Explainability for Human-Robot Collaboration

Elmira Yadollahi
KTH Royal Institute of Technology
Stockholm, Sweden
elmiray@kth.se

Marta Romeo
Heriot-Watt University
Edinburgh, UK
m.romeo@hw.ac.uk

Fethiye Irmak Dogan
KTH Royal Institute of Technology
Stockholm, Sweden
fidogan@kth.se

Wafa Johal
University of Melbourne
Melbourne, Australia
wafa.johal@unimelb.edu.au

Maartje De Graaf
Utrecht University
Utrecht, Netherlands
m.m.a.degraaf@uu.nl

Shelly Levy-Tzedek
Ben-Gurion University of the Negev
Beer Sheva, Isreal
shelly@bgu.ac.il

Iolanda Leite
KTH Royal Institute of Technology
Stockholm, Sweden
iolanda@kth.se

ABSTRACT

In human-robot collaboration, explainability bridges the communication gap between complex machine functionalities and humans. An active area of investigation in robotics and AI is understanding and generating explanations that can enhance collaboration and mutual understanding between humans and machines. A key to achieving such seamless collaborations is understanding end-users, whether naive or expert, and tailoring explanation features that are intuitive, user-centred, and contextually relevant. Advancing on the topic not only includes modelling humans' expectations for generating the explanations but also requires the development of metrics to evaluate generated explanations and assess how effectively autonomous systems communicate their intentions, actions, and decision-making rationale. This workshop is designed to tackle the nuanced role of explainability in enhancing the efficiency, safety, and trust in human-robot collaboration. It aims to initiate discussions on the importance of generating and evaluating explainability features developed in autonomous agents. Simultaneously, it addresses various challenges, including bias in explainability and downsides of explainability and deception in human-robot interaction.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Collaborative and social computing*; • **Computer systems organization** → **Robotics**; • **Applied computing** → *Psychology*; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Explainable Robotics, XAI, Human-Centered Robot Explanations

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3638154

ACM Reference Format:
Elmira Yadollahi, Marta Romeo, Fethiye Irmak Dogan, Wafa Johal, Maartje De Graaf, Shelly Levy-Tzedek, and Iolanda Leite. 2024. Explainability for Human-Robot Collaboration. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3610978.3638154

1 INTRODUCTION

Powered by the new generation of Artificial Intelligence (AI) algorithms, robots are becoming evermore autonomous, making decisions on collaborative actions to perform and how to socially interact with their users. The complexity of these AI and Machine Learning (ML) based systems calls for innovative ways to make them transparent and understandable to their users. The AI community has directed many efforts in this direction by starting a new field of research called Explainable AI (XAI), mainly focused on making the decisions taken by neural networks as intelligible as possible [12] to advance systems' intrinsic functioning [3, 10]. However, when the AI system is deployed on a physical platform in environments that are shared with humans, the challenges surrounding explainability deepen and transform [7].

Human partners have expectations and mental models of the agents they interact with, often ascribing them human-like characteristics and behaviours [2]. In this context, classical XAI solutions for explainability might not be effective in providing the desired transparency. The problem is exacerbated when an unexpected event, or an event that is perceived as unexpected, occurs during the execution of a joint task [1]. The ability to provide explanations and transparency during a human-robot interaction is linked to trust [5] and rapport building that ultimately leads to higher acceptability of robotics solutions [14] and more effective human-robot collaborations [4, 11]. Providing explanations is a social act, most of the time involving communication [6] in varying forms, e.g., textual/visual [13], or multi-model [8].

What makes a good explanation, when and how to deliver it, and how to evaluate its effectiveness and understanding are pressing open problems that this workshop sets out to investigate. In addition, we aim to understand and define the nuances between several

concepts associated with explainability, including transparency, interpretability [9] and legibility [4]. By gathering experts in human-robot interaction and participatory design, AI and explainability, we aim to identify the major challenges in explainability for human-robot collaboration and discuss ways to bridge the existing gaps. This workshop builds on the success of the X-HRI workshop[1] held at ROMAN 2023.

2 WORKSHOP OVERVIEW

In this workshop, we want to gather and share insights across the wider HRI community on the pitfalls and opportunities of understanding and implementing explainability. The objective of the workshop is to systematically discuss (1) what are the challenges that HRI researchers are facing to date when it comes to explainability for human-robot collaborations; (2) how multidisciplinarity and participatory design may help navigate different solutions for the identified challenges; (3) what metrics researchers in HRI could use to validate how explainable and transparent their systems are; (4) how can the definitions of different concepts associated with explainability translate to HRI scenarios. From the conversations that will arise on these points during the workshop, we aim to write a report summarising some future work action points.

2.1 Main Topics of Interests

Topics of interest include, but are not limited to:

- Using participatory design to achieve explainability
- The downsides of explainability
- The connection between explainability and trust
- What makes an interaction explainable
- Metrics to evaluate explainability
- Deception in Human-Robot Interaction
- Unintended biases in explainability & how to deal with those
- Transparency in Trustworthy autonomous systems
- Explanation generation as a model reconciliation process
- Adapting explanation through forming a mental model
- Explanation generation
- Ethics

2.2 Target Audience and Solicitation Plan

The call for participation will be distributed via mailing lists (i.e., hri-announcement, chi-announcements, robotics-worldwide) and social media. To encourage the participation of a multidisciplinary and diverse audience (expected 30 participants), we will also advertise the workshop to mailing lists such as corpora-list (aimed at researchers working on NLP), CVML (aimed at researchers working on computer vision and machine learning), HMC (aimed at researcher working on human-machine communication), the Women in Machine Learning mailing list, and the Women in Robotics Slack channel. A website (based on https://sites.google.com/view/x-hri/home) will be created to provide information about the workshop. Here, prospective participants will find the format and schedule of the workshop, as well as information on the keynotes. Prospective participants will be invited to submit 2-4 pages of extended abstracts on research related to the topics described above. We will encourage the submission of papers describing work in progress or preliminary results to encourage the discussion and peer review of new ideas. All manuscripts will be peer-reviewed based on originality, relevance, technical soundness, and clarity. At submission time, as discussion starters during the workshop day, we will gauge prospective participants to provide us with what they believe the biggest challenge in explainability for human-robot collaboration is.

[1] https://sites.google.com/view/x-hri

2.3 Invited Speakers

We have secured two keynote speakers and one backup speaker working with topics closely related to the workshop scope. The keynote presentations last 20 minutes, followed by 10 minutes of Q&A.

(1) Sonia Chernova, Georgia Tech (tentatively confirmed)
(2) Tathagata Chakraborti, IBM Research AI (tentatively confirmed)

2.4 Format and Activities

This is a half-day hybrid workshop. To encourage participation and improve the accessibility of the workshop, it will run in person and over Zoom. This way, speakers and participants who cannot travel to the conference will still be able to present and participate in the discussion. We will support remote participation in the discussions by appointing a member of the organising committee to monitor the chat, encourage remote participants to engage in the conversation and perform in-person tasks on their behalf when needed. Participants of the workshop who submitted an extended abstract will be given the possibility to present it to the rest of the group first, as a lightning talk and then as a poster during the coffee break to encourage further discussion and networking. The interactive activity will start after the second keynote. Participants will be divided into four small groups, each under the guidance of one of the organisers. The topic to be discussed in each group will be chosen based on participants' submissions and keynote presentations. All participants will take part in all groups' discussions by rotating every 25 minutes across the tables. The rotation will be organised so that each participant will participate in each discussion as part of a different group of people. This is done as a way to encourage participants with different interests and different backgrounds to engage with each other and with all the topics. This will foster new discussions and serve as an inspiring activity that leverages each participant's expertise. The last 20 minutes of activity will open the discussion to the full group, where each group can highlight their discussions and share their takeaways.

2.5 Structure

09:00 - 09:05 Welcome and introduction to the workshop
09:05 - 09:20 Ice-breaking activity
09:20 - 09:50 Keynote (20 + 10 Q/A): **Sonia Chernova**
09:50 - 10:10 Lightening talks
10:10 - 10:40 coffee and poster
10:40 - 11:10 Keynote (20 + 10 Q/A): **Tathagata Chakraborti**
11:10 - 12:45 Activity
12:45 - 13:00 Concluding remarks

3 DOCUMENTATION

The workshop's outcomes will be documented on our website, https://sites.google.com/view/x-hri/home and will include submitted papers, abstracts, as well as event logs generated from the discussions. Furthermore, we are planning to organize a special issue on the topics related to explainability and transparency in human-robot collaboration in a related journal. All accepted authors and other participants will be invited to contribute.

4 ORGANIZING COMMITTEE

Elmira Yadollahi is a Postdoctoral fellow at KTH Royal Institute of Technology, Sweden and obtained her PhD in Robotics and Computer Science from École polytechnique fédérale de Lausanne (EPFL), Switzerland and Instituto Superior Técnico, Portugal. Her research tackles novel approaches that enhance the transparency and efficiency of the collaboration between humans and machines. She is an associate editor of the International Journal of Child-Computer Interaction (IJCCI) and has served in several conferences.

Marta Romeo is an Assistant Professor in Computer Science at Heriot-Watt University, Edinburgh (UK). She received her PhD in socially assistive robotics for independent living within the H2020 MoveCare project at the Cognitive Robotics Lab at the University of Manchester (UK). Her current research, within the UKRI TAS Node on Trust, focuses on investigating how trust between humans and robots can be built, maintained, and recovered.

Fethiye Irmak Doğan is a Postdoctoral researcher in the Robotics, Perception, and Learning (RPL) division at KTH Royal Institute of Technology, Sweden. She received her PhD degree in Computer Engineering from KTH in 2023. Her research interests include Human-Robot Interaction, Machine Learning, and Computer Vision. Her current research focuses on endowing robots with the ability of semantic reasoning and scene understanding for autonomously operating them in human environments.

Wafa Johal is a Senior Lecturer and an ARC DECRA Fellow at the University of Melbourne, Australia. 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 an acceptable and useful assistive robot that can learn and teach using multi-modal interaction and cognitive reasoning. She is an Associate Editor of RA-L, and she served in various organisational roles (for conferences and workshops). She recently received the Women in AI Award in the Education category for the Asia-Pacific Region (2023).

Maartje De Graaf is an Assistant Professor of Human-Computer Interaction at Utrecht University, Netherlands. As a Communication Scientist, her research focuses on people's affective, behavioural, 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, Twente University, Netherlands) investigating the long-term acceptance of social robots in home environments. She is on HRI Steering Committee, Associate Editor of THRI, and co-organized 20+ workshops at HRI, RO-MAN, ICSR, including two previous workshops on Explainability.

Shelly Levy Tzedek is an Associate Professor and head of the Cognition, Aging & Rehabilitation Laboratory at Ben Gurion University, Israel. She studied Biomedical Engineering at UC Berkeley and at MIT, and her research focuses on the use of robotic technology for rehabilitation.

Iolanda Leite is an Associate Professor at the School of Electrical Engineering and Computer Science at KTH Royal Institute of Technology, Sweden. She holds a PhD in Information Systems and Computer Engineering from IST, University of Lisbon. Her research goal is to develop robots that can perceive, learn from, and respond appropriately to people in real-world situations.

ACKNOWLEDGMENTS

At KTH, this workshop is partially funded by grants from the Swedish Foundation for Strategic Research (SSF FFL18-0199), the S-FACTOR project from NordForsk, Digital Futures, and the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications. Dr Johal is supported by the Australian Research Council Discovery Early Career Research Award (Grant No. DE210100858). Dr Romeo's contribution is supported by the UKRI TAS Node on Trust (Grant EP/V026682/1).

REFERENCES

[1] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan Explanations as Model Reconciliation: Moving beyond Explanation as Soliloquy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (Melbourne, Australia) *(IJCAI'17)*. AAAI Press, 156–163.

[2] Maartje MA De Graaf and Bertram F Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

[3] Fethiye Irmak Doğan, Gaspar I. Melsión, and Iolanda Leite. 2023. Leveraging explainability for understanding object descriptions in ambiguous 3D environments. *Frontiers in Robotics and AI* 9 (2023). https://doi.org/10.3389/frobt.2022.937772

[4] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. 2013. Legibility and predictability of robot motion. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 301–308.

[5] Mark Edmonds, Feng Gao, Hangxin Liu, Xu Xie, Siyuan Qi, Brandon Rothrock, Yixin Zhu, Ying Nian Wu, Hongjing Lu, and Song-Chun Zhu. 2019. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics* 4, 37 (2019).

[6] Zhao Han, Elizabeth Phillips, and Holly A. Yanco. 2021. The Need for Verbal Robot Explanations and How People Would Like a Robot to Explain Itself. *J. Hum.-Robot Interact.* 10, 4, Article 36 (sep 2021), 42 pages. https://doi.org/10.1145/3469652

[7] Marco Matarese, Francesco Rea, and Alessandra Sciutti. 2021. A user-centred framework for explainable artificial intelligence in human-robot interaction. *arXiv preprint arXiv:2109.12912* (2021).

[8] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8779–8788. https://doi.org/10.1109/CVPR.2018.00915

[9] Ornnalin Phaijit, Claude Sammut, and Wafa Johal. 2023. User Interface Interventions for Improving Robot Learning from Demonstration. In *Proceedings of the 9th International Conference on Human-Agent Interaction*. 31–38.

[10] Ramprasaath Ramasamy Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. 2019. Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Vol. 2019-Octob. IEEE, 2591–2600. https://doi.org/10.1109/ICCV.2019.00268 arXiv:1902.03751

[11] Mohan Sridharan and Ben Meadows. 2019. Towards a Theory of Explanations for Human–Robot Collaboration. *KI-Künstliche Intelligenz* 33, 4 (2019), 331–342.

[12] Giulia Vilone and Luca Longo. 2020. Explainable Artificial Intelligence: a Systematic Review. arXiv:2006.00093 [cs.AI]

[13] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable Embodied Agents Through Social Cues: A Review. *ACM Transactions on Human-Robot Interaction* 10, 3 (July 2021), 27:1–27:24. https://doi.org/10.1145/3457188 arXiv:2003.05251 [cs.RO]

[14] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 408–416. https://doi.org/10.1145/2909824.3020230