Ongoing results

Ferran Gebelli
ferran.gebelli@pal-robotics.com
PAL Robotics
Barcelona, Spain

Anaís Garrell
anais.garrell@upc.edu
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
Barcelona, Spain

Séverin Lemaignan
severin.lemaignan@iiia.csic.es
Artificial Intelligence Research Institute (IIIA-CSIC)
Barcelona, Spain

Raquel Ros
raquel.ros@iiia.csic.es
Artificial Intelligence Research Institute (IIIA-CSIC)
Barcelona, Spain

Abstract

We present an ongoing longitudinal in-the-wild study of an office delivery robot that adapts the level of detail of its explanations to users of failure or unexpected events. We compare three explanation strategies: minimal "what happened" explanations, fully detailed including "what + why" reasons, and a personalised variant that tailors detail based on tracked user knowledge. Additionally, some users can request on-demand extra follow-up explanations. Ongoing results from the first two weeks indicate that personalised explanations preserve both subjective and objective understanding while providing a level of detail closer to correct when compared to fully detailed explanations. Moreover, users who receive personalised explanations request fewer extra follow-up explanations compared to the group receiving minimal explanations. Full study results will confirm these results and provide their evolution through the next 2 weeks of deployment.

CCS Concepts

• **Computing methodologies** → **Cognitive robotics**; • **Human-centered computing** → **Empirical studies in interaction design**.

Keywords

Explainable Human-Robot Interaction, Longitudinal in-the-wild studies

1 Introduction

In Human–Robot Interaction (HRI), explainability is broadly acknowledged as a crucial element for improving human comprehension of robotic behaviour and decision-making processes [47, 69]. From a *Theory of Mind* (ToM) standpoint, the purpose of explainability is to foster more accurate human mental models of the robot [24, 66]. Since robots must infer separate models for each user, personalisation becomes an essential component [59]. Nonetheless, only a small number of works have directly investigated personalised explanations in HRI [4].

Accounting for the dynamic nature of a user's knowledge presents significant challenges in integrating direct and indirect estimation methods. However, user models may be iteratively refined by leveraging interaction data to tailor several dimensions of communication, such as the level of detail, technical nomenclature, and content structure [11]. While detailed explanations can enhance understanding, excessive information may require more time and attention needed to make sense of the explanation [39] or overwhelm the users [31, 49]; thus, explanations must remain concise and appropriate [48], since simpler explanations tend to be more comprehensible [30].

In this work, we focus on the personalisation of explanations by adapting the level of detail to individual users. While prior work has explored this aspect in non-embodied AI systems [32, 38, 39], its application to robotics remains underexplored. Robots present unique challenges due to their physical and social presence: users tend to anthropomorphise robots, attributing intentions and agency, which elevates expectations for explanations of physical movements, failures, and norm violations [10, 67]. Furthermore, different explanation types and modalities (e.g., counterfactuals, verbosity, non-verbal cues) can significantly influence user attributions of robot behavior [27].

Moreover, adapting explanation detail is particularly critical in longitudinal in-the-wild interactions, where repeating already-mastered information proves inefficient in an ecologically valid task. Despite the prevalence of long-term robot deployments in real-world settings, longitudinal, in-the-wild studies in HRI are scarce [18], including the ones adapting the level of detail of explanations. Our prior work [17] introduced an algorithm that dynamically adjusts explanation detail based on tracked user knowledge while accounting for forgetting, and this work evaluates the effects of such personalised explanations on end-users.

The main contributions can be summarised as follows:

- We perform a longitudinal and in-the-wild study of a robot performing internal deliveries within an office for two weeks. We present ongoing results, as the robot will continue running for two more weeks.
- We compare the effects of minimally detailed, fully detailed and adaptive explanations, with results pointing to a benefit of explanations that adapt the level of detail.

This work is organised as follows: In Sec. 2, we present the related work. In Sec. 3 we detail the study design. In Sec. 4, we present and discuss the preliminary results. Sec. 5 concludes the paper and outlines future work.

2 Related Work

This section reviews the need for explanations in HRI, the contextual factors to adapt these explanations, how such adaptation affects users and the methods to generate personalised explanations.

2.1 Need for explanations in HRI

In HRI, explanations and other transparency mechanisms help users, especially non-experts, to understand a robot's internal logic and support basic recovery in failure cases [13, 21, 70, 76]. By improving users' mental models, explanations can foster trust, usability, task performance, and perceived accountability [25, 35, 64], but their usefulness depends on context and user characteristics.

Explanations are most valuable in situations with negative consequences, such as unforeseen events, inabilities, errors, or social norm violations [14, 73]. A recent taxonomy distinguishes user-centred needs (e.g., incomplete mental models or social concerns) from robot-centred ones (e.g., inabilities, errors, unforeseen situations [71]), and links explanation seeking to user mindsets such as utilitarian, interrogative, or critical [9]. Several approaches estimate when explanations are needed, including Large Language Models (LLMs) [72], non-verbal cues [29], eye-tracking and performance-based prediction [74], and in-situ co-design with end users [19].

At the same time, more explanations are not always better. People often have only coarse intuitions about complex systems [26], and explanation preference varies with personality and technological comfort [54] as well as cognitive style, with low need for cognition users benefiting the most and high need for cognition users sometimes losing confidence [46]. Transparency can also pose problems by exposing algorithmic errors [65]. Progressive and interactive disclosure approaches, which provide hierarchically organised, on-demand information, mitigate these issues and can improve perceived usefulness, task performance, and trust, especially when explanations are timely and proactively provided [6, 36, 65].

2.2 Context criteria to adapt explanations

Many works have employed the user role as the primary criterion to adapt explanations, distinguishing experts from non-experts' information needs [53, 56]. LLM-based systems can explicitly query expertise and tune technical depth [45], and explanation preferences have been linked to demographic and psychological factors rather than app-specific knowledge [52]. Related work further connects preferred detail levels to characteristics such as technical expertise, innovativeness, need for cognition, and problem-solving style [1, 12], and shows that user type moderates the effect "why" versus "what" explanations [15, 63].

Other studies de-emphasise user profiles and instead focus on situational factors [51]. For instance, some adapt detail to the type of explained information (input, process, output) [23], or to predicted confusion, mental state, and familiarity with the issue [50].

A third line of work highlights user preferences. Allowing users to choose can better accommodate diverse goals and improve transparency, trust, and satisfaction [22]. Preferences may concern not only the level of detail but also modality, vocabulary, and explanation goals [7, 61]. Since preferred explanations do not always maximise performance, some approaches combine user preferences with workload and performance estimates [68].

2.3 Effects on users of personalised explanations

Many works have explored the effect of changes in the level of detail on users of a system. For non-robotic systems, it has been suggested that good explanations must be both useful and understandable for a given user [38]; too much detail can reduce trust or demand excessive cognitive effort [28, 39], while oversimplification can also increase mental demand and erode trust when understanding is very low [33]. Some forms of detail are perceived as more transparent and satisfying despite higher effort [20], and varying initial information levels do not necessarily increase cognitive overload [32]. Personalisation may also incur a "predictability cost", as highly adaptive systems become harder to learn [60].

Similar patterns appear in HRI works. Increasing explanation detail can improve perceived and objective understanding, but users may still regard explanations as overly detailed [48, 49]. Detailed explanations can raise confidence in navigation and perceived understanding of behaviour and decisions, whereas simpler ones are easier to grasp [30]. Context and history matter: when robots fail (e.g., cannot find an object), non-experts need sufficient but not overwhelming detail to diagnose and resolve problems [13]. A longitudinal work shows that the effectiveness of explanation level depends on failure type and prior exposure: explanation strategies with fixed or decaying levels of detail across rounds affect novices' satisfaction differently, with satisfaction shaped jointly by current and past explanations [27]. Other studies separate what- and why-information [34] and compare how-, why-, and what-if-explanations for failure detection, finding that why-explanations are often preferred yet not always the most understandable [29].

Although ecologically valid longitudinal studies are needed to move beyond novelty effects [5], most prior works on explainability in HRI remain online or lab-based and not longitudinal [18], leaving open questions about how personalised explanation detail plays out in everyday HRI, which we address in this work.

2.4 Generation of personalised explanations

Despite substantial work on the effects of personalised explanations on users, fewer contributions address the issue of how to generate them automatically.

Generally in HRI, personalisation has mostly targeted robot decision-making and behaviour via user models [2, 57], including recent LLM-based approaches [78], while explanation personalisation remains comparatively understudied [4]. LLMs enable personalised dialogue through prompting [78], retrieval-augmented generation over external knowledge [55, 58], fine-tuning [77], and memory-based user modelling [43]. Existing LLM-based explanation systems for robots use templates [37], RAG over robot logs [62], or episodic memory [40, 75], but typically do not adapt to prior user knowledge. Our recent work [17] proposes a method that stores previously explained concepts for each user in an embedding space and uses a heuristic to drive an LLM prompt that adapts the level of detail to this tracked knowledge. This personalisation generation algorithm is employed in the present study to adapt the level of detail of explanations generated with the hierarchical generation framework HEXAR [41] that provides specialised explainers tailored to specific robot software components.

Figure 1: Delivery robot within the office

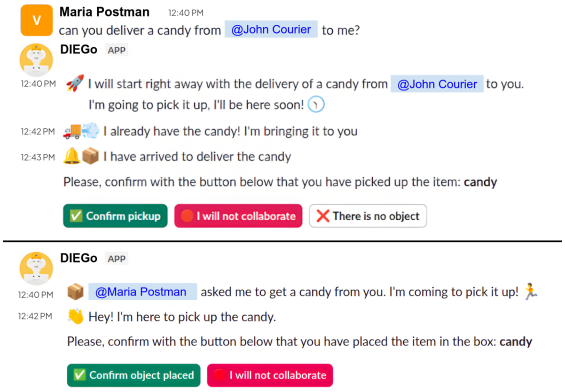

Figure 2: Strip of conversations for a user that requested a delivery (top) and the sender (bottom).

3 Study Design

3.1 Environment and task

The robot operates in a large, open space accommodating around 100 employees (Fig. 1). This environment is particularly relevant, as a recent review highlights a lack of longitudinal research in workplace settings, especially offices [44]. The company engineers and manufactures robots, and the employees in the office are from all departments, such as human resources, business, project management, field technicians, IT, and engineering.

The DIEGo (Deliver Internal Employee Goods) robot's primary function is to deliver items among employees, who can communicate with the robot via the messaging app "Slack", which they already internally use. DIEGo is an additional profile with which users can engage privately in natural language, utilising an LLM that detects when they wish to schedule, delete, or inspect deliveries, or to request explanations. The app features buttons that allow users to confirm or decline the placement or pickup of deliveries.

The robot emits sounds upon arrival and when its path is obstructed, but it does not engage in verbal communication. Users can initiate three types of deliveries: requesting the robot to come to them to pick up an item for someone else, asking the robot to go to another person to retrieve an item for them, or requesting a delivery between two other users. The robot consistently informs both the sender and receiver regarding its approach and arrival through the messaging app. Fig. 2 illustrates a conversation flow between two users during a successful delivery. The top strip shows the receiver's (and requester's) perspective, while the bottom strip shows the sender's perspective.

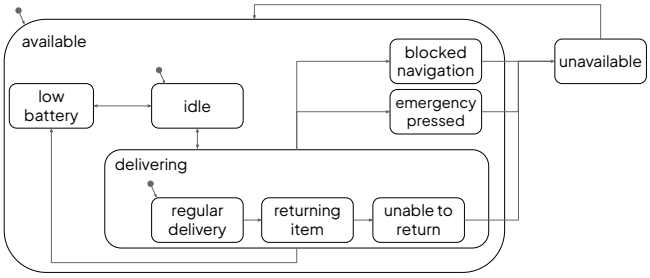

Figure 3: Simplified state machine of the delivery robot.

In the event that the queue is empty, the robot will autonomously proceed to its docking station to charge while awaiting new delivery requests. Conversely, it will sequentially execute deliveries when requests are present. Should the robot's battery level fall below a critical threshold, it will also navigate to the charging station and defer further requests until an adequate battery margin is restored.

The queue of tasks has priorities based on historical usage. This approach ensures that novel or less engaged users do not find themselves relegated to the end of the queue, thereby mitigating potential demotivation. Users have the capability to view the queue and can delete future deliveries they have requested, as well as cancel ongoing deliveries that they initiated. In instances where the robot is unable to pick up an object, it will proceed to the next scheduled deliveries. However, if the object has already been collected and cannot be delivered, it will be returned to the sender.

Some situations require resolution by the research team, who will be present throughout the deployment. The first of these situations is triggered when the red emergency button is pressed, allowing personnel to halt the robot immediately in the event of a potential collision, cancelling the current delivery and resuming operations only after the emergency mode has been reset. If the robot encounters navigation difficulties over multiple attempts, it will classify the current delivery as failed. In many instances, the robot will be able to proceed to the next deliveries, as it might not be entirely obstructed but rather unable to access a specific corridor. However, if the robot consistently fails to reach various destinations, it will designate itself as blocked until assistance from the research team is provided. Finally, if the robot attempts to return an item that could not be delivered, but the original sender does not confirm receipt, the research team will be notified by the robot to effectively resolve the return of the item.

If the aforementioned situations (emergency button pressed, navigation blocked, unable to return item) remain unresolved by the research team within a specified timeframe, the robot will transition into an unavailable state. In this state, the robot will cease to accept new requests while retaining the existing queue in anticipation of potential intervention by the research team.

Requests are only accepted between 9:30 AM and 5:00 PM (up to 14:00 PM on Fridays). Before being turned off at the end of the day, a message is sent to all requesters of pending deliveries, informing them that the robot will be shut down and that the queue will be cleared, as it may no longer be relevant for the following day.

A summary overview of all the robot states can be seen in Fig. 3.

3.2 Explanation generation

The robot proactively provides explanations whenever it identifies a failure or unexpected situation. A comprehensive list of these supported situations is presented in Tab. 1, which includes a summarised version of the explanations detailing the "what" occurred and the "why" it transpired, a classification inspired by previous studies on explanation's level of detail [15, 34].

Explanations are generated automatically in real-time through the HEXAR hierarchical approach developed by the authors [41], which incorporates specialised component explainers aligned with each of the "what" categories from Tab. 1 that leverage LLMs or template-based explanations from the robot's logs. The robot employs the same explanation generation system to respond to specific questions in natural language.

For some users, the robot supports personalisation of the level of detail [17]. This algorithm takes an explanation in natural language, calculates its embeddings, and performs a thresholded retrieval in a per-user database that contains embeddings of previously received explanations. Depending on the relatedness of the retrieved knowledge, an heuristic guides an LLM to reduce the detail level.

To ensure experimental control, two specific failures (starting later due to low battery and cannot complete due to blocked path) will be deliberately triggered up to two times throughout the experiment for each of the participants. This guarantees that participants receive the corresponding explanations, ensuring that certain failures are presented. A target of 0, 1, 2, and 3 occurrences for each of these failure types is established for each participant for each week, respectively, facilitating their distribution across the study. If these failures occur for real, the forced triggering is skipped.

3.3 Apparatus and technical implementation

The robot (as seen in Fig. 1) is a TIAGo from PAL Robotics running ROS2 Humble middleware on an Intel Core i7-10700. It uses the nav2 framework [42] to navigate autonomously using two 2D LIDARs and a depth camera mounted on the head. The robot has a box on its tray where the participants can put the items to be delivered.

The robot operates through a "slack_bolt" backend that processes user messages utilising an LLM that generates structured outputs comprising two main components: user intent and a verbal response. These components are mutually exclusive. The intent is populated only when the user requests actions such as creating or cancelling a delivery, or viewing the schedule. This intent is then transmitted via ROS to the mission controller, which controls the schedule, updates the state machines and triggers the necessary skills (e.g. navigation, pickup, dropoff, and explain), which will render the user message updates through the delivery process. Conversely, if the user does not make a specific request or provide incomplete details, the LLM will populate the verbal response field to maintain a natural conversational flow.

3.4 Participants

A total of N=102 participants are participating in the study (76 males, 26 females). All the employees of the company who work in the area reachable by the robot have agreed to participate. Not all of them are always in the office due to hybrid schedules and holidays, with a mean office attendance of approximately 70%.

What	Why
Invalid request	Pickup and dropoff users are the same
	Pickup and dropoff users are very close
	Pickup or dropoff user not indicated with @tag
	Outside of operating hours.
Unavailable robot	The robot waited for assistance for too long
Deleting all the schedule	The robot is going to be turned off for today
Starting later a request	There are other scheduled deliveries
	The robot has a low battery and needs to charge
	The emergency button was pressed
	There is an unresolved unreturned item
	The robot movements are stuck
Delivery further delayed	The robot has a low battery and needs to charge
	The robot is waiting for assistance to continue
Cannot cancel	There are no active deliveries to cancel
	More than 1 cancellable delivery, which one?
	Task is already being cancelled
	Returning item after an unsuccessful delivery
	Could not return item
Delivery could not be completed	Robot path was blocked
	User declined to place
	User did not confirm placement
	User declined to pick
	User did not confirm picking
	The object was missing at dropoff
	User requested cancellation
	Emergency button pressed
Return could not be completed	Robot path was blocked
	User declined to pick
	User did not confirm picking
	The object was missing at return
	Emergency button pressed

Table 1: Overview of failure types explainable by the robot.

Although employees are accustomed to seeing robots, only a subset has expertise in robotics, and many have never interacted with an autonomous robot. Therefore, the participants have also been segmented on whether they are roboticists or not, that is, if they have technical backgrounds in robots' application layers (e.g. excluding non-engineers, but also firmware or electronics engineers). A total of 53 participants have been considered as roboticists, and 49 as non-roboticists. A total of 93 participants have interacted at least once with the system at the end of the second week.

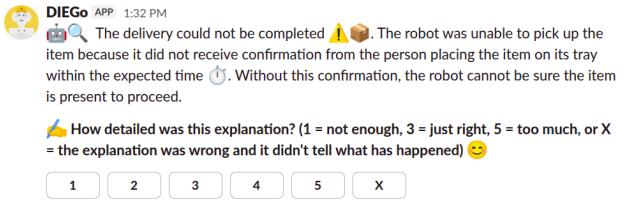

Figure 4: Example explanation with the level of detail rating.

3.5 Procedure

The study is initiated by a formal recruitment announcement and the distribution of a conceptual infographic to participants. To ensure experimental control, a researcher provides a standardised briefing regarding the study's scope; however, the provided information is intentionally limited to prevent disclosing specific functionalities that the robot should communicate autonomously. Participants are then required to complete a consent process.[1]

Throughout the deployment, the system captures ratings in the messaging app regarding the preferred level of detail in the robot's communications. These continuous data points are supplemented by three primary evaluation points: at the conclusion of the first week, the beginning of the third week, and the beginning of the fifth week, with the last two still pending. These longitudinal questionnaires are designed to measure both objective and subjective understandability of the system. The final questionnaire will also include a usability questionnaire [8].

3.6 Independent variables

This study follows a between-subjects design with 2 main criteria distributed evenly throughout the background *roboticist* variable.

Explanations level of detail. Participants in the *WHAT* group receive only the minimal "what" explanations from Tab. 1. Participants in the *WHY* group receive both the "what" and the "why" components. Participants in the *PER* group receive the "what" and the "why" components, but processed by a personalisation algorithm from the authors [17] that takes into account the user's previous knowledge. Notably, due to space limitations, Tab. 1 contains only a summary of the situations, whereas the actual "what" and "why" explanations are further elaborated (example in Fig. 4).

Possibility to request extra explanations. The participants in the *EXTRA* group can see a button right after the explanations named "I want to know more". Pressing this button presents the full *WHY* explanation to participants in the *WHAT* and *PER* groups, whereas for participants in the *WHY* group, the robot replies that no additional information can be provided. The use of this button is entirely optional. Moreover, for users in the *EXTRA* group, the robot can respond to any new or follow-up questions, while for users in the *NO_EXTRA* group, no such button is shown, and all explanation-related questions are answered with a message informing that the system cannot respond to them. The *NO_EXTRA* group aims to avoid providing additional information, especially to the *WHAT* and *PER* groups, while the *EXTRA* group enables results on the proportion of explanations with requested further explanations.

[1]Ethical approval 210/2025 from CSIC Ethical Committee

The robot's overall functioning is a mystery to me (R)
It is hard to make sense of the robot's general functioning (R)
It is difficult to get a clear picture of the robot's overall operations (R)
I am confused about the robot's general objectives (R)
I am unsure what the robot does (R)
I cannot comprehend the robot's inner processes (R)
I cannot explain the robot's behavior (R)
It is impossible to know what the robot does (R)
It is clear to me what the robot does
I have a clear understanding of how the robot operates in general

Table 2: Subjective understandability statements for a 7 Likert, reversed from TOROS [3] illegibility sub-scale.

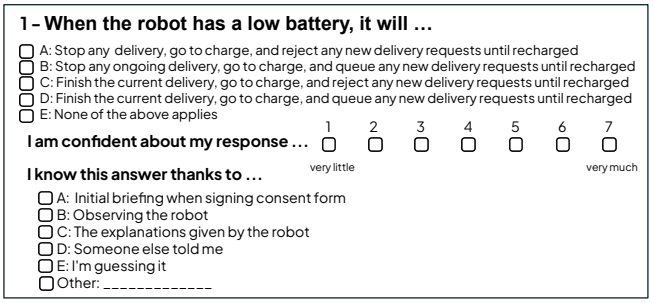

Figure 5: One of the 3 objective understandability questions.

3.7 Dependent variables

The following measurements have been recorded.

Perceived level of detail. It is recorded on a scale from 1 to 5 (as in Fig. 4), with 3 being a correct level of detail, 5 too much detail and 1 not enough detail. It was developed by merging in a single scale similar statements from other works [25, 49]. Moreover, we added the possibility to state that the explanation is not correct.

Requests of extra explanations. We automatically record the count of button presses to request further explanations.

Subjective understandability. We use the reverse of the illegibility subscale from the TOROS scale [3] (as seen in Tab. 2), which was found to be correlated with the meta-understanding construct.

Objective understandability. We use 3 single-choice questions. Two of them are about the two forced failures, and the third one is about a non-responsive person during pickup, one of the most frequent failures. Similar to our previous work[16], we also record the confidence and reasoning behind their answers (Fig. 9).

3.8 Hypotheses

The following hypotheses have been made:

H1: For the *NO_EXTRA* participants, the perceived of detail is *WHY* > *PER* > *WHAT*.

H2: For the *EXTRA* participants, *WHAT* participants will request more explanations than *PER* and *WHY* participants, which is a proxy measure for not having enough detailed information.

H3: For the *NO_EXTRA* participants, the objective and subjective understandability of *PER* participants is similar to that of *WHY* participants, since personalisation avoids repeating only known information, and is higher than that of *WHAT* participants, who might receive insufficient information.

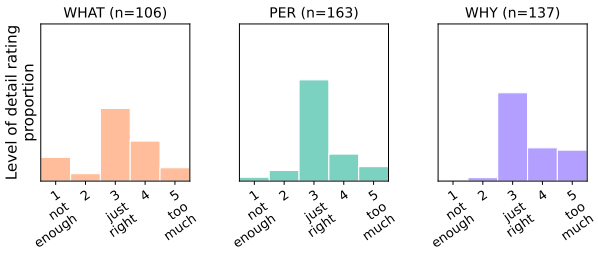

Figure 6: Rating of the perceived level of detail per group.

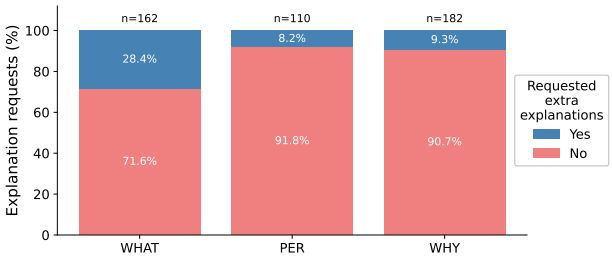

Figure 7: Presses of the button to know more when an explanation is displayed for the *EXTRA* group.

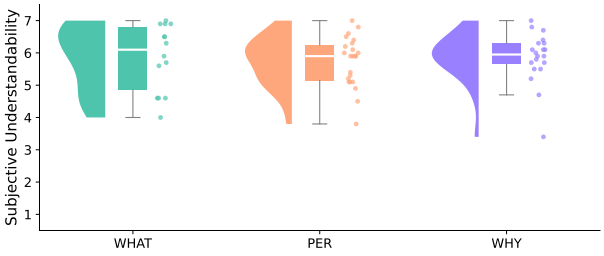

Figure 8: Subjective understandability per detail group.

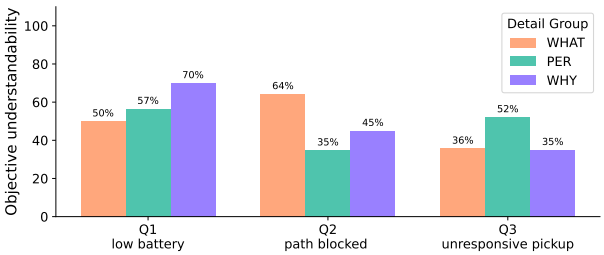

Figure 9: Objective understandability per detail group.

4 Ongoing results

By the second week, 247 items had been successfully delivered, and 1,030 explanations had been sent related to 769 situations. Explanations can be sent only to the requester (e.g. for an invalid request), or both to the sender and receiver. From those 769 situations, the main "what" categories were a request starting later (406), an incomplete delivery (139) or a delivery further delayed (118). Within the incomplete delivery, the main "why" subcategories were obstructed paths (36), timeouts to pick (31) and timeouts to drop (29).

First, we use the data from the participants who cannot get further explanations (*NO_EXTRA*). This ensures that those participants have only received the information related to *WHAT*, *PER* or *WHY* groups. We proceed to evaluate the *perceived level of detail*. We discard ratings marked "X", as they refer to explanations that participants rated as wrong. After validating with a Shapiro-Wilk test that they are not normal, a Kruskal-Wallis H-test proves significant differences ($p = .001$) and pairwise Mann-Whitney U tests with Bonferroni corrections demonstrate that *WHY* users perceive the level of detail higher than *WHAT* and *PER* users ($p = .010$ and $p = .003$ respectively), but that there are no significant differences for *PER* vs *WHAT* users ($p = 1.0$). Therefore, we can **partially support H1**. It can be observed from Fig. 6 that the *PER* group presents the highest count of "just right" ratings, and that, unexpectedly, *WHAT* participants reported excessive detail in many cases, although presenting more insufficiently detailed explanations. *WHY* participants generally considered explanations as too detailed. These results will be further analysed by explanation categories from Tab. 1 and evolution over repeatedly received explanations.

Regarding the *request of extra explanations*, we can **accept H2**, since a chi-squared test proves the significance of the count of presses for extra explanations ($p < .001$), and pairwise Fisher exact tests with Bonferroni correction show that *WHAT* participants press it with more frequency than *PER* ($p < .001$) and *WHY* ($p < .001$) participants. This suggests that *WHAT* participants do not have enough information within initial explanations.

Finally, regarding the levels of *objective and subjective understandability*, we include only the *NO_EXTRA* participants–presenting a consistent amount of received information–who have responed the initial questionnaire between the end of the first week and the beginning of the second. A total of 79 participants responded it, being n=39 from the *NO_EXTRA* group. Both metrics do not pass a normality test, and neither presents significant differences between *explanation level of detail* groups in a Kruskal-Wallis H-test ($p = 0.761$ for subjective, $p = 0.874$ for objective), **partially supporting H3** (Fig. 8 and 9). As expected, *PER* explanations do not lead to a decrease in understanding respect from *WHY* explanations, but unexpectedly, users receiving only *WHAT* explanations do reach similar levels of both objective and subjective understandability. Interestingly, results suggest an overestimation of subjective vs objective understanding. An extended analysis considering the subquestions of confidence and reasoning combined with the actual information received by each of the participants related to each of the questions, as well as the two future rounds after prolonged interaction with the system, will clarify these results.

5 Conclusions

We reported an ongoing longitudinal in-the-wild study of an office delivery robot that adapts the level of detail of its explanations. Ongoing results indicate that personalised explanations can preserve understanding comparable to fully detailed explanations while providing a more correct level of detail, while purely minimal explanations are often insufficient. Future work will analyse the full deployment results based on usage profiles, confidence in the responses, evolution in time and usability.

Acknowledgments

This work has been supported by Horizon Europe Marie Skłodowska-Curie grant agreement No. 101072488 (TRAIL).

References

[1] Andrew Anderson, David Piorkowski, Margaret Burnett, and Justin Weisz. 2025. An LLM's Attempts to Adapt to Diverse Software Engineers' Problem-Solving Styles: More Inclusive and Equitable? *arXiv preprint arXiv:2503.11018* (2025).

[2] Antonio Andriella, Carme Torras, and Guillem Alenyà. 2019. Learning robot policies using a high-level abstraction persona-behaviour simulator. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1–8.

[3] Georgios Angelopoulos, Dimitri Lacroix, Ricarda Wullenkord, Alessandra Rossi, Silvia Rossi, and Friederike Eyssel. 2025. Measuring transparency in intelligent robots. *Scientific Reports* (2025).

[4] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1078–1088.

[5] Tony Belpaeme. 2020. Advice to new human-robot interaction researchers. *Human-robot interaction: Evaluation methods and their standardization* (2020), 355–369.

[6] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R Eagan, and Winston Maxwell. 2023. On selective, mutable and dialogic XAI: A review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[7] Kayla Boggess, Shenghui Chen, and Lu Feng. 2020. Towards personalized explanation of robot path planning via user feedback. *arXiv preprint arXiv:2011.00524* (2020).

[8] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8, 2 (2013).

[9] Daniel Buschek, Malin Eiband, and Heinrich Hussmann. 2022. How to support users in understanding intelligent systems? an analysis and conceptual framework of user questions considering user mindsets, involvement, and knowledge outcomes. *ACM Transactions on Interactive Intelligent Systems* 12, 4 (2022), 1–27.

[10] Harriet R Cameron, Simon Castle-Green, Muhammad Chughtai, Liz Dowthwaite, Ayse Kucukyilmaz, Horia A Maior, Victor Ngo, Eike Schneiders, and Bernd C Stahl. 2024. A Taxonomy of Domestic Robot Failure Outcomes: Understanding the impact of failure on trustworthiness of domestic robots. In *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*. 1–14.

[11] Alison Cawsey. 1993. User modelling in interactive explanations. *User Modeling and User-Adapted Interaction* 3, 3 (1993), 221–247.

[12] Mohamed Amine Chatti, Mouadh Guesmi, Laura Vorgerd, Thao Ngo, Shoeb Joarder, Qurat Ul Ain, and Arham Muslim. 2022. Is more always better? The effects of personal characteristics and level of detail on the perception of explanations in a recommender system. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 254–264.

[13] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery. In *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*. 351–360.

[14] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[15] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. The who in XAI: how AI background shapes perceptions of AI explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–32.

[16] Ferran Gebellí, Anaís Garell, Séverin Lemaignan, and Raquel Ros. 2025. Dynamics of Mental Models: Objective Vs. Subjective User Understanding of a Robot in the Wild. *IEEE Robotics and Automation Letters* (2025).

[17] Ferran Gebellí, Anaís Garrell, Jan-Gerrit Habekost, Séverin Lemaignan, Stefan Wermter, and Raquel Ros. 2025. Personalised Explanations in Long-term Human-Robot Interactions. In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 775–782.

[18] Ferran Gebelli, Pradip Pramanick, Tamlin Love, Raquel Ros, Anais Garrell, Silvia Rossi, Antonio Andriella, Guillem Alenya, et al. 2025. Measuring User Understanding in Explainable Human-Robot Interaction: A Systematic Review. *Zenodo* (2025).

[19] Ferran Gebellí, Raquel Ros, Séverin Lemaignan, and Anaís Garrell. 2024. Co-designing Explainable Robots: A Participatory Design Approach for HRI. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1564–1570.

[20] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies* 72, 4 (2014), 367–382.

[21] Stephen R Grimm. 2010. The goal of explanation. *Studies in History and Philosophy of Science Part A* 41, 4 (2010), 337–344.

[22] Mouadh Guesmi, Mohamed Amine Chatti, Shoeb Joarder, Qurat Ul Ain, Rawaa Alatrash, Clara Siepmann, and Tannaz Vahidi. 2024. Interactive explanation with varying level of details in an explainable scientific literature recommender system. *International Journal of Human–Computer Interaction* 40, 22 (2024), 7248–7269.

[23] Mouadh Guesmi, Mohamed Amine Chatti, Laura Vorgerd, Shoeb Ahmed Joarder, Qurat Ul Ain, Thao Ngo, Shadi Zumor, Yiqi Sun, Fangzheng Ji, and Arham Muslim. 2021. Input or Output: Effects of Explanation Focus on the Perception of Explainable Recommendation with Varying Level of Details.. In *Intrs@ recsys*. 55–72.

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

[25] 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.

[26] Frank C Keil. 2003. Folkscience: Coarse interpretations of a complex reality. *Trends in cognitive sciences* 7, 8 (2003), 368–373.

[27] Parag Khanna, Elmira Yadollahi, Mårten Björkman, Iolanda Leite, and Christian Smith. 2023. Effects of Explanation Strategies to Resolve Failures in Human-Robot Collaboration. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, Busan, Korea, Republic of, 1829–1836.

[28] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2390–2395.

[29] Dimosthenis Kontogiorgos and Julie Shah. 2025. Questioning the Robot: Using Human Non-verbal Cues to Estimate the Need for Explanations. In *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*. 717–728.

[30] Raj Korpan, Daniel Tiourine, Sami Chen, and Susan Epstein. 2025. Evaluation of a Robot Navigatior's Explanations. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*.

[31] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[32] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*. 1–10.

[33] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[34] Shikhar Kumar, Yisrael Parmet, and Yael Edan. 2024. Exploratory user study on verbalization of explanations. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*. IEEE, 1–7.

[35] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.

[36] Gregory LeMasurier, Alvika Gautam, Zhao Han, Jacob W Crandall, and Holly A Yanco. 2024. Reactive or proactive? how robots should explain failures. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 413–422.

[37] Gregory LeMasurier, Christian Tagliamonte, Jacob Breen, Daniel Maccaline, and Holly A Yanco. 2024. Templated vs. Generative: Explaining Robot Failures. In *33rd IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 1346–1353.

[38] Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* (2023).

[39] Rhema Linder, Sina Mohseni, Fan Yang, Shiva K Pentyala, Eric D Ragan, and Xia Ben Hu. 2021. How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters* 2, 4 (2021), e49.

[40] Zeyi Liu, Arpit Bahety, and Shuran Song. 2023. Reflect: Summarizing robot experiences for failure explanation and correction. *arXiv preprint arXiv:2306.15724* (2023).

[41] Tamlin Love, Ferran Gebellí, Pradip Pramanick, Antonio Andriella, Guillem Alenyà, Anais Garrell, Raquel Ros, and Silvia Rossi. 2026. HEXAR: a Hierarchical Explainability Architecture for Robots. *arXiv preprint arXiv:2601.03070* (2026).

[42] Steve Macenski, Francisco Martín, Ruffin White, and Jonatan Ginés Clavero. 2020. The marathon 2: A navigation system. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2718–2725.

[43] Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022. Memory-assisted prompt editing to improve GPT-3 after deployment. *arXiv preprint arXiv:2201.06009* (2022).

[44] Kayla Matheus, Rebecca Ramnauth, Brian Scassellati, and Nicole Salomons. 2025. Long-Term Interactions with Social Robots: Trends, Insights, and Recommendations. *ACM Transactions on Human-Robot Interaction* (2025).

[45] Philip Mavrepis, Georgios Makridis, Georgios Fatouros, Vasileios Koukos, Maria Margarita Separdani, and Dimosthenis Kyriazis. 2024. XAI for all: Can large language models simplify explainable AI? *arXiv preprint arXiv:2401.13110* (2024).

[46] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th international conference on intelligent user interfaces*. 397–407.

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

[48] Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi, Abdeljalil Abbas-Turki, Stéphane Galland, and Christophe Nicolle. 2022. The quest of parsimonious XAI: A human-agent architecture for explanation formulation. *Artificial intelligence* 302 (2022), 103573.

[49] Yazan Mualla, Igor Haman Tchappi, Amro Najjar, Timotheus Kampik, Stéphane Galland, and Christophe Nicolle. 2020. Human-agent Explainability: An Experimental Case Study on the Filtering of Explanations.. In *ICAART*. 378–385.

[50] Andreas Naoum, Parag Khanna, Elmira Yadollahi, Mårten Björkman, and Christian Smith. 2025. Adapting robot's explanation for failures based on observed human behavior in human-robot collaboration. *arXiv preprint arXiv:2504.09717* (2025).

[51] Robert Nimmo, Marios Constantinides, Ke Zhou, Daniele Quercia, and Simone Stumpf. 2024. User characteristics in explainable AI: The rabbit hole of personalization?. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–13.

[52] Martin Obaidi, Jannik Fischbach, Marc Herrmann, Hannah Deters, Jakob Droste, Jil Klünder, and Kurt Schneider. 2025. How Does Users' App Knowledge Influence the Preferred Level of Detail and Format of Software Explanations?. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 106–122.

[53] Guglielmo Papagni and Sabine Koeszegi. 2020. Understandable and trustworthy explainable robots: A sensemaking perspective. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2020), 13–30.

[54] Sabid Bin Habib Pias, Alicia Freel, Timothy Trammel, Taslima Akter, Donald Williamson, and Apu Kapadia. 2024. The Drawback of Insight: Detailed Explanations Can Reduce Agreement with XAI. *arXiv preprint arXiv:2404.19629* (2024).

[55] Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081* (2023).

[56] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Autonomous agents and multi-agent systems* 33, 6 (2019), 673–705.

[57] Silvia Rossi, François Ferland, and Adriana Tapus. 2017. User profiling and behavioral adaptation for HRI. *Pattern Recognition Letters* 99 (2017), 3–12.

[58] Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization methods for personalizing large language models through retrieval augmentation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 752–762.

[59] Rossitza Setchi, Maryam Banitalebi Dehkordi, and Juwairiya Siraj Khan. 2020. Explainable robotics in human-robot interactions. *Procedia Computer Science* 176 (2020), 3057–3066.

[60] Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.

[61] Andrew Silva, Pradyumna Tambwekar, Mariah Schrum, and Matthew Gombolay. 2024. Towards balancing preference and performance through adaptive personalized explainability. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 658–668.

[62] David Sobrín-Hidalgo, Miguel A González-Santamarta, Ángel M Guerrero-Higueras, Francisco J Rodríguez-Lera, and Vicente Matellán-Olivera. 2024. Explaining Autonomy: Enhancing Human-Robot Interaction through Explanation Generation with Large Language Models. *arXiv preprint arXiv:2402.04206* (2024).

[63] Utkarsh Soni, Sarath Sreedharan, and Subbarao Kambhampati. 2021. Not all users are the same: Providing personalized explanations for sequential decision making problems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 6240–6247.

[64] Timo Speith and Markus Langer. 2023. A new perspective on evaluation methods for explainable artificial intelligence (XAI). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*. IEEE, 325–331.

[65] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.

[66] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. 2021. Foundations of explanations as model reconciliation. *Artificial Intelligence* 301 (2021), 103558.

[67] Sophie van der Woerdt and Pim Haselager. 2016. Lack of effort or lack of ability? robot failures and human perception of agency and responsibility. In *Benelux conference on artificial intelligence*. Springer, 155–168.

[68] Ruben S Verhagen, Mark A Neerincx, Can Parlar, Marin Vogel, and Myrthe L Tielman. 2023. Personalized agent explanations for human-agent teamwork: Adapting explanations to user trust, workload, and performance. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*. 2316–2318.

[69] 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.

[70] Lennart Wachowiak, Oya Celiktutan, Andrew Coles, and Gerard Canal. 2023. A Survey of Evaluation Methods and Metrics for Explanations in Human–Robot Interaction (HRI). In *ICRA2023 Workshop on Explainable Robotics*.

[71] Lennart Wachowiak, Andrew Coles, Gerard Canal, and Oya Celiktutan. 2024. A taxonomy of explanation types and need indicators in human–agent collaborations. *International Journal of Social Robotics* 16, 7 (2024), 1681–1692.

[72] Lennart Wachowiak, Andrew Coles, Oya Celiktutan, and Gerard Canal. 2024. Are Large Language Models Aligned with People's Social Intuitions for Human–Robot Interactions?. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2520–2527.

[73] Lennart Wachowiak, Andrew Fenn, Haris Kamran, Andrew Coles, Oya Celiktutan, and Gerard Canal. 2024. When do people want an explanation from a robot?. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 752–761.

[74] Lennart Wachowiak, Peter Tisnikar, Gerard Canal, Andrew Coles, Matteo Leonetti, and Oya Celiktutan. 2024. Predicting when and what to explain from multimodal eye tracking and task signals. *IEEE Transactions on Affective Computing* 16, 1 (2024), 179–190.

[75] Zihan Wang, Brian Liang, Varad Dhat, Zander Brumbaugh, Nick Walker, Ranjay Krishna, and Maya Cakmak. 2024. I Can Tell What I am Doing: Toward Real-World Natural Language Grounding of Robot Experiences. In *Annual Conference on Robot Learning*.

[76] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. 2017. Robot transparency: Improving understanding of intelligent behaviour for designers and users. In *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Guildford, UK, July 19–21, 2017, Proceedings 18*. Springer, 274–289.

[77] Stanisław Woźniak, Bartłomiej Koptyra, Arkadiusz Janz, Przemysław Kazienko, and Jan Kocoń. 2024. Personalized large language models. *arXiv preprint arXiv:2402.09269* (2024).

[78] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots* 47, 8 (2023), 1087–1102.