Towards Speech-based Interactive Post hoc Explanations in Explainable AI

STEFAN HILLMANN, TU Berlin, Germany
SEBASTIAN MÖLLER, TU Berlin, and German Research Center for Artificial Intelligence (DFKI), Germany
THILO MICHAEL, TU Berlin, Germany

CCS Concepts: • **Human-centered computing → Natural language interfaces**.

Additional Key Words and Phrases: XAI, Explainable AI, post hoc explanations, spoken dialog, argumentation

1 INTRODUCTION

AI-based systems offer solutions for information extraction (e.g., finding information), information transformation (e.g., machine translation), classification (e.g., classifying news as fake or true), or decision support (e.g., providing diagnoses and treatment proposals for medical doctors) in many real-world applications. The solutions are based on machine leaning (ML) models and are commonly offered to a large and diverse group of users, some of them experts, many others naïve users from a large population. Nowadays and in particular deep neural network architectures are black-boxes for users and even developers [1, 4, 9] (also cp. [6]). A major goal of Explainable Artificial Intelligence (XAI) is making complex decision-making systems more trustworthy and accountable [7, p. 2]. That is why XAI seeks to ensure transparency, interpretability, and explainability [9].

Common to most users is that they are not able to understand the functioning of the AI-based systems, i.e., those are perceived as black-boxes. Humans are confronted with the results, but they cannot comprehend what information was used by the system for reaching this result (interpretability), and in which way this information was processed and weighted (transparency). The underlying reason is that an explicit functional description of the system is missing or even not possible in most Machine-Learning-(ML)-based AI systems – the function is trained by adjusting the internal parameters, and sometimes also the architecture is learned from a basic set of standard architectures. However, natural language and speech-based explanations allow better explainability [1, p. 11] due to interactive post hoc explanations in from of an informed dialog [7, p. 2]. Additionally, AI is also addressed by regulations, e.g., of the EU [2, 3], and thus becomes even more relevant for industry and research. Here, not at least recognition of bias in AI systems' decision plays an important role.

2 WHAT DO WE KNOW ALREADY?

XAI has attracted much attention to increase trust in AI-based systems, and finally to enable acceptance. Here, our perspective on trust follows Ribeiro's et al. [8] approach. They differentiate two different aspects of trusting in an (AI) model: "(1) trusting a prediction" (or decision), and "(2) trusting a model, i.e. whether the user trusts a model to behave in reasonable ways if deployed". Trusting in a prediction is the crucial aspect for a user affected by the decisions of an AI system. Trust means, the user trusts to a sufficient degree in the (factual) correctness of the model's decision [8]. Especially when interacting with an AI system in a human-like manner (e.g., by natural language), trust is strongly related to the user's perceived competence of an AI system. Systems able to provide comprehensible explanations will potentially show a higher competence than systems just providing decisions or predictions, without any explanation.

Authors' addresses: Stefan Hillmann, TU Berlin, Berlin, Germany, stefan.hillmann@tu-berlin.de; Sebastian Möller, TU Berlin, and German Research Center for Artificial Intelligence (DFKI), Berlin, Germany; Thilo Michael, TU Berlin, Berlin, Germany.

1st AI-Debate Workshop © OvGU, Editors; DOI: 10.25673/38473

XAI frequently focuses on what information is encoded in a model's intermediate representation. One way to do this inspection is through different kinds of probing tasks that enable to study the behavior of the AI model towards different input. Another way is to use saliency methods that show activation of model features when providing a certain output. The saliency maps can be applied to the input space, thus highlighting which input information has served to what degree in providing the desired output.

Evaluations in the context of textual explanations show that automatically generated explanations can increase human understanding, trust and confidence in the AI system for certain tasks [5]. However, whereas graphical highlighting might be helpful for experts, it comes to its limits when more complex relationships are to be analyzed. Such complex relationships could better be presented via natural language, which is our common way to express relationships between entities, and to pronounce judgments. Natural explanations are contrastive, selective and social [7] and argumentation-based dialogs are the most suitable for this purpose [10, p. 14]. We thus advocate for explanations about the behavior of AI-based systems which are generated in the form of natural language, and preferably using speech.

Let's take the example of a decision support system for medical doctors, which provides diagnosis support for kidney transplant patients. The doctor would have to find and justify his own diagnosis; knowing the reasons why the system proposes a certain diagnosis (e.g., the physiological parameters, comparable cases, risk factors, etc.) would enable him to judge his own and the system's diagnosis, and to better consider all relevant information for the decision. Another example is an AI-based system that identifies fake news on a public news channel: The system might provide reasons on why it judges certain news to be fake, e.g., by citing contradicting trustworthy information, highlighting the source of the fake news, or others.

A speech-based system could provide such explanations in a rather unobtrusive way. It could be triggered by the user who would like to know more about the basis for the decision. The user could ask a rather general question ("What makes you confident that this information is fake?") or a more specific one ("Did the author of this news article already spread fake news?" or "Which political party does he belong to?").

3 HOW DO WE STUDY THE PHENOMENON?

Research on this topic requires empirical analysis with a range of systems showing different degrees of performance, and offering different ways of providing explanations. The explanations could be given via text or speech; they could be given one-shot, as a user-driven question-answer dialog, or via a more system-initiative conversation. Explanations could include external sources of information which could be distinguished via voice characteristics, and use prosodic features, e.g., to gradually express confidence . Ultimately, speech and natural language could be combined with other modalities to enable multimodal explanations, adapted towards the user's needs and usage situation.

With such systems, comparative empirical studies should be carried out, comparing different system versions with different ways of generating explanations. The empirical studies should be performed with a representative group of users of the target application (if possible), as the users' needs for explanations might duffer according to their background knowledge.

4 WHAT WE WOULD LIKE TO KNOW?

The aim of this research is to enable AI systems which provide helpful and comprehensible explanations to the users. A proper form and of post hoc explanations could help to increase trust in the system, and finally its acceptance. Speech-based systems are particularly adequate for this purpose, as speech characteristics could be dedicatedly manipulated to increase explainability.

5 WHAT DO I WANT TO LEARN FROM DIFFERENT DISCIPLINES?

For appropriate generation (form and degree of detail) of explanations we count input from cognitive psychology and psychology of explanations as well as argumentation (as part of NLP). Social sciences are important concerning acceptance beyond pure performance of AI-systems (cp. [1, 7]). AI and XAI are needed for extraction of information provided in conversational explanations.

6 WHAT DO WE WANT TO TEACH OTHER DISCIPLINES?

Knowledge about how speech and language technology can be used to provide (and evaluate) explanations in AI will open a powerful tool to the (X)AI research community. We expect that knowledge on speech technology, dialog design, argumentation and natural language generation will enable better explanations for different types of applications.

REFERENCES

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[2] European Commission. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council (GDPR).

[3] European Commission. 2021. Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.

[4] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable Artificial Intelligence. *Science Robotics* 4, 37 (Dec. 2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[5] Piyawat Lertvittayakumjorn and Francesca Toni. 2019. Human-Grounded Evaluations of Explanation Methods for Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5194–5204. https://doi.org/10.18653/v1/D19-1523

[6] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[7] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 279–288. https://doi.org/10.1145/3287560.3287574

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[9] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. 2020. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* 8 (2020), 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

[10] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and Explainable Artificial Intelligence: A Survey. *The Knowledge Engineering Review* 36 (2021), e5. https://doi.org/10.1017/S0269888921000011