Semantic coordination in conversational explanations of predictive models: Preliminary findings

Alexander Berman
Centre for Linguistic Theory and Studies in Probability (CLASP)
Dept. of Philosophy, Linguistics and Theory of Science
University of Gothenburg
alexander.berman@gu.se

Abstract

Most work in explainable artificial intelligence (XAI) focuses on causes for predictions from machine-learning (ML) based models. In other words, XAI typically aims to address questions such as "Why does the model predict that I have a high risk of developing heart disease?". To investigate what kinds of explanations that humans actually communicate in ML-related contexts, we collect human dialogues revolving around model predictions. A preliminary analysis reveals that causes for predictions is indeed a common topic, but that the *meaning (or nature) of target labels* is even more frequent. This finding suggests that conversationally explainable AI systems may need to be able to "teach" human users the meaning of words or expressions that they use and to repair potential problems related to semantic coordination of such words.

1 Introduction

Most research in explainable artificial intelligence (XAI) focuses on causes for predictions (see e.g. (Barredo Arrieta et al., 2020; Miller, 2019)). In principle, this enables XAI methods to address questions such as "Why does the model predict that I have a high risk of developing heart disease?" (request for local explanation of specific prediction) or "How does the model predict risk of developing heart disease?" (request for global explanation of how the model generally makes its predictions). However, not much work has studied what kinds of explanations that humans actually communicate in scenarios involving predictive or statistical modeling. Previous work has collected dialogues where the explainer is a dialogue system (Kuźba and Biecek, 2020) or a researcher acting as the system (Hernandez-Bocanegra and Ziegler, 2021), as well as explanatory dialogues that do not specifically involve statistical estimates (Moore and Paris, 1993; Madumal et al., 2019; Alshomary et al., 2024; Fisher et al., 2023; Götze and Schlangen, 2023). As far as we are aware, no previous work has collected explanatory dialogues revolving around model predictions, with human participants in both roles.

2 Data collection

Our web-based experiment (Berman and Howes, 2022) collects human dialogues about model predictions of personality traits (openness, extraversion, etc.) from music preferences. Firstly, participants listen to 30-second excerpts of 10 tracks and rate them on a 4-point hedonic scale (like/dislike slightly/very much). In a second part, participants are paired up with each other and are randomly assigned the role of either explainee or explainer. They then chat with each other using an interface where explainers, but not explainees, are given access to prediction results (estimated personality traits), information about the statistical model, descriptions of personality traits, global and local feature contribution plots, and feature values (plots of the explainee's music preferences). In a third part, participants are once again paired up with each other, but this time in opposite roles.

The experiment does not involve any personal data such as participant's names. Participants were recruited via various channels such as the university's web page, newsletters, posters at campus, and social media.

3 Preliminary results

A preliminary analysis of 27 collected dialogues reveals that causes for predictions is a fairly common explanandum category. For example, in one of the dialogues, an explainee utters: "I really want to know what these results are based on...why am I so low on openness? kind of disagree with that". However, the *meaning/nature of target label* is an even more frequent topic. This latter kind of explanatory exchange constitutes a form of *semantic coordina-*

tion (Larsson, 2008; Larsson and Myrendal, 2017), where the meaning of specific words or expressions form a topic of conversation. Observed strategies for initiating coordination include signaling non-understanding ("I think about the word agreeablenes, don't know what to think about that"), inquiring about implications ("Is a lower score on agreeableness a negative quality to have?"), raising explicit word meaning questions ("what do 'agreeableness' mean?") and self-initiation (explaining target labels unpromptedly). Strategies for repairing or apprehending coordination problems (i.e. explaining target labels) include copying content from the web interface ("Openness to experience describes a dimension of cognitive style that distinguishes imaginative, creative people from down-to-earth, conventional people."), elaborating implications ("You seem to be a person that seldom end up in conflicts, that is easy to do business with", "Looks like your openness score would make you both creative and down to earth"), and referring to a higher-order concept ("Do you know about the OCEAN scale?").

Below is an example of an excerpt where an explainee (A) and explainer (B) together coordinate the meaning of one of the target labels:

A: I think about the word agreeablenes, don't know what to think about that :)
B: You probably are not so concerend with working together with other people eithr
A: Maybe not, but I do every day and have done so for many years
B: It means cooperation with others and concern with social harmony

4 Discussion and future work

The collected data indicates that meaning of target labels is an important topic in human conversations about model predictions. Although the collected dataset is small and only concerns a single task, it does not seem far-fetched to expect similar findings for other tasks and in other domains, at least when explainees are not domain experts. The finding also resonates with the frequently cited XAI question bank (Liao et al., 2020), which includes "What does the system output mean?" as one of the prototypical questions that explainees may want to ask about an AI.

The results suggest that conversationally explainable AI systems may need to be able to "teach" their human users the meaning of terms or expressions used by the system and to repair potential coordination problems that emerge during the interaction. Future work will need to investigate how an AI system might explain meaning of target labels without producing presupposition violations that could potentially give users an inaccurate mental model of the system's capabilities. For example, if an AI system explains what agreeableness means, this might presuppose that a predictive model trained on tabular data has such knowledge, when in fact the model only learns correlations from linguistically unlabeled data distributions. In other words, to avoid presupposition violations, the linguistic design of such explanations may need to be carefully crafted (Berman, 2024). Furthermore, interesting challenges are raised by the potential need to semantically coordinate vague and context-sensitive explanatory expressions. For example, if an AI explains a prediction by stating that a person likes *loud* music (Berman, forthcoming), this may raise the question what the system means by "loud" (in this context). Depending on the type of predictive model, it may not be evident how to best answer such a question.

Acknowledgements

This work was supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Milad Alshomary, Felix Lange, Meisam Booshehri, Meghdut Sengupta, Philipp Cimiano, and Henning Wachsmuth. 2024. Modeling the quality of dialogical explanations. *Preprint*, arXiv:2403.00662.

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*, 58:82–115.

Alexander Berman. 2024. Too Far Away from the Job Market – Says Who? Linguistically Analyzing Rationales for AI-based Decisions Concerning Employment Support. *Weizenbaum Journal of the Digital Society*, 4(3).

Alexander Berman. forthcoming. Argumentative Dialogue As Basis For Human-AI Collaboration. In

Proceedings of the Communication in Human-AI Interaction Workshop (CHAI-2024).

Alexander Berman and Christine Howes. 2022. "Apparently acousticness is positively correlated with neuroticism": Conversational explanations of model predictions. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Dublin, Ireland. SEMDIAL.

Josephine B Fisher, Amelie S Robrecht, Stefan Kopp, and Katharina J Rohlfing. 2023. Exploring the semantic dialogue patterns of explanations–a case study of game explanations. In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*, pages 35–46.

Jana Götze and David Schlangen. 2023. "Why Do You Say So?" Dialogical Classification Explanations in the Wild and Elicited Through Classification Games. In *Proceedings of the 27th Workshop on the Semantics and Pragmatics of Dialogue*.

Diana C Hernandez-Bocanegra and Jürgen Ziegler. 2021. Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *CUI 2021-3rd Conference on Conversational User Interfaces*, pages 1–11.

Michał Kuźba and Przemysław Biecek. 2020. What would you ask the machine learning model? Identification of user needs for model explanations based on human-model conversations. In *ECML PKDD 2020 Workshops*, pages 447–459, Cham. Springer International Publishing.

Staffan Larsson. 2008. Formalizing the dynamics of semantic systems in dialogue. *Language in Flux-Dialogue Coordination, Language Variation, Change and Evolution*, pages 1–21.

Staffan Larsson and Jenny Myrendal. 2017. Dialogue acts and updates for semantic coordination. In *Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue*, page 59.

Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1033–1041.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Johanna D Moore and Cécile L Paris. 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. Technical report, University of Southern California, Marina Del Rey Information Sciences Institution.