Personalising Explanations and Explaining Personalisation

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

***Abstract*— Both personalisation and explainability have become popular research topics in social robotics, each capable of improving human-robot interactions. However, challenges have been identified in both fields, from issues of transparency, bias and privacy in personalisation to issues of identifying and communicating relevant explanations in explainability. In this work, we examine the intersection of these two fields - using personalisation to improve explanations and explainability to improve personalisation - and identify a number of research directions that could be of benefit to both communities.**

I. INTRODUCTION

The personalisation of social robots - that is, adapting their behaviour to the needs and preferences of individual users - has been shown to improve perceptions of competence and trust [1], interaction quality, engagement and motivation [2], and can improve public acceptance of these technologies [3]. For example, consider a robot placed in a domestic environment to assist an elderly patient in daily living. Personalisation can be employed to adapt the robot's behaviours, such as offering personalised reminders for meals and medication, or factoring in preferences when guiding the patient through cognitive or physical exercises. In this way, the robot could improve the interaction quality for the patient and foster trust and acceptance for both the patient and caregiver.

However, drawbacks have been identified for personalisation, such as introducing biases in the robot's behaviour [4], lack of transparency [5], and the privacy concerns surrounding the collection of personal information [6]. To ensure transparency and traceability, and to instil confidence and trust in the robot, all stakeholders should be able to understand exactly how a robot's behaviour is impacted by personalisation.

Explainability seeks to improve a user's understanding of a decision-making system by explaining the reasons for its decisions, and has seen a recent surge in popularity, for both machine learning [7] and robotics [8]. Robots that can explain their decisions have the potential to address the identified drawbacks of personalisation, by exposing biases and communicating how personal information is used in decision-making. However, explainability has its own challenges. In a review of explanations in the social sciences, Miller [9] argues that explanations are **contrastive** (i.e. that "Why X?" questions are better understood as "Why X and not Y?" questions, even if the contrast is implicit) and **selected** (i.e. that explanations should focus on a few, relevant causes rather than overloading recipients with all possible causes). Automatically identifying the most appropriate contrast and selecting the most relevant causes remain open challenges for explainability. Additionally, in Human-Robot Interaction (HRI) settings, it can be challenging to resolve the communication ambiguities in expressing and interpreting queries and explanations [10]. Personalisation could prove useful in addressing these challenges by considering the unique needs and preferences of individual users.

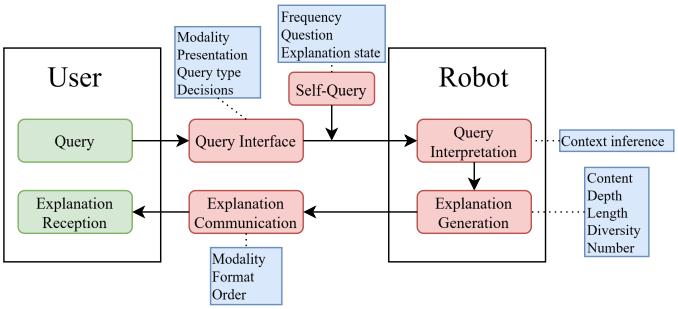

Fig. 1. Our framework for the explanation process, adapted from the literature [11], [12], with possible personalisation strategies identified for each component.

Clearly, HRI practitioners working in personalisation and explainability stand to gain from combining approaches in both fields. Thus, the aim of this work is to identify works at the intersection of these fields and propose research directions towards realising interpretable and user-aware social robots.

II. EXPLAINING PERSONALISATION

To improve transparency, the factors used by the personalisation system (such as the user's needs and preferences) can be incorporated into the state used by explainability algorithms to identify the reasons for the robot's decisions. While such systems have not been employed for HRI scenarios, they have been utilised in intelligent tutoring systems [13], recommendation systems [14] and robot mission planning [15].

Given the contrastive nature of explanations [9], an intuitive way of explaining the impact of personalisation is through *counterfactual explanations*, which examine how the robot's behaviour would change given a change

[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 antonio.andriella@iiia.csic.es

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; by the EU-founded project grant agreement No 101070930 (VALAWAI); and by the Research Council of Norway under the project SECUROPS (INT-NO/0875).

to its input [16]. Using counterfactual explanations, a user could pose a query to the system (e.g. "Why did you recommend I exercise now and not in the evening?") and an explanation can be generated that identifies both a reason and how a different set of preferences or needs could result in a different decision (e.g. "Because I think you prefer exercising in the morning. If not, I would have suggested exercising in the evening."). Given that preferences are often obtained from user data, we can in turn explain the robot's beliefs about preferences (e.g. "I think you prefer exercising in the morning because you seem happier when exercising in the morning versus in the evening."), thus fostering transparency over multiple levels of decision-making. Causal models of the decision-making system may be particularly fruitful, allowing for counterfactual scenarios to be interrogated through interventions on the model [17].

Explainability for bias detection has seen some attention [18], [19], and this can extend to detecting biases introduced by personalisation. Explanations of behaviour can be generated and assessed by various stakeholders (e.g. patients and caregivers) to determine if the reasons for these decisions align with their expectations and values [20]. Likewise, explainability can improve transparency surrounding personal information, allowing users to understand how their data is used to make decisions [21], though precautions should be taken to ensure that privacy is preserved during explanations.

III. PERSONALISING EXPLANATIONS

Just as explainability can address some drawbacks of personalisation, so too can personalisation address challenges in explainability. To this end, we present a framework for explainability in HRI settings inspired by Anjomshoae et al. [11] and Matarese et al. [12], depicted in Fig. 1. For each component of this framework over which the robot has control, we identify opportunities for personalising explanations.

The process typically begins with the user requesting an explanation from the robot, mediated by an **query interface** that determines the communication channels between the human and the robot. The robot could also generate explanations unprompted through *self-query*. After receiving a query, the robot must interpret it (**query interpretation**), transforming it into a formal query that can impose conditions on the search for explanations. Such a process involves resolving any ambiguities in the query, such as those created by implicit contrasts in a "why (not)" question. Once the query has been interpreted, one or more suitable explanations must be generated that match the query (**explanation generation**). Finally, once one or more suitable explanations have been found, they must be communicated in a human-understandable format (**explanation communication**).

Query Interface: The query interface can be personalised in several ways. Firstly, the interface may support multiple question-asking modalities (e.g. natural language, a GUI, etc.), and the use of one over the other can reflect user roles or preferences. The exact presentation of the interface could also be adapted, leveraging personalisation of user interfaces [22].

The types of query supported could be adapted to the role of the user. For example, a programmer might be able to ask a range of technical questions to the robot for debugging purposes which may not be presented to lay users. Similarly, the decisions allowed for querying could be restricted to certain users.

If the robot is capable of *self-query* then the frequency of explanations, situations in which explanations are warranted, and the question the robot asks itself could all be personalised [23].

Query Interpretation: Personalisation can be used to infer the context implicit in user queries. Herbold et al. [24] address the problem of automatically detecting implicit contrasts in the context of rule-based systems. Among other things, they factor the user's relationship to the triggered rule (e.g. as the creator of the rule) in whether or not a user is likely to expect that rule to be fired. Such personalisation could be extended, for example, to consider the user's history of interactions with the robot, their role (e.g. a caregiver might have different expectations than a patient) and preferences (e.g. a user might want to know why their preferred behaviour wasn't triggered).

Explanation Generation: Within this component, there are several opportunities for personalisation. Firstly, the choice of state variables featured in the explanation can be affected by user needs or preferences. For example, to preserve privacy, explanations involving protected variables might be restricted to certain users. If a layered system is used, with features ranging from low-level to high-level, the "depth" of the explanation could be personalised (e.g. a patient might be interested in the high-level human activity the robot detected, while a programmer might be interested in the individual keypoints of the detected skeleton). However, care should be taken not to "over-personalise" explanations using unimportant variables [25]. The length (in terms of number of variables), diversity and number of explanations could also be personalised based on user roles and preferences [26].

Explanation Communication: Explanations can be communicated in a number of modalities (e.g. text, images, embodied actions, etc.), and the choice of modality could be informed by user needs and preferences. After a modality is chosen, the format of the explanation can be personalised, adapting the explanation template[27], choice of words [15], types of graphical elements [28] or the presentation of a user interface [29]. If multiple explanations are provided, the order in which they are presented could be personalised [26].

IV. CONCLUSION

In conclusion, there is room for personalisation's drawbacks to be addressed by explainability, while similarly, explanations can potentially be improved through personalisation. This work represents a step towards bringing together the two fields, identifying research directions at their intersection. Our intention is that each of these directions can be explored in future work, especially in HRI, to facilitate the development of interpretable, user-aware social robots.

REFERENCES

[1] S. Schneider and F. Kummert, "Comparing robot and human guided personalization: adaptive exercise robots are perceived as more competent and trustworthy," *International Journal of Social Robotics*, vol. 13, no. 2, pp. 169–185, 2021.

[2] B. Irfan, N. Céspedes, J. Casas, E. Senft, L. F. Gutiérrez, M. Rincon-Roncancio, C. A. Cifuentes, T. Belpaeme, and M. Múnera, "Personalised socially assistive robot for cardiac rehabilitation: Critical reflections on long-term interactions in the real world," *User Modeling and User-Adapted Interaction*, vol. 33, no. 2, pp. 497–544, 2023.

[3] K. Liu and D. Tao, "The roles of trust, personalization, loss of privacy, and anthropomorphism in public acceptance of smart healthcare services," *Computers in Human Behavior*, vol. 127, p. 107026, 2022.

[4] A. Kubota, M. Pourebadi, S. Banh, S. Kim, and L. Riek, "Somebody that I used to know: The risks of personalizing robots for dementia care," *Proceedings of We Robot*, 2021.

[5] N. Fronemann, K. Pollmann, and W. Loh, "Should my robot know what's best for me? Human–robot interaction between user experience and ethical design," *AI & SOCIETY*, vol. 37, no. 2, pp. 517–533, 2022.

[6] B. A. Yilma, Y. Naudet, and H. Panetto, "Introduction to personalisation in cyber-physical-social systems," in *On the Move to Meaningful Internet Systems: OTM 2018 Workshops: Confederated International Workshops: EI2N, FBM, ICSP, and Meta4eS 2018, Valletta, Malta, October 22–26, 2018, Revised Selected Papers*. Springer, 2019, pp. 25–35.

[7] W. Saeed and C. Omlin, "Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110273, 2023.

[8] 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.

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

[10] J. Leusmann, C. Wang, M. Gienger, A. Schmidt, and S. Mayer, "Understanding the uncertainty loop of human-robot interaction," *arXiv preprint arXiv:2303.07889*, 2023.

[11] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.

[12] M. Matarese, F. Rea, and A. Sciutti, "A user-centred framework for explainable artificial intelligence in human-robot interaction," *arXiv preprint arXiv:2109.12912*, 2021.

[13] C. Conati, O. Barral, V. Putnam, and L. Rieger, "Toward personalized XAI: A case study in intelligent tutoring systems," *Artificial intelligence*, vol. 298, p. 103503, 2021.

[14] S. Arnórsson, F. Abeillon, I. Al-Hazwani, J. Bernard, H. Hauptmann, and M. El-Assady, "Why am I reading this? Explaining personalized news recommender systems," in *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, 2023, pp. 67–72.

[15] R. Wohlrab, M. Vierhauser, and E. Nilsson, "What impact do my preferences have? A framework for explanation-based elicitation of quality objectives for robotic mission planning," in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2024, pp. 111–128.

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

[17] T. Love, A. Andriella, and G. Alenyà, "Towards explainable proactive robot interactions for groups of people in unstructured environments," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 697–701.

[18] A. Mikołajczyk, M. Grochowski, and A. Kwasigroch, "Towards explainable classifiers using the counterfactual approach-global explanations for discovering bias in data," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 11, no. 1, pp. 51–67, 2021.

[19] K. Alikhademi, B. Richardson, E. Drobina, and J. E. Gilbert, "Can explainable AI explain unfairness? A framework for evaluating explainable AI," *arXiv preprint arXiv:2106.07483*, 2021.

[20] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion*, vol. 58, pp. 82–115, 2020.

[21] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable artificial intelligence: objectives, stakeholders, and future research opportunities," *Information Systems Management*, vol. 39, no. 1, pp. 53–63, 2022.

[22] H. Al-Samarraie, S. M. Sarsam, and H. Guesgen, "Predicting user preferences of environment design: a perceptual mechanism of user interface customisation," *Behaviour & Information Technology*, vol. 35, no. 8, pp. 644–653, 2016.

[23] Z. Gong and Y. Zhang, "Behavior explanation as intention signaling in human-robot teaming," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 1005–1011.

[24] L. Herbold, M. Sadeghi, and A. Vogelsang, "Generating context-aware contrastive explanations in rule-based systems," *arXiv preprint arXiv:2402.13000*, 2024.

[25] R. Nimmo, M. Constantinides, K. Zhou, D. Quercia, and S. Stumpf, "User characteristics in explainable AI: The rabbit hole of personalization?" in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–13.

[26] M. Naiseh, N. Jiang, J. Ma, and R. Ali, "Personalising explainable recommendations: literature and conceptualisation," in *Trends and Innovations in Information Systems and Technologies: Volume 2 8*. Springer, 2020, pp. 518–533.

[27] M. Sadeghi, L. Herbold, M. Unterbusch, and A. Vogelsang, "SmartEx: A framework for generating user-centric explanations in smart environments," in *2024 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2024, pp. 106–113.

[28] S. Schömbs, S. Pareek, J. Goncalves, and W. Johal, "Robot-assisted decision-making: Unveiling the role of uncertainty visualisation and embodiment," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–16.

[29] J. Schneider and J. Handali, "Personalized explanation in machine learning: A conceptualization," *arXiv preprint arXiv:1901.00770*, 2019.