HRI for Explainable Robotics

Wafa Johal
School of Computer Science and IS
University of Melbourne
Melbourne, Australia
wafa.johal@unimelb.edu.au

Lina Phaijit
School of Computer Science and IS
University of New South Wales
Sydney, Australia
l.phaijit@unsw.edu.au

Fethiye Irmak Doğan
Div. of Robotics, Perception and Learning
KTH Royal Institute of Technology
Stockholm, Sweden
fidogan@kth.se

Aaquib Tabrez
Department of Computer Science
University of Colorado Boulder
Boulder, CO, USA
mohd.tabrez@colorado.edu

Maartje de Graaf
Information & Computing Sciences
Utrecht University
Utrecht, The Netherlands
m.m.a.degraaf@uu.nl

***Abstract*—With the new wave in artificial intelligence (AI) powered by machine learning and novel algorithms, more complex robotic systems are being designed to function around humans. With this, the question of trust has gained attention in the past decade in the human-robot interaction community. While trust is a central aspect of a collaborative relationship, it has been found that transparency of the robot's reasoning and decision process were crucial predictors of trust. Transparency, explainability, interpretability, legibility, etc. all these concepts aim to provide a way for the robot to gain trust by making the internal model of the robot less opaque to the human user. This workshop will develop a multidisciplinary vision for defining, designing, and evaluating HRI for Explainable Robotics.**

***Index Terms*—human-robot interaction, explainable robotics, interpretability, decision making**

I. Overview

Explainable AI (XAI) is a growing and broad field [1]. The advances in applications of deep learning systems in critical areas such as health or defense, have made explanations essential for the users to trust and judge appropriately the outcome of the AI system. But while in XAI, the focus of researchers has been to build tools to generate explanations about the system's decision making, in robotics, the decision may not be reflected in the observable action taken by the robot (see Figure 1). Indeed, robots are *physical* autonomous agents which reduces the relevance of the information presented by XAI as a tool for achieving effective communication between the autonomous agent and the human user [2]. This complexifies the idea of explainable robotics (XAR) and necessitates to offer transparency for the different components of robotic systems (e.g., perception, decision, and action) [3]. The implications for explaining and what needs to be stated are numerous. Foremost, the explaining must be done by the robot itself [4] because, in terms of their embodiment, robots serve as the interaction partners themselves. Explanation in interaction must take place inside the boundaries of the interaction itself, placing restrictions on how actions can be explained.

How can HRI answer this need for explainability? How is explainability defined/perceived from the HRI point of view?

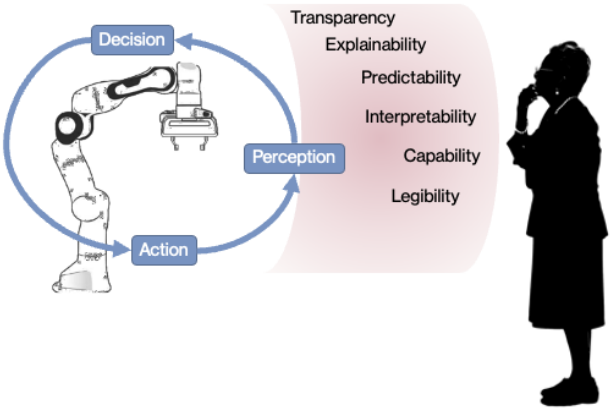

Fig. 1. Illustration of the workshop's topic

How can we design HRI systems and interfaces to allow explainability? How to design user-centered subjective and objective measures to evaluate robot's explanations? How do the concepts of trust, interpretability, predictability, transparency, intelligibility, understandability, and comprehensibility relate to each other? In the end, the field of XAR encompasses more than just AI and robotics because designing user interfaces that foster appropriate user trust in robotics systems is a genuine human-robot interaction challenge.

The first part of the workshop aims at defining the scope of actions for HRI research in designing explainable robotic systems. After working on defining the concepts around explainability for robotics systems, we will discuss the design space. This second part aims at investigating the possible modalities to provide explanations and to discuss how HRI can support this with novel interfaces. The last part aims to discuss subjective and objective metrics that could be used to evaluate explainability from a user-centered point of view.

II. Organizers

Main organizer **Wafa Johal** is a Senior Lecturer at Faculty of Engineering and IT, Wafa obtained her PhD from the University of Grenoble (France) focusing on bodily signals in child-robot interaction and affective reasoning for social agents. Her research aims at creating acceptable and useful assistive robot interactions using social signal sensing, affective and cognitive reasoning and natural expressivity.

Lina Phaijit is a PhD candidate in the Faculty of Engineering at UNSW. Her research interests lie at the intersection of human-centred robot learning and augmented reality.

Fethiye Irmak Doğan is a Ph.D. candidate in the Robotics, Perception, and Learning (RPL) division at KTH Royal Institute of Technology. She received her B.Sc. and M.Sc. degrees in Computer Engineering from Middle East Technical University, Turkey. Her research interests include Human-Robot Interaction, Machine Learning, and Computer Vision.

Aaquib Tabrez is a Ph.D. student in the Collaborative AI and Robotics Lab at CU. His research focuses on developing and operationalizing novel explainable AI techniques within human-robot teaming scenarios, primarily focusing on human-robot communication, reinforcement learning, and mental modeling to improve trust, transparency, and fluency among teammates.

Maartje De Graaf, the main organizer of a similar workshop at HRI 2018 [5], is an Assistant Professor of Human-Computer Interaction at Utrecht University. As a Communication Scientist, her research focuses on peoples affective, behavioral, and cognitive responses to robots aiming for the development of socially acceptable robots. She obtained her PhD in Communication Science and Human-Robot Interaction (2015) investigating the long-term acceptance of social robots in home environments.

III. Workshop Overview

We propose to organize a half-day workshop on the topic of *HRI for Explainable Robotics*, partially as a follow-up of the *Explainable Robotic Systems* workshop organized at HRI 2018 [5]. The workshop will be highly interactive with oral presentations of accepted papers, keynote speakers, and groups discussion around the sessions' themes: *Defining Explainability in HRI*, *Designing HRI solutions for Explainability*, and *Evaluating Explainability in HRI*. The workshop will have:

- **Keynotes**: invited senior researchers will share their perspectives and experiences on the field. They will be recorded to allow all participants watch them on-demand.
- **Interactive Sessions**: Each session will aim to look at the participants' work with a certain angle, and will start with a short intro on the topic followed by lighting talks and group discussion.

Provisional schedule

Depending on the registered participants, we will align the schedule to the adequate timezones if the workshop runs in a hybrid fashion. We also investigate the possibility of having the workshop in several instances to suit several time zones. The keynote times will be fixed, but the rest is kept open depending on participants.

TABLE I
Tentative of Schedule

10min	Introduction and Welcome
35min	Keynote 1: Explainable AI, *Speaker TBC*
	Session 1 - Defining Explainability in HRI
5min	introduction
15min	contributed short talks
20min	group discussion
15min	Break
	Session 2 - Designing HRI for Explainability
5min	introduction
15min	contributed short talks
20min	group discussion
	Session 3 - Evaluation of Explainability in HRI
5min	introduction
15min	contributed short talks
20min	group discussion
35min	Keynote 2: Perspectives for Explainability in HRI, *Speaker TBC*
20min	Wrap-up and Conclusion

A. Target Audience and Approach for recruiting participants

We invite participants to report previous or planned research, practice and interest in developing solutions for explainability in human-robot interaction. Researchers from a variety of backgrounds will be invited to contribute and participate, including but not limited to HRI, robotics, computer science, psychology, communication science, and cognitive science. The workshop will be advertised by sending a call-for-interest on robotics and technology for learning mailing lists (e.g., HRI-ANNOUNCEMENT) and using social networks. Participants to the Winter School of Embodied AI[1] will also be encouraged to participate, and participants of the previous workshop on *Explainable Robotics Systems* at HRI 2018 [5] will be contacted individually to submit their newest work to the workshop.

B. Schedule

- Call for interest 30 April 2023
- Submission deadline 10 July 2023
- Notifications 1 August 2023

Participants will be invited to submit extended abstracts (2-4 pages, excluding references) in PDF using the IEEE/RO-MAN two-column template, and will be peer-reviewed based on their originality, relevance, technical soundness, and clarity.

C. Plan for documenting the workshop

The proceedings of the workshop will be made available on our website, https://x-hri.github.io/ Depending on the quality of submissions, a special issue will be proposed to the journal Frontiers Robots and AI Research.

[1] https://hriwinterschool.com/

D. List of topics

- Explainability / Transparency in HRI
- Trust in HRI
- Legibility / Predictability in HRI
- Metrics to evaluate robot's explanation, interpretability and/or predictability
- Theories and methods for designing explainable embodied systems
- Modalities of explanations in HRI
- Impacts of explanation in HRI
- Technical innovation in designing HRI tools for explainable robotics
- Autonomous explanation generation in HRI
- Adaptive explanation generation in HRI
- Explainable planning and decision-making in HRI
- Ethical quandaries and considerations for designing explainable robotic systems

IV. Acknowledgements

This research is partially supported by the Australian Research Council Discovery Early Career Research Award (Grant No. DE210100858).

References

[1] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, p. eaay7120, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.aay7120

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

[3] 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.

[4] M. M. De Graaf and B. F. Malle, "How people explain action (and autonomous intelligent systems should too)," in *2017 AAAI Fall Symposium Series*, 2017.

[5] M. M. de Graaf, B. F. Malle, A. Dragan, and T. Ziemke, "Explainable robotic systems," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 387–388.