"Apparently acousticness is positively correlated with neuroticism" Conversational explanations of model predictions

Alexander Berman and **Christine Howes**
Department of Philosophy, Linguistics and Theory of Science
University of Gothenburg
alexander.berman@gu.se, christine.howes@gu.se

Abstract

This paper describes an experiment that collects human dialogues about predictions of participants' personality traits on the basis of their music preferences, and presents preliminary results. This type of data can inform the design of explanatory dialogue systems, and the method can straightforwardly be adapted to other domains and statistical models.

1 Introduction

When machine-learning models inform high-stakes decisions, such as in healthcare, it is important to understand what the models' estimates are based on. Under the umbrella term "explainable AI" (XAI), various techniques have been developed for explaining estimates from models that are otherwise considered opaque, such as deep neural networks. One of the most popular techniques involves constructing a simpler, linear approximation of the prediction to be explained (Ribeiro et al., 2016). However, most work in XAI has primarily targeted machine-learning experts, has not assessed explainability in naturalistic settings, and has not accounted for the interactive nature of human explanations (Miller, 2019; Arya et al., 2019; Weld and Bansal, 2018; Simkute et al., 2021). Specifically, Lakkaraju et al. (2022) report that users of current explanation techniques lack interactivity and conversational possibilities.

This paper presents a method for collecting human dialogues revolving around judgements by statistical models, as a basis for informing the design of explanatory dialogue systems and yielding requirements for XAI techniques. In a similar vein, 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 dialogues that do not specifically involve statistical estimates (Moore and Paris, 1993; Madumal et al., 2019). As far as we are aware, no previous work has collected explanatory dialogues revolving around model predictions to inform the design of XAI, with human participants/informants in both roles.

2 Experiment

Our experiment collects human explanatory dialogues about a model's predictions of personality traits 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 the 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 online text chat interface (see figure 1). Explainers, but not explainees, are given access to prediction results (estimated personality traits), information about the statistical model and what the personality traits mean, global and local feature contribution plots, and feature values (plots of the explainee's music preferences), as well as an interactive exploration enabling the explainer to make predictions for hypothetical feature values.

Since participants are paired up with each other, we avoid known issues of bias when using confederates (Kuhlen and Brennan, 2013), enabling an open-ended investigation. A high level of data protection is achieved by not asking participants about their names or contact information, not logging information that could link data to persons, and by screening collected utterances before storing them.

Tracks are featurised on the basis of 10 audio properties (energy, loudness etc.), and an explainee's ratings are aggregated into a fixed-size vector using weighted averaging. For each big-five personality trait (John et al., 1999), we train a logistic regression model to predict polarity (e.g. introverted or extraverted). As training data, we use listening histories from Last.fm and Spotify, audio features extracted from Spotify API, and psycho-

Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue, August, 22-24, 2022, Dublin.

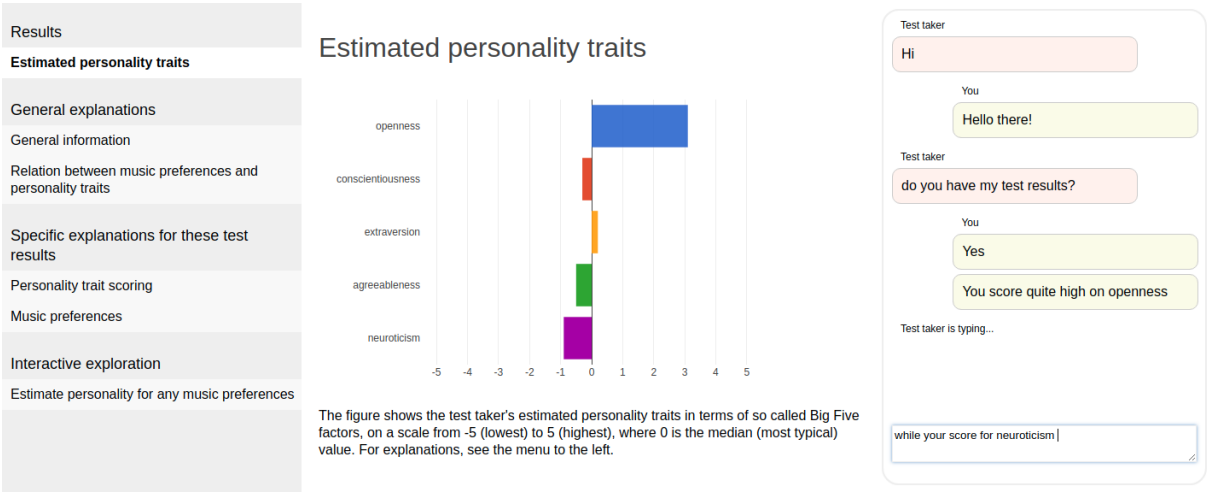

Figure 1: Screenshot of explainer's main view during chat. Explainees only see a chat window (similar to right-most part of explainer's view). Neither the personality prediction or the chat utterances are authentic.

metric test results from the MyPersonality dataset, assembled by Melchiorre and Schedl (2020). Explainers see the log odds of the predictions on a scale from -5 to 5 (see figure 1).

3 Preliminary results

Pilots have been performed with 6 colleagues from the department as participants, resulting in 3 collected dialogues (303 utterances in total). The data encompasses a range of topics including the meaning of labels ("what does agreeableness entail?"), validity of predictions ("conscientiousness is a bit too low I think"), trust ("it is hard to trust these ratings nevertheless"), causation ("I wonder if music influences the personality or if it's only the other way") and the activity as such ("It's a really fun experiment"), as well as different dialogue strategies, exemplified by the two excerpts below (A=explainer, B=explainee):

(1)

A: in terms of the "big five" factors
A: apparently, you are very open
A: almost 5 (out of -5 to 5 where 0 is the median)
B: It's interesting, I wonder what song would give this trait
A: well I actually can tell you something about that I think
A: not which song in particular, but how openness relates to features of the music
B: Oh great I'm interested

(2)

A: um apparently accousticness is positively correlated with neuroticism
B: Haha I'm almost surprised I scored low
B: And openness as well?
A: openness is the opposite with respect to acousticness
A: so I guess if you want to be more open and less neurotic the answer is to develop a preference for acoustic music

These short excerpts demonstrate that explanations given by people for the results provided by the statistical model do not necessarily adhere to the types of explanations usually considered by XAI. In the first excerpt, the explainee seems to target an examplar-based explanation; the explainer offers a correlational explanation instead, which the explainee accepts. The second excerpt exemplifies a logically incomplete explanation (Breitholtz, 2020), drawing on a shared assumption which is not explicitly stated in the dialogues (that being open and non-neurotic is desirable) and that would not necessarily be available to an AI.

4 Future work

In future work, we plan to collect more dialogues with the same setup and perform an analysis of the data. It could also be useful to focus on simpler models – e.g. rule lists or small decision trees – as well as more opaque models such as deep neural nets, with or without the support of a simpler explanation model.

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. Howes was additionally supported by VR 2016-0116, Incremental Reasoning in Dialogue (IncReD).

References

Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv preprint arXiv:1909.03012*.

Ellen Breitholtz. 2020. *Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation*. Brill, Leiden, The Netherlands.

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.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Anna K Kuhlen and Susan E Brennan. 2013. Language in dialogue: When confederates might be hazardous to your data. *Psychonomic bulletin & review*, 20(1):54–72.

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.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:2202.01875*.

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.

Alessandro B. Melchiorre and Markus Schedl. 2020. *Personality Correlates of Music Audio Preferences for Modelling Music Listeners*, page 313–317. Association for Computing Machinery, New York, NY, USA.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Auste Simkute, Ewa Luger, Bronwyn Jones, Michael Evans, and Rhianne Jones. 2021. Explainability for experts: A design framework for making algorithms supporting expert decisions more explainable. *Journal of Responsible Technology*, 7-8:100017.

Daniel S. Weld and Gagan Bansal. 2018. Intelligible artificial intelligence. *CoRR*, abs/1803.04263.