What Would I Do If...? Promoting Understanding in HRI through Real-Time Explanations in the Wild

Tamlin Love[1], Antonio Andriella[2] and Guillem Alenyà[1]

***Abstract*— As robots become more and more integrated in human spaces, it is increasingly important for them to be able to explain their decisions to the people they interact with. These explanations need to be generated automatically and in real-time in response to decisions taken in dynamic and often unstructured environments. However, most research in explainable human-robot interaction only considers explanations (often manually selected) presented in controlled environments. We present an explanation generation method based on counterfactuals and demonstrate its use in an "in-the-wild" experiment using automatically generated and selected explanations of autonomous interactions with real people to assess the effect of these explanations on participants' ability to predict the robot's behaviour in hypothetical scenarios. Our results suggest that explanations help aid one's ability to predict the robot's behaviour, but also that the addition of counterfactual statements may add some burden and counteract this beneficial effect.**

I. INTRODUCTION

As advances in the field of robotics continue to be made, people are increasingly afforded the opportunity to interact with robots in public spaces such as hospitals [1], stores [2] and restaurants [3]. However, as the potential impact robot decisions can have on people's daily lives grows, so too does the need for the decision-making of these robots to be understood and explained [4][5], with the European Union going so far as to designate a "right to explanation" [6].

Given the breadth of the topic of explainability, from explainable AI (XAI) [7] to explainable robotics [8], it is important to specify what exactly is meant by "explainability". Here we adopt the definition argued for by Miller that a decision being *explainable* describes the degree to which the causes of that decision can be understood [9]. Our focus in this work is on improving explainability via *explanation* - the explicit statement of the reasons for a decision - and specifically pose the following research question: *Do explanations of a robot's decision in a given context improve one's ability to understand (and thus predict) the decision-making of the robot in similar contexts?*

To answer this question, we need to evaluate how explanations can promote understanding of a decision-making system. When evaluating explanations, researchers often conduct experiments that simplify and abstract a target domain to more tightly control the conditions of an experiment and remove the influence of confounding factors

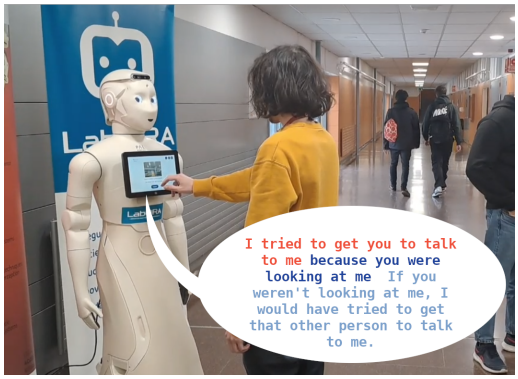


Fig. 1: A person interacting with the robot and being presented with an explanation. In the explanation, red indicates the decision the robot is explaining, dark blue is the explanation itself, and light blue is a counterfactual statement building from the explanation.

[10]. This type of approach (dubbed "human-grounded" by Doshi-Velez and Kim [10]) has its upsides, namely by making an evaluation of explanations easier to perform, but the controlled conditions of a lab may not sufficiently match the dynamic and unstructured conditions of some real environments. For example, if a study is evaluating an explanation of a human-robot interaction, then the person involved in the interaction (if indeed the interaction is real at all) may not necessarily be the same person involved in the evaluation. If a robot is to explain its decisions to the people it interacts with, then the fact that the person is involved in both the interaction and the explanation may be crucial in examining how such explanations may be interpreted.

In this work, we aim to assess explanations in the context of real interactions in unstructured "in-the-wild" environments. Our contributions are as follows. Firstly, we develop a system that allows a robot to elicit interactions from people in a multi-person environment and then provide automatically generated and online counterfactual explanations for the specific elicitation decisions it makes, based on previous work [11]. Secondly, we design and conduct an experiment in which said robot is left unsupervised in a public space to elicit interactions from participants (N=92), explain its decisions to those same participants and assess their understanding of the decision-making system through their ability to predict the robot's decisions in similar, hypothetical cases.

We find that explanations are indeed helpful and improve participants' ability to predict the robot's decision-making, but also that the additional of counterfactual statements in an explanation, while providing more information, may

[1]Tamlin Love and Guillem Alenyà are with the Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028, Barcelona, Spain tlove@iri.upc.edu, galenya@iri.upc.edu

[2]Antonio Andriella is with the Artificial Intelligence Research Institute (IIIA-CSIC), Campus de la UAB, 08193 Bellaterra, Barcelona, Spain aandriella@iri.upc.edu

counteract this effect.

To improve reproducibility, the ROS1 implementations of the perception, decision-making, explanation generation and robot behaviour are made publically available[1]. All components apart from the robot's behaviour are robot-agnostic.

II. RELATED WORK

There is a wide range of methods for automatically generating explanations within XAI [7], some of the most popular being LIME [12] and SHAP [13], which compute the importance of each feature to a particular decision. Given the contrastive nature of explanations [9], many approaches exploit counterfactual reasoning to generate suitable explanations of a decision [14]. Counterfactual explanations, either explicitly or implicitly, suggest how an input can be changed to change the decision. Examples include the counterfactual explanations of Wachter et al. [6] and Albini et al. [15].

Explanation generation has seen some attention in robotics contexts [8][5][4]. For example, Sobrín-Hidalgo et al. [16] use an LLM to generate natural language explanations from the logs of a robot performing a navigation task. Leveraging counterfactual reasoning, Diehl and Ramirez-Amaro [17] use a causal model learned in simulation to predict and explain potential failures in a block stacking task. Some works focus explicitly on explaining human-robot interactions, such as Stange et al. [18], who use episodic memory to explain a robot's decisions in terms of the strategies and needs of the robot at the moment of decision-making. Our approach to generating explanations is most similar to that of Diehl and Ramirez-Amaro [17], but differs significantly in the counterfactual search (which more closely resembles that of Albini et al. [15]) and in the number of counterfactual situations encoded by the explanations returned.

Given the goal of explanations to improve understanding, several studies have proposed behavioural evaluations to measure understanding. For example, Chandran Nair et al. [19] have participants reconstruct the order of video clips of a robot performing a task to evaluate the effectiveness of different types of explanations. Van der Waa et al. [20] use the ability of participants to predict the output of an advice-giving system, given hand-designed contrastive rule-based or example-based explanations, to measure the construct of *system understanding*. Our approach differs from these works as we evaluate explanations generated automatically and in real-time, presented to participants who are also the ones interacting with the robot, in an "in-the-wild" environment.

III. PERCEPTION AND DECISION-MAKING

In this section, we describe the way the robot perceives the people it interacts with and how it autonomously makes decisions, adapted from a previous publication [11].

The robot is equipped with an RGBD camera. From the raw image, OpenDR's OpenPose implementation is used to detect the 2D pose of each visible person [21]. Together with the depth information, 3D poses for each person can be determined, which in turn can be used to calculate low-level features such as the distance of each person to the robot (denoted D_{XR} for person X and robot R) and the direction they are facing. The pose estimator also outputs a confidence value (PEC_X) for each detected person. High-level features such as each person's mutual gaze score with the robot (MG_{XR}) and a derived engagement score (EV_{XR}) are then calculated [22].

To decide how and with whom to interact, the robot follows a set of heuristic rules. If the robot does not detect anyone, it will do nothing (action *NOTHING*). Otherwise, it will calculate a score $Score(X) = PEC_X \times EV_{XR}$ for each detected person X. The robot targets an elicitation towards person $T = \text{argmax}_X\, Score(X)$ if $Score(T)$ is above a specified threshold (action *ELICIT_TARGET* on target T), otherwise performing a general elicitation targeted at no one in particular (action *ELICIT_GENERAL*).

Translating these decisions to actions performed by the robot, the *ELICIT_TARGET* decision results in the robot looking at the target with a happy expression while performing one of several waving gestures, while the *ELICIT_GENERAL* decision results in the robot looking straight ahead with a neutral expression while performing simple idle gestures. For both of these decisions, the robot also utters a short sentence (e.g. "Hi" or "Good day").

IV. EXPLANATION GENERATION

In this section, we describe our approach to automatically generating explanations of the decisions the robot takes. In our approach, explanations answer the question "Why $\mathcal{D} = \langle \mathcal{A}, \mathcal{T} \rangle$?", where the decision $\mathcal{D}$ is composed of the action $\mathcal{A}$ that the robot decides to perform and the target $\mathcal{T}$ of that action (or $\varnothing$ if $\mathcal{A}$ has no target). The answer to this question consists of the assignments of one or more variables. For example, if the robot makes the decision $\langle$*ELICIT_TARGET,A*$\rangle$, a valid explanation for this decision might be $D_{AR} < 3m$, implying both that any distance less than 3 metres would result in the same decision and that, if person A would have been 3 metres away or further, the robot would have taken either a different action or the same action with a different target.

To find these explanations, we employ a counterfactual search on a causal model of the perception system described in Sec. III and how it relates to the decision-making algorithm, depicted in Fig. 2, encoding the causal dynamics between the relevant features. The use of a causal model allows interventions to be performed on variables such that the effect of these interventions is propagated down the graph to the decision. This differs from a sampling approach (e.g. as in LIME [12]) as no change occurs to ancestors of the intervened variable. By using a graphical model, the graph can be expanded or contracted at each decision point to accommodate a dynamic number of people, as opposed to limiting the number of people by storing every variable in a fixed-length feature vector.

The counterfactual search then consists of applying interventions on each variable in the causal graph to find a

[1] https://github.com/tamlinlove/engage

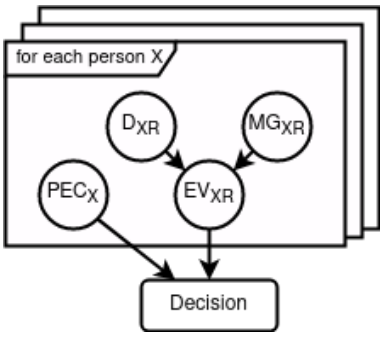


Fig. 2: A causal model of the perception system and how it relates to the decision-making algorithm. The algorithm relies on the pose estimation confidence PEC_X and the engagement value with the robot EV_{XR} for each person X. EV_{XR} is determined by the distance (D_{XR}) and mutual gaze score with the robot (MG_{XR}).

threshold above or below which any setting of that variable would result in a different decision from the real one. To make this search tractable, the variables in the model are discretised. For example, continuing with the example above, the causal search may find that any intervention setting the distance D_{AR} to a value of 3 or more metres would result in a different decision to the one the robot took, while any intervention setting $D_{AR} < 3m$ would result in the same decision, hence arriving at the explanation $D_{AR} < 3m$.

The time complexity of this counterfactual search to find all explanations consisting of a single variable is $\mathcal{O}(|V|^2NC)$, where V is the set of variables for each person in the causal model, N is the number of people detected and C is the maximum cardinality of the discrete variables in V. V and C are fixed on model specification, while N varies between interactions. In our experiment, these values are all relatively small, with $|V| = 4$, $C = 13$ (13 discrete distance values, 4 discrete values for every other variable) and the largest number of people encountered being $N = 7$.

Once an explanation is selected, it can be converted to natural language using an authored mapping, as can its associated counterfactual statement. Fig. 3 illustrates some examples of explanations.

V. EXPERIMENT

In this section, we present an experiment for evaluating our explanation generation method (Sec. IV) in the context of a robot eliciting interactions from people in an unstructured, "in-the-wild" environment. This study has been approved by the ethical committee of the Spanish National Research Council (CSIC).

In our experiment, we positioned a robot (a Pal ARI [23]) in the entrance hall of a university building, as depicted in Fig. 1, facing the entrance of the building (as can be seen in Fig. 3). The robot remained in this position for 8 hours between 09:30 and 17:30 every day for four days, totalling 32 hours of operation. Computations were performed on an MSI Stealth GS77 laptop with an Intel Core i7-12700H CPU, 32 GB RAM and an Nvidia GeForce RTX 3070 Ti GPU. Given the "in-the-wild" setting of the experiment, participants consisted of people who voluntarily approached the robot and completed a short test on the robot's tablet interface. The experiment was conducted in Catalan.

A. Experimental Condition

The experiment was set up as a between-subject study, in which we manipulated the form of explanation generated by the robot as a consequence of its decision during its interaction with the participant. Each participant was randomly assigned to one of the three conditions:

Control (group **C**) - participants in this group were only given a description of the decision

- e.g. "I tried to get Person A to talk to me."
- e.g. "I tried to get anyone to talk to me."

Explanation (group **E**) - participants in this group were given a natural language explanation detailing the reason why the decision was made, obtained from the process described in Sec. IV. Where multiple explanations were available from the explanation module, one was selected randomly such that each explanation consisted of only one variable.

- e.g. "I tried to get Person A to talk to me because they were looking at me."
- e.g. "I tried to get anyone to talk to me because I was not confident in my detection of Person A's skeleton."

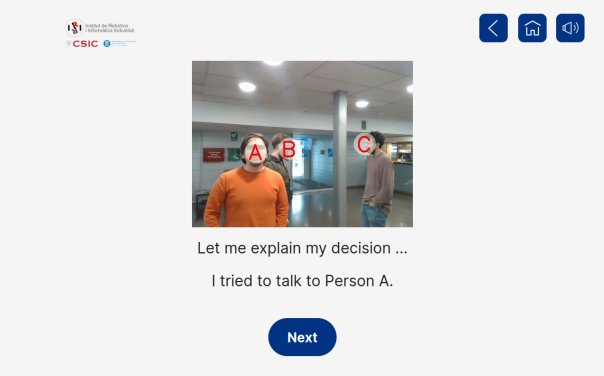
Explanation + Counterfactual (group **CF**) - this group was identical to group **E**, except that they also received a counterfactual statement indicating what the robot would have done if the stated reason was not so.

- e.g. "I tried to get Person A to talk to me because they were looking at me. If Person A was not looking at me, I would have tried to get Person B to talk to me."
- e.g. "I tried to get anyone to talk to me because I was not confident in my detection of Person A's skeleton. If I was at least a little bit confident of my detection of Person A's skeleton, I would have tried to get Person A to talk to me."

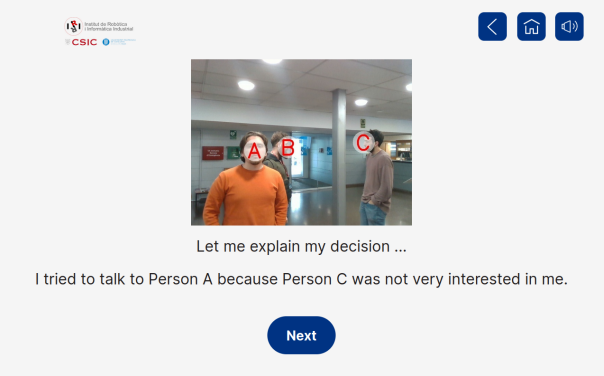
The total number of participants who performed the test was 115. However, we excluded 23 who interacted more than once with the robot. Therefore, $N = 92$ participants were considered for the study: 30 in Group **C**, 31 in Group **E** and 31 in Group **CF**). Given the number of participants, and an effect size, $f^2 = 0.1$ we expected a statistical power of 0.77 for a $p < 0.05$ when considering 2 predictors.

B. Experimental Measure

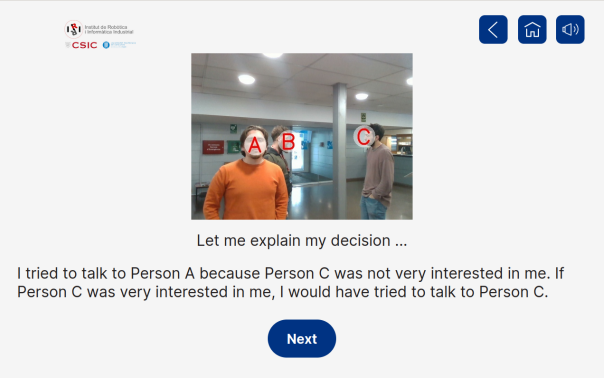
The behavioural measure we used to assess participants' understanding of the robot's decisions is their *ability to predict the decision-making* of the robot in a hypothetical scenario, inspired by the forward simulation metric proposed by Doshi-Velez and Kim [10] and the *advice prediction* measure used by van der Waa et al. [20]. Given that our experiment took place "in-the-wild" where attracting participants necessarily interrupts their normal schedule, we opted to focus on this measure alone, rather than including additional subjective questionnaires, to shorten the duration of each test. Results from this experiment can suggest the



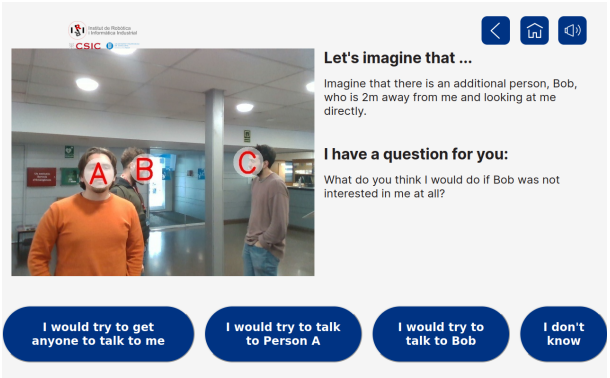
(a) Explanation for Group **C** (Control)



(b) Explanation for Group **E** (Explanation)



(c) Explanation for Group **CF** (Explanation + Counterfactual)



(d) Corresponding question (identical for each group)

Fig. 3: Example of the explanations (3a - 3c) and questions (3d) presented to participants.

kinds of subjective measures we may wish to investigate in future experiments (as discussed in Sec. VII).

To convert the participants' answers into a score for use in the statistical analysis, we assigned -1 for incorrect answers, 0 for selecting the "I don't know" option and 1 for correct answers.

C. Research Hypotheses

We aimed to test the following hypotheses.

H 1 *Participants who interact with a robot that provides explanations of its decisions (groups **E** and **CF**) can better predict how the robot would act in a similar situation in comparison to those who are provided with no explanation (**C**).*

H 2 *Participants who interact with a robot that provides counterfactual statements (group **CF**) can better predict how the robot would act in a similar situation in comparison to those who are provided with no explanation (**C**) or only explanation (**E**).*

H 3 *Participants provided with explanations using physically-grounded variables such as distance from the robot (D_{XR}) and mutual gaze (MG_{XR}) with the robot are able to better predict the robot's behaviour than those provided with explanations using more abstract variables, such as engagement with the robot (EV_{XR}) and the person's pose estimation confidence (PEC_X).*

D. Procedure

To attract participants, the robot executed the decision-making procedures described in Sec. III. Once a participant approached the robot and began to interact with its tablet, all decision-making behaviour was paused. The participant was given the option to consent to participating in the experiment, and if they accepted, they were asked whether or not they had interacted with the robot before. If participants had already interacted with the robot, we excluded them from the experiment, but they were still welcome to play with the robot and take another test. After a short contextual description of the experiment provided to first-time participants, the participant was then shown a picture of the moment the last decision was made, with detected people labelled alphabetically. Accompanying the picture was an explanation of the robot's decision, determined by the participant group (see Fig. 3a - 3c).

Participants were then asked to imagine a hypothetical scenario in which an additional person (here given the name "Bob") was present in the same scene as the participant. Some context relating to the additional person was provided, namely that "Bob" was 2 metres away from the robot (and whether or not this put them closer or further than anyone else) and looking directly at it (although one of these may be omitted if the explanation provided relates to this variable). Participants were then asked what the robot would do if some intervention was made on the additional person,

using the same variable as presented in the explanation for groups **E** and **CF**. Fig. 3d shows an example question corresponding to the explanations in Fig. 3a - 3c.

The participant was presented with four multiple-choice answers, the first three of which contained the correct answer and two incorrect answers in the form of decisions the robot would make, with the last answer always being "I don't know". Upon selecting an answer, the participant was thanked, the test ended, and after some time the robot resumed its elicitation behaviour.

VI. RESULTS AND DISCUSSION

To test hypothesis H1, we ran a logistic regression with the type of explanation (**C**, **E**, **CF**) as predictors controlling for participants' ability to predict the correct robot's decision. The overall model was significant ($F(2, 90) = 3.37$, $R^2 = 0.07$, $p < 0.05$). Results indicate that participants who received explanations from a robot (group **E**) were better able to predict the robot's decisions ($M = 0.23$ $SD = 0.89$) than those who received no explanation (group **C**) ($M = 0.06$ $SD = 0.94$, $\beta = 0.43$, $p < 0.05$). However, counterfactual statements (group **CF**) did not have a statistically significant effect on participants' ability to predict the robot's decisions ($M = 0.03$, $SD = 0.91$, $\beta = -0.14$, $p = 0.43$).

To address hypothesis H2, we ran a logistic regression excluding group **C**. The results indicate that the explanation type (**E** vs **CF**) had a significant effect on the participants' ability to predict the robot's decision. Participants who were provided with additional counterfactual statements were less able to predict the robot's decision compared to those who received only explanations ($F(1, 61) = 6.62$, $R^2 = 0.1$, $\beta = -0.57$, $p < 0.05$, $f^2 = 0.1$, statistical power 0.69 with one predictor).

Overall, these results suggest that providing an explanation for the robot's decision helped participants predict the robot's behaviour in the hypothetical scenario, suggesting that the explanations led to an improved understanding of the robot's decision. This partially provides evidence for hypothesis H1. However, the results also suggest that participants who received explanations and counterfactual statements were not significantly different in their ability to predict the robot's behaviour than the control group, performing worse than those who only received explanations, going against hypothesis H2. These results appear to agree with similar results from van der Waa et al. [20], who found no significant difference in participants' ability to predict a system's output given contrastive rule-based explanations, whose form is similar to the explanations and counterfactual statements received by group **CF**, when compared to a control who received no explanations.

While it is impossible to ascertain the reasons why the addition of counterfactual statements harmed performance given the data we have collected, one possible reason for this effect is that adding the counterfactual statement to the explanation added too much information for participants to parse, increasing cognitive load and confusion. Another possibility is that the counterfactual statement may have reinforced false impressions of the robot's "intentions". For example, consider the explanations depicted in Fig. 3. In this example, Person A did not have the maximum possible engagement value (EV_{AR}), and thus one valid counterfactual explanation, as presented here, is to suggest that Person C would be a preferred target if their engagement value EV_{CR} was very high. Some participants may interpret such an explanation as indicating that the robot would somehow prefer to speak to Person C but had to settle for Person A. While this confusion could be communicated by the explanation alone, the counterfactual statement may possibly serve to reinforce it.

To test hypothesis H3, we included the variables used for the explanation as an additional predictor to our logistic regression model described for H1. The overall model was significant when the explanation variable was "Mutual Gaze" ($F(5, 87) = 3.4$, $R^2 = 0.17$, $p < 0.05$, $f^2 = 0.1$, statistical power 0.6 with five predictors), while we did not find any significance for the other three variables. The results show a significant effect of the explanation variable for participants who belonged to Group **CF** who performed worse compared to participants who belonged to Group **C** ($\beta = -1.24$, $p < 0.05$). Given the small sample sizes when combining group and explanation variable, we also report the descriptive statistics to further speculate on the effect of each explanation variable, "Engagement Value" ($M = -0.03$, $SD = 0.92$), for "Mutual Gaze" ($M = -0.04$, $SD = 0.88$), "Pose Estimation Confidence" ($M = -0.05$, $SD = 0.93$), and "Distance" ($M = 0.37$, $SD = 0.88$).

While these results hint at distance-based questions being overall easier to predict, which partially supports hypothesis H3, the lack of statistical significance and the fact that questions based on mutual gaze appear to be harder to predict (especially for those in group **CF**, where results are significant), may prove damning for the hypothesis. Further studies with larger sample sizes would be needed to more definitively determine the effect of the explanation variable on the ability to predict the robot's decision-making, accounting for the types of explanations provided.

Finally, as an exploratory hypothesis, we evaluated whether the number of people in the depicted scene had an impact on the participants' ability to guess the correct answer regardless of the groups to which they belonged, speculating that the more people, the higher the chance to make a mistake, due to more complexity in the scene, more possibilities for errors in the perception system (e.g. due to occlusions in a crowd of people) and due to the effect of misinterpreted robot intentions discussed above. Results show a trend that could support our hypothesis. Indeed, when the scene contained more than 1 person, the participants tended to guess the incorrect answer ($F(1, 91) = 2.99$, $R^2 = 0.03$ $\beta = -0.33$, $p = 0.08$) more often than in cases where only 1 person was detected.

VII. CONCLUSION

In this work, we have devised a method for generating explanations that allows robotic systems to provide real-time

counterfactual explanations of their decisions in real-world settings that involve multiple people. To assess the efficacy of our approach in improving participants' understanding of the robot's decision-making process, we conducted an "in-the-wild" user study with 92 participants.

Our findings suggest that explanations of a robot's decisions are indeed useful in aiding people's abilities to predict the robot's behaviour in similar situations, which would suggest an affirmative answer to our research question. However, our results also indicate that the addition of counterfactual statements to an explanation erases the benefits of such explanations, though further work is needed to ascertain the mechanism by which this takes place. This future work could explicitly measure factors such as confusion and cognitive load. Further research on generating different types of explanations (e.g. comparing to a past decision rather than a counterfactual), different approaches to presenting explanations (e.g. multi-modal explanations, proactive vs. reactive, etc.), and making the process of receiving an explanation more interactive (e.g. by allowing prolonged interactions involving follow-up questions and explanations) may also help us to understand how explanations can be used to promote understanding of robot behaviour in real-world environments.

We consider our work a necessary first step to the evaluation of automatically generated explanations of real interactions in real-world domains. However, it does come with some limitations, foremost of which are those associated with "in-the-wild" studies on explainable HRI. These include the large number of confounding factors present in real-world domains as well as the logistical and engineering difficulties associated with developing and deploying these systems outside of the lab. To address the latter concern, improvements can be made to the perception system to make measurements of features such as distance and gaze direction more reliable, and to the decision-making system and robot actions in order to make interactions more responsive, purposeful and meaningful to the users. Different evaluation metrics can also be investigated, provided they are practical to measure "in-the-wild". To address the former, more measurements can be made to capture as many factors as possible, such as human motions and group dynamics. User modelling can be used to incorporate the intentions, goals and mental states of the people involved in interactions and explanations.

VIII. ACKNOWLEDGEMENTS

This work was supported by Horizon Europe under the MSCA grant agreement No 101072488 (TRAIL); by the "European Union NextGenerationEU/PRTR" project CHLOE-GRAPH PID2020-118649RB-I00 funded by MCIN/ AEI /10.13039/ 501100011033.

REFERENCES

[1] A. Andriella, C. Torras, C. Abdelnour, and G. Alenyà, "Introducing CARESSER: A framework for in situ learning robot social assistance from expert knowledge and demonstrations," *User Modeling and User-Adapted Interaction*, vol. 33, no. 2, pp. 441–496, 2023.

[2] S. Edirisinghe, S. Satake, D. Brscic, Y. Liu, and T. Kanda, "Field trial of an autonomous shopworker robot that aims to provide friendly encouragement and exert social pressure," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2024, p. 194–202.

[3] H. Knight, D. Flynn, T. M. Oo, and J. Hansen, "Iterative robot waiter algorithm design: Service expectations and social factors," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2024, p. 394–402.

[4] F. Sado, C. K. Loo, W. S. Liew, M. Kerzel, and S. Wermter, "Explainable goal-driven agents and robots-a comprehensive review," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–41, 2023.

[5] T. Sakai and T. Nagai, "Explainable autonomous robots: A survey and perspective," *Advanced Robotics*, vol. 36, no. 5-6, pp. 219–238, 2022.

[6] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, p. 841, 2017.

[7] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, *et al.*, "Benchmarking and survey of explanation methods for black box models," *Data Mining and Knowledge Discovery*, pp. 1–60, 2023.

[8] R. Setchi, M. B. Dehkordi, and J. S. Khan, "Explainable robotics in human-robot interactions," *Procedia Computer Science*, vol. 176, pp. 3057–3066, 2020.

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

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

[11] T. Love, A. Andriella, and G. Alenyà, "Towards explainable proactive robot interactions for groups of people in unstructured environments," in *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2024.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?"," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016.

[13] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[14] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

[15] E. Albini, A. Rago, P. Baroni, and F. Toni, "Relation-based counterfactual explanations for bayesian network classifiers." in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2020, pp. 451–457.

[16] D. Sobrín-Hidalgo, M. A. González-Santamarta, Á. M. Guerrero-Higueras, F. J. Rodríguez-Lera, and V. Matellán-Olivera, "Explaining autonomy: Enhancing human-robot interaction through explanation generation with large language models," *arXiv preprint arXiv:2402.04206*, 2024.

[17] M. Diehl and K. Ramirez-Amaro, "A causal-based approach to explain, predict and prevent failures in robotic tasks," *Robotics and Autonomous Systems*, vol. 162, p. 104376, 2023.

[18] S. Stange, T. Hassan, F. Schröder, J. Konkol, and S. Kopp, "Self-explaining social robots: An explainable behavior generation architecture for human-robot interaction," *Frontiers in Artificial Intelligence*, vol. 5, p. 87, 2022.

[19] N. Chandran Nair, A. Rossi, and S. Rossi, "Impact of explanations on transparency in HRI: A study using the HRIVST metric," in *International Conference on Social Robotics*. Springer, 2023, pp. 171–180.

[20] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating xai: A comparison of rule-based and example-based explanations," *Artificial Intelligence*, vol. 291, p. 103404, 2021.

[21] N. Passalis, S. Pedrazzi, R. Babuska, W. Burgard, *et al.*, "OpenDR: An open toolkit for enabling high performance, low footprint deep learning for robotics," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2022, pp. 12 479–12 484.

[22] N. Webb, M. Giuliani, and S. Lemaignan, "Measuring visual social engagement from proxemics and gaze," in *31st IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 2022, pp. 757–762.

[23] S. Cooper, A. Di Fava, C. Vivas, L. Marchionni, and F. Ferro, "ARI: The social assistive robot and companion," in *29th IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 2020, pp. 745–751.