HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230099

Towards Robots That Meet Users' Need for Explanation

Sonja STANGE [a,1], Stefan KOPP [a]
[a] *Faculty of Technology, CITEC, Bielefeld University*
ORCiD ID: Sonja Stange https://orcid.org/0000-0001-5696-101X, Stefan Kopp https://orcid.org/0000-0002-4047-9277

Abstract. While humans are used to reason about other humans' behavior, they are not readily able to understand the decision processes of artificial agents. This can be harmful in human-robot interaction (HRI) settings where a user may suspect erroneous or, even worse, intentionally non-cooperative behavior, resulting in reduced acceptance of the robot. In order to mitigate such negative effects, autonomous robots may be equipped with the ability to adequately explain their behavior. To that end, a robot is required to have the ability to (1) robustly detect a user's need for explanation and (2) identify the situation-specific nature of the explanation need. Further it needs to be endowed with (3) communicative capabilities in order to deliver suitable explanations and ensure sufficient understanding. This extended abstract presents recent work towards endowing a social robot with such qualities and discusses how robots can meet users' explanation needs more adequately.

Keywords. explainable robots, XAR, explainability, HRI, transparency

1. Introduction

Many reasons have been put forth to offer explanations of A.I. systems' decisions, e.g., to solve a lack of transparency, convey knowledge to users, help developers to debug, or increase user trust towards and thus acceptability of A.I. systems. Researchers have thus started to develop methods for finding the best possible explanations of systems' decisions, leading to the field of explainable artificial intelligence (XAI). The advances in this field, however, have also shown that one explanation does "not fit all" [1], but rather that explanations need to be crafted to the application scenario, the recipient and the concrete explanatory goal [2]. As a result, different subfields are starting to emerge such as the field of explainable autonomous robots (XAR). Such human-robot interaction (HRI) settings often focus on human lay users who have a tendency to anthropomorphize robots and to treat them as social interaction partners. Thus, expectations to generate explanations adapted to users' explanatory needs are voiced increasingly, also taking into account insights from the social sciences. However, a clear path to meeting users' need for explanation in HRI has not yet emerged and there are several open questions as to how a robot can (1) reliably detect a users' need for explanation, (2) identify the nature of this need, and (3) then adequately address it with a suitable explanation.

[1] Corresponding Author: Sonja Stange, sstange@techfak.uni-bielefeld.de.

2. Problems and Challenges

Detecting Explanation Need in HRI Generally, by providing an explanation, an explainer, who is in possession of information, conveys relevant parts of this information to a recipient, the explainee [3]. Most research in the field of explainable AI and human-robot interaction looks at this with the motivation to increase transparency of and trust in the artifact's functioning [4]. In social interaction settings, in contrast, a behavior explanation does not only aim for mere understanding but also manages a relationship by influencing how an interaction partner is perceived [5]. This is supported by the finding that humans explain events that are perceived as surprising and negative [6], e.g. trying to justify a behavior and positively influence how it is perceived. De Graaf & Malle [7] present findings that humans explain robot behavior similar to human behavior, and thus argue for taking perceived intentionality, surprisingness and desirability of the robot behavior into account.

While humans may have an intuitive understanding of which behavior they expect to be negatively surprising, the task of detecting such a need for explanation is a complex issue for social robots as it strongly depends on the context the behavior is executed in and dynamically evolves within an interaction. A straightforward approach to detect an explanation need was implemented by Koeman et al. [8] who implemented a graphical user interface that offers buttons for users to request explanations. A less reliable but also less disruptive approach is to equip a robot with a cognitive model of the user's understanding and offer explanations in case of an estimated lack of it. For instance, Chakraborti et al. [9] equip a robot with a mental model of the user's understanding of its behavior which allows the robot to plan its path in a human-aware manner. This way the robot can either choose a path that matches the human's mental model or choose a path that diverges from users' expectations and offer an explanation in order to update the user's model. Further, a robust understanding of humans' explanation needs can benefit from taking into account user feedback. One intuitive means for humans to indicate a need for explanation are verbal explanation requests [10], other approaches propose to access users' explanation need via dialogue [11].

Understanding Explanation Need in HRI Once a need for explanation is detected, a robot needs to be able to identify the nature of the explanation need and thus figure out what about its own behavior may need to be explained. This means that robots need to be endowed with a self-situation awareness in order to classify their behavior in the current interaction setting. Instead of giving a complete set of reasons, an explainer selects a set of reasons that is most useful to the explainee [10]. In this, verbal explanation requests hold particular epistemic value: one way to break down a user's explanatory need is to carefully consider the question asked, which informs about the explainee's knowledge gap and thus provides insight into what may be helpful information [10]. In addition to enabling correct identification of the aspect of behavior that calls for an explanation, the robot's self model further needs to enable access to the reasons and thus be structured in an interpretable manner [12].

Addressing Explanation Need in HRI The question of how to adequately meet a particular users' explanation need is again context-dependent and different explainers have different means for explanation at their disposal. While disembodied intelligent systems use written text or imagery, embodied agents and especially humanoid robots have the

advantage of being able to use social cues such as speech and movement to explain system decisions [13]. Moreover, using natural language explanations is advantageous in HRI settings as it enables smooth integration in the interaction and intuitive communication such as verbal explanation requests [14]. For this, a robot needs to be able to map decisive factors in its decision process to verbal explanation. Based on findings that people explain agent behavior similarly to human behavior, a prominent choice for explainable agents and robots is to verbalize selected reasons in line with belief-desire-intention principles [15,4]. Lastly, the robot needs to have basic dialogue abilities to ensure mutual understanding.

3. Towards Robots that Meet Users' Need for Explanation

In previous and recent work, we have started to address these challenges of meeting users' need for explanation in HRI.

Detecting the Need for Explanation In order to assess whether humans' need for robot behavior explanations is comparable to the findings in human-human interaction, we conducted an empirical investigation of the effects of a social robot's behavioral self-explanations depending on behavioral attributes such as perceived intentionality, surprisingness and desirability of the robot behaviors [16]. Explanatory success was assessed in terms of an increase of users' understandability and desirability of robot behaviors. In an online video study, participants ($N = 97$) watched a set of six robot behavior videos, evaluated as surprising in a pre-study, paired with verbal behavior explanations. While understandability was significantly increased for all behaviors except the most understandable one, desirability ratings were increased to a statistically significant extent for three out of four behaviors that were previously as undesirable, while not having a statistically significant impact on desirable behaviors. The robot's explanations were thus particularly helpful for negatively perceived behaviors which suggests that people's strategy to explain intentional and observable behaviors that are surprising and negative [6] may be transferable to human-robot interaction. Based on the view of the process of explaining as a dialogue, and considering the prospect of deducing not only when an explanation is needed but also what type of information a user might be missing, an explanation dialogue model that offered explanation as a response to users' requests was developed and employed as part of an explainable interaction architecture [17]. An interaction study revealed that, even though specifically instructed to do so, overall, participants were hesitant to request explanations [17].

Identifying the Need for Explanation To be able to classify a user's explanation need regarding its own behavior in retrospect, in our explainable interaction architecture [17] the decision process was structured in an interpretable manner. The robot was equipped with internal needs that dynamically change over time and, in combination with external influencing factors such as a user entering the room, led to the autonomous selection of certain behavioral strategies (driving towards the user / the charger). Strategies consisted of low-level actions which were represented as behavior trees (BTs) [18]. The factors that led to the initiation of a strategy were saved as a decision snapshot in the robot's self model, enabling access of this information in case of an explanation request. Which reasons to select for the explanation was deduced from users' natural language explana-

tion requests. In the first implementation this was realized as a differentiation between what- and why-requests [17]. While this first implementation only enabled explanation of the currently active behavior, a user study revealed the necessity to refer to past behaviors [17]. In order to match the selection of reasons with the explainee's knowledge gap more specifically, a more elaborate reference resolution component was introduced, which detects temporal adverbial and verb constraints in the syntactical dependency tree of utterances, executes a query in an episodic memory, and then scores the resulting entries to find the referred behavior [19]. For this purpose, the robot's self model and, more specifically, its episodic memory were equipped with a graph database that stores and queries representations of the internal execution. This enables inference of reasons for past events, as well as access to and thus explanation of failed strategies [19].

Addressing the Need for Explanation In order to communicate explanations in a suitable manner, we decided to give natural language explanations that build on folk-theory of how people explain intentional behavior. After confirmation of positive effects of five explanation types in an empirical study [16] we developed an explanation dialogue model that incorporated the pre-evaluated explanation types and step-wise adaptation of explanation strategies according to user requests [17]. This model was integrated in the robot's dialogue management and enabled the robot to verbalize its decision process in terms of intention or action explanations. In case of subsequent elaboration requests, it added a more complex, causally structured explanation. Additional to these content specific communicative capabilities, the robot was equipped with basic dialogue abilities such as repetition of misunderstood utterances and saving information in the context which enable grounding, repair and feedback and to ensure reciprocal understanding (cf. [17]).

4. Discussion and Implications

While the contributions described here present first achievements in enabling a robot to meet users' explanation need, they also have demonstrated the complexity thereof (for a more detailed analysis see [20]). In order to further extend robots' abilities to meet users' explanatory needs, not only perception-related advances (speech recognition, face perception) are indispensable. At best, robots' explanation capabilities could not only account for users' explanation need at one specific moment in the interaction, but rather adapt to user preferences at a number of levels. An explanation situation constitutes a complex interplay of the participants and their relationships to each other and the explanandum in an explanation situation. Within this construct, explanatory depth could, for instance, develop over time, or according to the relationship between user and robot. Further, explanation initiative (reactive vs. proactive) and timing (before or after execution of a behavior) could be varied based on user preferences or explanandum attributes (noise, disturbance, first execution). Essentially, this calls for the development of a framework that incorporates more fine graded definition of the determinants of an explanation situation, and careful consideration of the influence and interplay thereof. For this, findings on explanatory preferences need to be carefully categorized based on who explains what to whom and with which aim, and how this plays out in HRI. Only then, predictions about the applicability of insights from human-human interaction and XAI in HRI can be made and interdependencies detected. Simultaneously, feasibility to deploy such a framework in the complex task of autonomous and explainable behavior generation should be given consideration continually and validated in actual human-robot interaction studies.

References

[1] Sokol K, Flach P. One explanation does not fit all: The promise of interactive explanations for machine learning transparency. KI-Künstliche Intelligenz. 2020;34(2):235-50.

[2] Rosenfeld A, Richardson A. Explainability in human–agent systems. Autonomous Agents and Multi-Agent Systems. 2019 11;33(6):673-705. doi:10.1007/s10458-019-09408-y.

[3] Lewis D. Causal Explanation. In: Lewis D, editor. Philosophical Papers Volume II. Oxford University Press; 1986. p. 214-40. doi:10.1093/0195036468.003.0007.

[4] Anjomshoae S, Najjar A, Calvaresi D, Främling K. Explainable Agents and Robots: Results from a Systematic Literature Review. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. AAMAS '19. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems; 2019. p. 1078–1088.

[5] Malle BF. How the Mind Explains Behavior: Folk Explanation, Meaning and Social Interaction. 6. MIT Press; 2004.

[6] Malle BF, Knobe J. Which behaviors do people explain? A basic actor–observer asymmetry. Journal of Personality and Social Psychology. 1997;72(2):288-304. doi:10.1037/0022-3514.72.2.288.

[7] de Graaf MMA, Malle BF. People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. In: Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI '19). IEEE; 2019. p. 239-48. doi:10.1109/HRI.2019.8673308.

[8] Koeman VJ, Dennis LA, Webster M, Fisher M, Hindriks K. The "Why Did You Do That?" Button: Answering Why-Questions for End Users of Robotic Systems. In: Dennis LA, Bordini RH, Lespérance Y, editors. Engineering Multi-Agent Systems - 7th International Workshop. Cham: Springer International Publishing; 2020. p. 152-72. doi:10.1007/978-3-030-51417-4$_8$.

[9] Chakraborti T, Sreedharan S, Grover S, Kambhampati S. Plan Explanations as Model Reconciliation – An Empirical Study. In: Proceedings of the 14th ACM / IEEE International Conference on Human-Robot Interaction (HRI '19). IEEE; 2019. p. 258-66. doi:10.1109/HRI.2019.8673193.

[10] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence. 2019;267:1-38. doi:10.1016/j.artint.2018.07.007.

[11] Madumal P, Miller T, Sonenberg L, Vetere F. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19); 2019. p. 1033-41. doi:10.48550/arXiv.1903.02409.

[12] Malle BF, Scheutz M. In: Bendel O, editor. Learning How to Behave. Wiesbaden: Springer Fachmedien Wiesbaden; 2018. p. 1-24. doi:10.1007/978-3-658-17484-2$_1$7 – 1.

[13] Wallkötter S, Tulli S, Castellano G, Paiva A, Chetouani M. Explainable Embodied Agents through Social Cues: A Review. ACM Transactions on Human-Robot Interaction. 2021 7;10(3):1–24. doi:10.1145/3457188.

[14] Sado F, Loo CK, Kerzel M, Wermter S. Explainable goal-driven agents and robots-a comprehensive review and new framework. arXiv. 2020;180. doi:10.48550/arXiv.2004.09705.

[15] Harbers M, Van Den Bosch K, Meyer JJC. A study into preferred explanations of virtual agent behavior. In: Ruttkay Z, Kipp M, Nijholt A, Vilhjálmsson HH, editors. Intelligent Virtual Agents. vol. 5773 of Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, Berlin, Heidelberg; 2009. p. 132-45. doi:10.1007/978-3-642-04380-2$_1$7.

[16] Stange S, Kopp S. Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior. In: Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI '20); 2020. p. 619-27. doi:10.1145/3319502.3374802.

[17] Stange S, Hassan T, Schröder F, Konkol J, Kopp S. Self-Explaining Social Robots: An Explainable Behavior Generation Architecture for Human-Robot Interaction. Frontiers in Artificial Intelligence. 2022 4;5(4):1-19. doi:10.3389/frai.2022.866920.

[18] Colledanchise M, Ögren P. Behavior trees in robotics and AI: An introduction. CRC Press; 2018. doi:10.1201/9780429489105.

[19] Schröder F, Stange S, Kopp S. Resolving References in Natural Language Explanation Requests about Robot Behavior in HRI. In: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI'23 Companion); 2023. doi:10.1145/3568294.3579981.

[20] Stange S. Tell Me Why (and What)! Self-Explanations for Autonomous Social Robot Behavior [Ph.D. thesis]; 2022. doi:10.4119/unibi/2967737.