Causal Explanations for Robot Decisions and Beliefs

Tamlin Love[*]
tlove@iri.upc.edu
Institut de Robòtica i Informàtica
Industrial (CSIC-UPC)
Llorens i Artigas 4-6, 08028,
Barcelona, Spain

Antonio Andriella
aandriella@iri.upc.edu
Institut de Robòtica i Informàtica
Industrial (CSIC-UPC)
Llorens i Artigas 4-6, 08028
Barcelona, Spain

Guillem Alenyà
galenya@iri.upc.edu
Institut de Robòtica i Informàtica
Industrial (CSIC-UPC)
Llorens i Artigas 4-6, 08028
Barcelona, Spain

Abstract

Explainability is an important tool for human-robot interaction (HRI). By explaining its decisions and beliefs, a robot can promote understandability and thereby foster desiderata such as trust, acceptance and usability. However, HRI domains pose challenges to automatic explanation generation. In such domains, a robot must consider the causal reasons for behaviour embedded in temporal sequences of decisions, all while factoring in noise and uncertainty inherent to these kinds of domains. Additionally, as explainability itself constitutes a human-robot interaction, it is important for robots to be able to properly interpret user questions and effectively communicate explanations in order to improve understanding. In our work, we address these challenges from a causal perspective, developing methods that use causal models in order to automatically generate causal, counterfactual explanations in HRI domains. We also produce some insights into embedding such a system in a human-robot interaction in order to maximise understandability.

CCS Concepts

• **Computing methodologies** → **Reasoning about belief and knowledge**; • **Human-centered computing** → *Human computer interaction (HCI).*

Keywords

Explainability, Human-Robot Interaction, Causal Models, Counterfactual Explanations

ACM Reference Format:
Tamlin Love, Antonio Andriella, and Guillem Alenyà. 2026. Causal Explanations for Robot Decisions and Beliefs. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3776734.3794634

1 Introduction

The ability for systems to explain the reasons for their decisions, has been deemed critically important by researchers [17, 20] and policy makers [7]. This is especially true in human-robot interaction (HRI) scenarios, where improved understandability can enhance factors such as trust, usability, and collaborative performance [9, 22].

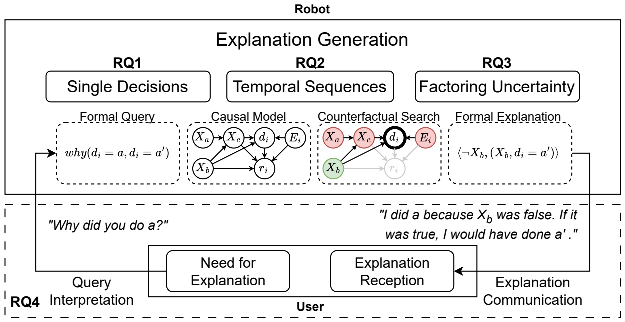

Figure 1: An overview of our explainability approach, illustrating how we generate explanations from user queries.

There are many different approaches for automatically generating explanations in explainable artificial intelligence (XAI) [21] and explainable robotics [19] contexts, relying on different formulations of explanations. We follow the formulation provided by Miller [17], who identifies that explanations are inherently contrastive, and answer questions of the form "Why X and not Y?", where X is a real event and Y is a hypothetical event. Pearl and Mackenzie [18] have argued that answering such "Why" questions requires counterfactual reasoning grounded in notions of causality. By maintaining a causal model of an environment, counterfactual scenarios can be examined by intervening on the model, and thus, causes for particular decisions or events can be determined. Some work has taken place in using causal models to generate explanations in both XAI [3, 4] and robotics [5] contexts, but the field remains open to further research.

Robotics and HRI domains present unique challenges for explainability. For one, robot executions represent a temporal sequence of decisions, and thus a robot should be able to explain its behaviour in terms of past decisions and states. Additionally, robot decisions factor in several sources of uncertainty [11, 25], such as noisy observations, hidden state variables (where human internal states are not observable directly), and uncertain outcomes of physical and social actions. Finally, the robot's explainability system needs to be situated within an HRI, where the robot must estimate which information to provide to a user in response to a query [16], and how best to communicate that information to maximise understandability [2].

Given these challenges, our research seeks to address the following research questions:

- **RQ1** - How can we produce accurate, causally-grounded explanations of robot decisions in HRI environments?
- **RQ2** - How can this approach be extended to complex temporal sequences of robot decisions and beliefs?

[*] Corresponding author.

HRI Companion '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2321-6/2026/03
https://doi.org/10.1145/3776734.3794634

- **RQ3** - How can the inherent uncertainty present in HRI be incorporated into explanation generation approaches?
- **RQ4** - How can such a explanation generation be embedded into an HRI to maximise the system's understandability?

2 Methodology and Results to Date

Here we divide our methodology and current work by research question. A complete overview is provided in Fig. 1.

RQ1: Explaining Single Decisions - Our initial work focused on adapting causal, counterfactual explanation generation (e.g. [5]) to unstructured HRI domains [12]. We considered a scenario where a robot was placed in a public space and proactively elicited interactions from passers-by. We devised a two-level perception-decision pipeline. In layer 1, the robot used pose estimations to calculate the engagement of each person with the robot, via features such as distance and gaze direction. In layer 2, those features are used to make decisions regarding eliciting an HRI. We modelled the pipeline using a causal model, with the novel contribution that model subgraphs could be dynamically added/removed based on the number of people detected. A counterfactual search is then performed to generate templated explanations of the form "I did X because R. If instead R^*, then I would have done Y". This system was subsequently deployed in a user study "in the wild" [13], where the robot successfully generated explanations automatically and in real time, in order to measure their effects on understandability (see RQ4). These initial results demonstrated that causal, counterfactual explanations could be effectively deployed in HRI environments.

RQ2: Explaining Temporal Behaviour - Many real-world HRI scenarios require robots to execute intricate sequences of actions. To represent such sequential decision-making within a causal model, we turned to Behaviour Trees (BTs) as a well-suited control architecture [10]. The novelty of our approach is that we automatically construct a causal model (representing decisions, node executions, return statuses and environment states) directly from the structure of a BT and domain knowledge in the form of a causal model of the state [14]. This model can then be instantiated using episodic memory and queried for explanations similarly to our previous approaches [12, 13]. While there are other methods for producing explanations from BTs [8, 23], our method is capable of generating causal, contrastive explanations, which we show to be accurate in a simulated cognitive exercise use case [14].

RQ4: Explainability as an Interaction - When conceptualising explainability as an HRI, we can divide it into a few stages [2, 16], some of which are addressed by our work (see Fig. 1). Firstly, the robot must interpret a user's question (*query interpretation*). In [14] we provide a formal representation of a contrastive "Why A and not B?" query that can be used to initiate a counterfactual search of a causal model. In a work under review [15], we employ an LLM-based orchestrator for selecting an appropriate explanation module given a user's query and other relevant context (e.g. task execution statuses). Evaluated in a home assistance robot domain, our approach successfully selected the correct component in 99.4% of execution-query pairs.

Secondly, the robot must communicate the explanation's contents in a human-understandable format (*explanation communication*). Most of our current work has leveraged template-based approaches [12–14], but we have also investigated using an LLM to convert a formal explanation to natural language [15]. By generating the explanation using a causal model and then converting the formal explanation to natural language, we maintain a higher explanation accuracy than an end-to-end LLM-based approach.

We have also explored how explanations affect understandability (*explanation reception*). In a systematic review [6], we examined how understandability is measured in HRI. The review reveals a lack of standardisation in measuring understandability across all types of measures and a lack of longitudinal studies on the temporal effects of explanations on understandability. We found only one instance of "in the wild" evaluations — our previous work addressing RQ1 [13].

3 Future Work

RQ3: Factoring Uncertainty - Given the uncertainty in robotic domains and decision-making [11, 25], we have identified three ways in which such uncertainty should be factored into explanation generation. Firstly, explanations should be able to express the effect of uncertainty on decision-making or beliefs. For example, if the robot selects one path over another because it is more confident in the first path being clear, this confidence should be expressed in the explanation. Secondly, even if uncertainty is not directly factored into decision-making, explanations should be able to express that the robot's beliefs may not correspond with the user's understanding of the environment (e.g. "I *believed/thought* you were frustrated"). By addressing the perceptual belief problem in this way [24], a robot can not only help a user locate discrepancies in its world model but also foster a shared understanding of the situation. Finally, uncertainty should be factored into counterfactual scenarios (e.g. "If I had fetched a pizza instead of the salad, you *most likely* would have been happier"), allowing the robot to express the uncertainty that arises from stochastic processes in the causal model itself. In all these cases, extending the approach we developed for RQ1 and RQ2 to incorporate probability (e.g. using a causal Bayesian network) may prove to be successful, although this introduces the problem of correctly estimating the conditional probability distribution. One potential solution may lie in learning the distribution through simulation [1, 5].

RQ4: Enhancing Explainability Interaction - While we have taken some steps towards addressing RQ4, more work can be done to better situate explanation generation in an HRI. We propose investigating LLM-based methods by which natural language questions can be converted into formal contrastive queries. By allowing a robot to interpret natural language in this way we can ensure smoother and more natural explanatory interactions. Such a natural language interface will allow us to conduct a user study in which users engage in prolonged, interactive explanatory sessions, enabled by the temporal explanations produced in [14] and incorporating uncertainty as discussed for RQ3. By conducting such a study, we hope to better understand how users form a complete understanding of complex robot behaviour in HRI settings.

Acknowledgments

This work has been supported by the European Union through Horizon Europe Marie Skłodowska-Curie grant agreement No. 101072488 (TRAIL) and by PID2023-152259OB-I00 CHLOE-MAP funded by MCIU/ AEI / 10.13039/ 501100011033, ERDF and UE.

References

[1] Antonio Andriella, Carme Torras, and Guillem Alenyà. 2019. Learning robot policies using a high-level abstraction persona-behaviour simulator. In *International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–8. doi:10.1109/RO-MAN46459.2019.8956357

[2] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: results from a systematic literature review. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.

[3] Nils Ole Breuer, Andreas Sauter, Majid Mohammadi, and Erman Acar. 2024. CAGE: Causality-aware Shapley value for global explanations. In *World Conference on Explainable Artificial Intelligence*. Springer, 143–162. doi:10.1007/978-3-031-63800-8_8

[4] Ricardo Miguel de Oliveira Moreira, Jacopo Bono, Mário Cardoso, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. 2024. DiConStruct: Causal Concept-based Explanations through Black-Box Distillation. In *Causal Learning and Reasoning*. PMLR, 740–768. doi:10.48550/arXiv.2401.08534

[5] Maximilian Diehl and Karinne Ramirez-Amaro. 2022. Why did I fail? A causal-based method to find explanations for robot failures. *Robotics and Automation Letters* 7, 4 (2022), 8925–8932. doi:10.1109/LRA.2022.3188889

[6] Ferran Gebelli, Pradip Pramanick, Tamlin Love, Raquel Ros, Anais Garrell, Silvia Rossi, Antonio Andriella, and Guillem Alenya. 2025. Measuring User Understanding in Explainable Human-Robot Interaction: A systematic Review. (2025). doi:10.5281/zenodo.15838566

[7] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. 2022. Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine* 17, 1 (2022), 72–85. doi:10.1109/MCI.2021.3129960

[8] Zhao Han, Daniel Giger, Jordan Allspaw, Michael S Lee, Henny Admoni, and Holly A Yanco. 2021. Building the foundation of robot explanation generation using behavior trees. *Transactions on Human-Robot Interaction* 10 (2021), 1–31. doi:10.1145/3457185

[9] 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. doi:10.3389/fcomp.2023.1096257

[10] Matteo Iovino, Edvards Scukins, Jonathan Styrud, Petter Ögren, and Christian Smith. 2022. A survey of behavior trees in robotics and AI. *Robotics and Autonomous Systems* 154 (2022), 104096. doi:10.1016/j.robot.2022.104096

[11] Jan Leusmann, Chao Wang, Michael Gienger, Albrecht Schmidt, and Sven Mayer. 2023. Understanding the Uncertainty Loop of Human-Robot Interaction. *arXiv preprint arXiv:2303.07889* (2023). doi:10.48550/arXiv.2303.07889

[12] Tamlin Love, Antonio Andriella, and Guillem Alenyà. 2024. Towards explainable proactive robot interactions for groups of people in unstructured environments. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 697–701. doi:10.1145/3610978.3640734

[13] Tamlin Love, Antonio Andriella, and Guillem Alenyà. 2024. What would I do If. . . ? Promoting understanding in HRI through real-time explanations in the wild. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 504–509. doi:10.1109/RO-MAN60168.2024.10731403

[14] Tamlin Love, Antonio Andriella, and Guillem Alenyà. 2025. Temporal Counterfactual Explanations of Behaviour Tree Decisions. *arXiv preprint arXiv:2509.07674* (2025). doi:10.48550/arXiv.2509.07674

[15] 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). doi:10.48550/arXiv.2601.03070

[16] 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). doi:10.48550/arXiv.2109.12912

[17] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38. doi:10.1016/j.artint.2018.07.007

[18] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

[19] Fatai Sado, Chu Kiong Loo, Wei Shiung Liew, Matthias Kerzel, and Stefan Wermter. 2023. Explainable goal-driven agents and robots-a comprehensive review. *Comput. Surveys* 55, 10 (2023), 1–41. doi:10.1145/3564240

[20] Rossitza Setchi, Maryam Banitalebi Dehkordi, and Juwairiya Siraj Khan. 2020. Explainable robotics in human-robot interactions. *Procedia Computer Science* 176 (2020), 3057–3066. doi:10.1016/j.procs.2020.09.198

[21] Timo Speith. 2022. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *ACM conference on fairness, accountability, and transparency (FAccT)*. 2239–2250. doi:10.1145/3531146.3534639

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

[23] Christian Tagliamonte, Daniel Maccaline, Gregory LeMasurier, and Holly A Yanco. 2024. A Generalizable Architecture for Explaining Robot Failures Using Behavior Trees and Large Language Models. In *Companion of the International Conference on Human-Robot Interaction*. ACM/IEEE, 1038–1042. doi:10.1145/3610978.3640551

[24] Sam Thellman and Tom Ziemke. 2021. The perceptual belief problem: Why explainability is a tough challenge in social robotics. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 3 (2021), 1–15. doi:10.1145/3461781

[25] Sebastian Thrun. 2002. Probabilistic robotics. *Commun. ACM* 45, 3 (2002), 52–57. doi:10.1145/504729.504754

Received 30 October 2025; accepted 18 December 2025