

# Why Do You Think This Person Is Introverted?

1

## *Towards Conversational Argumentative Explainability*

Alexander BERMAN<sup>a</sup>

<sup>a</sup>*Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg*

ORCID ID: Alexander Berman <https://orcid.org/0000-0003-0513-4107>

**Abstract.** This paper presents a system that supports conversational argumentative explainability between AI and human.

**Keywords.** human-AI collaboration, hybrid human-AI intelligence, conversational explainability, argumentation theory, explainable AI

## 1. Introduction

Previous work has shown that users of AI-based decision-support systems prefer conversational forms of explanations, i.e. interfaces that enable users to request and obtain explanations for model predictions in natural-language dialogues between user and system [1]. However, issues regarding how to design, implement and evaluate such interfaces remain relatively unexplored (exceptions include [2,3,4]). One of the challenges that requires more attention concerns *feature contributions*, i.e. how an AI prediction is underpinned by particular features (factors). In everyday conversation, relations between features and judgements are often expressed as rhetorical arguments, as in the utterance “Let’s walk along Walnut Street. *It’s shorter.*” [5, emphasis added], where a route’s short distance is presented as support for its goodness. Typical methods for explainable AI, such as LIME [6] and SHAP [7], only convey *how much* a feature contributes to a prediction (i.e. feature importance), not *how* they contribute. In natural language, this corresponds to merely referring to features by their name: “Based on *distance*, I suggest that we walk along Walnut Street.” (cf. [2,8]). This communicative strategy might be successful in cases where the explainee can easily infer the nature of the contribution. For example, if the explainee already knows that the route via Walnut street is shorter than other options at hand, merely referring to distance may suffice. But if the goal of an AI-based decision-support system is to facilitate hybrid human-AI intelligence [9] – including the possibility to make better decisions with the help of AI – explanations cannot assume that an AI always “reasons” in similar ways as humans. In line with this, one of the aims

---

<sup>1</sup>This is a preview of a demo paper accepted for publication in Proceedings of the Third International Conference on Hybrid Human-Machine Intelligence (2024).

of the prototype described in this paper is to show how feature contributions in the form of more explicit *arguments* can be supported in a dialogue interface between a human and a statistical model.

## 2. Prototype

We demonstrate **MindTone**<sup>2</sup>, a browser-based game featuring conversational argumentative explainability between AI and human. The task of the game is to estimate whether persons are extraverted or introverted based on their music preferences. In each round of the game, the player is shown specific tracks that a person has listened to frequently as well as audio statistics of music heard by the person, such as loudness and energy. The player is assisted by a chatbot that predicts whether the person is extraverted or introverted using a logistic regression model trained on audio statistics (mean values for loudness etc.) as independent variables and degree of extraversion as dependent variable [10]. The dialogue manager is implemented in a rule-based information-state update approach [11] inspired by conversation-oriented semantics [12] and argumentation theory [13,5]. It supports various dialogue strategies including feature contribution arguments, as exemplified by the following dialogue between user (U) and assistant (A):

- A: I think this person is introverted.  
U: Why?  
A: The person likes music with high energy.  
U: I don't understand  
A: Statistically, people that like music with high energy tend to be more introverted.  
U: OK, I see. Why do you think this person likes music with high energy?  
A: Music heard by the person has a higher average score for energy than music in general.  
U: OK, and does the the fact that the person likes music that is not danceable support the assessment that the person is introverted?  
A: Yes.

For natural-language understanding, the system uses GPT-4 [14] to parse user utterances into a formal language for dialogue moves, while templates are used for natural-language generation.

## 3. Future Work

In the future, it would be useful to empirically study how conversational strategies such as feature contribution arguments affect human-AI team performance, e.g. compared to a similar interface that only supports feature importance. Extending the prototype to support model-agnostic explanations (rather than only logistic regression) could also be an interesting avenue for further work.

---

<sup>2</sup>Live demo: <https://github.com/alex-berman/argumentative-explainability>

## 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

- [1] Lakkaraju H, Slack D, Chen Y, Tan C, Singh S. Rethinking explainability as a dialogue: A practitioner's perspective. *arXiv preprint arXiv:220201875*. 2022.
- [2] Slack D, Krishna S, Lakkaraju H, Singh S. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*. 2023;5(8):873-83.
- [3] Berman A, Breitholtz E, Howes C, Bernardy JP. Explaining predictions with enthymematic counterfactuals. In: *Proceedings of the 1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE*. vol. 22; 2022. p. 95-100.
- [4] Feldhus N, Ravichandran AM, Möller S. Mediators: Conversational Agents Explaining NLP Model Behavior; 2022.
- [5] Breitholtz E. Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation. Leiden, The Netherlands: Brill; 2020. Available from: <https://brill.com/view/title/58383>.
- [6] Ribeiro MT, Singh S, Guestrin C. "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*; 2016. p. 1135-44.
- [7] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17*. Red Hook, NY, USA: Curran Associates Inc.; 2017. p. 4768-4777.
- [8] Berman A. Why Does the AI Say That I Am Too Far Away From the Job Market? In: *Weizenbaum Conference" AI, Big Data, Social Media, and People on the Move"*. DEU; 2023. p. 1-11.
- [9] Kamar E. Directions in hybrid intelligence: complementing AI systems with human intelligence. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*; 2016. p. 4070-3.
- [10] Melchiorre AB, Schedl M. Personality Correlates of Music Audio Preferences for Modelling Music Listeners. In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. New York, NY, USA: Association for Computing Machinery; 2020. p. 313-317. Available from: <https://doi-org.ezproxy.ub.gu.se/10.1145/3340631.3394874>.
- [11] Larsson S. Issue-Based Dialogue Management. 2002.
- [12] Ginzburg J. The interactive stance: Meaning for conversation. Oxford University Press; 2012.
- [13] Toulmin SE. The uses of argument. Cambridge university press; 2003.
- [14] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al.. GPT-4 Technical Report; 2024.