Argumentative Dialogue As Basis For Human-AI Collaboration

Alexander Berman[1]

[1]*Dept. of Philosophy, Linguistics and Theory of Science, University of Gothenburg*

Abstract

Argumentation, by which we here mean the ability to give reasons or arguments for a claim, plays a central role in society generally and in collaborative decision-making specifically. However, the role of argumentation in human-AI collaboration and AI-assisted decision-making has received limited attention, despite the widespread interest in "explainable" AI. This paper aims to bridge this gap. First, it is shown that many kinds of AI models are not argumentative in the sense that they do not enable human-AI interfaces to provide reasons or arguments for AI predictions. Second, it is shown that some interpretable AI models encode a knowledge structure that can be harvested for the purpose of supporting argumentative human-AI interaction. Third, a method for extracting such structures from an interpretable model is outlined. Finally, a prototype supporting argumentative dialogue between AI and human user is presented.

Keywords

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

1. Introduction

Argumentation plays a crucial role in society generally and in collaborative decision-making more specifically [1]. By requesting and providing support for claims, we justify our beliefs and actions, and evaluate claims made by others. In the context of artificial intelligence (AI) based on machine learning (ML), it is natural to treat *predictions* made by AI systems as *claims*. For example, if a statistical model predicts that a certain individual is introverted, we can intuitively understand this prediction as a claim. It is then natural to also ask whether the model can provide *arguments* for its claim? In current discourses revolving around AI, this question is typically approached as a matter of *explainability* or *interpretability* (see e.g. [2]). A distinction is often made between black-box models whose predictions and inner workings can only be explained by means of inherently unreliable explanation methods [3], and interpretable models whose logic can in principle be understood by humans [4]. However, from the perspective of human-AI collaboration, the notions of explainability and interpretability are not necessarily crucial in and of themselves. In this paper, we instead hypothesize that human-AI collaboration yield more value when the AI systems can engage in *argumentation*. The main aims of the paper are to briefly discuss the conditions that support argumentative dialogue between a machine learning

HHAI-WS 2024: Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 10—14, 2024, Malmö, Sweden

alexander.berman@gu.se (A. Berman)
0000-0003-0513-4107 (A. Berman)

model and human, and to demonstrate how such a capacity can be conceived theoretically and implemented technically.

2. Theoretical Framework

The present work applies Toulmin's [5] theory of argumentation to ML-based AI.[1] According to this theory, argumentation is an interactive process through which presented claims are challenged and backed. Specifically, a *claim* (e.g. that Sam is introverted) can be backed with *data* (e.g. that Sam doesn't like danceable music). Data support claims by highlighting specific facts or circumstances. Furthermore, the backing of claims by data (often implicitly) rests on *warrants* (e.g. that people that like non-danceable music are generally introverted). While claims and data are specific, warrants are general; their argumentative function is to bridge data and claims by means of e.g. taxonomy (that instances of one category are always instances of another category) or statistics (that instances of one category also tend to be instances of another category). Warrants support conclusions with varying degrees of *force*, signalled linguistically with a *qualifier* (e.g., statistically, people that like non-danceable music are more introverted, and Sam doesn't like danceable music; so, *presumably* Sam is introverted).

3. Argumentation Affordances

The extent to which ML-based AI systems support argumentation differs across different kinds of ML models. Although a comprehensive assessment of this matter is beyond the scope of this paper, some brief remarks can be made. First, it can be noted that black-box models such as deep neural networks and random forests do not afford argumentation in any obvious manner. While claims (e.g. classifiying a person as introverted) can straightforwardly be qualified in terms of confidence (e.g. that the prediction is associated with a probability of 67%), data and warrants cannot readily be identified due to the complex inner workings of these models. To some extent, feature-importance based explanation methods such as LIME [7] and SHAP [8] can be conceived to identify *data*, since they highlight the features that were the most important for a particular prediction. For example, if a neural network predicts that a person is introverted based on the person's music preferences (measured as numerical values for features such as danceability and loudness), LIME may highlight danceability as the most important feature for the prediction at hand. From this information, one can construct the datum "On a scale from 0 to 1, Sam's preference for danceable music is 0.34" as a claim-backing datum. But what kind of warrant supports the conclusion from datum to claim? Has the model learned that people with a preference for danceable music of *exactly 0.34* are generally introverted? Or has it learned something more general, e.g. that introverts prefer music with a danceability value below a certain threshold? These questions cannot be answered by methods such as LIME or SHAP. In reality, a black box may combine preference for danceable music with preference for loud music and other feature values in non-linear and complicated ways that may be difficult or impossible to express in words. In argumentative terms, no warrant can be generated.

[1] For alternative theories of argumentation, see e.g. [6].

For a more interpretable option, we can consider linear additive models such as logistic regression. In contrast to black boxes, these models force features to affect output independently of each other, without any interactions.[2] Furthermore, features affect output monotonically; for example, a stronger preference for danceable music always increases the predicted probability that the person is extraverted. Due to these formal properties, warrants such as "statistically, people that like danceable music are more extraverted" will faithfully reflect the actual knowledge learned by the model. Below, we will formally show how to extract data and warrants for claims obtained from linear additive models.

4. Extracting Arguments From Linear Additive Models

We assume a linear additive model on the form

$$\hat{y} = \beta_0 + \sum_{i=1}^{m} \beta_i X_i$$

where β_0 is the intercept/bias, β_i are the coefficients (both of which are learned when fitting the model to training data), X is the instance (feature values), i denotes feature (e.g. 1 for energy, 2 for danceability, etc.), and $\hat{y}$ is the output (predicted value). We also assume that output and features are standardized continuous variables so that 0 corresponds to mean. Sticking to the example domain above, we say that if the prediction is positive ($\hat{y} > 0$), it is claimed that the person described by X is extraverted, or else introverted. Data supporting a positive claim can then be extracted by listing features with a positive value and a positive coefficient, and features with a negative value and a negative coefficient, since in both of these cases the feature contributes to a positive prediction. Conversely, a negative claim is supported by features with a positive value and a negative coefficient, and features with a negative value and a positive coefficient. A datum can be conveyed linguistically with reference to the feature and its polarity; for example, $X_2 > 0$ can be expressed as "The person likes danceable music".

As warrants, we construct support-relations between the polarity of a coefficient (e.g. $\beta_2 < 0$) and the polarity of the prediction (e.g. $\hat{y} < 0$). This can be expressed linguistically as "Statistically, people that like non-danceable music are more likely to be introverted". The force of each combination of datum and warrant can be defined as the magnitude of the respective addend ($|\beta_i X_i|$). In what follows, we will see how this extraction procedure can form the basis for an argumentative AI communicator.

5. Prototype

We briefly describe **MindTone**[3], a browser-based game featuring argumentative communication 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

---

[2]We here disregard the possibility to use combinations of variables as features.
[3]Live demo: https://github.com/alex-berman/argumentative-explainability

by the person, such as loudness and energy (see figure 1). The player is assisted by a chatbot that predicts whether the person is extraverted or introverted using a logistic regression model trained on audio features and personality traits [9]. The dialogue manager is implemented in a rule-based information-state update approach [10] inspired by conversation-oriented semantics [11]. Specifically, when a session starts, the system makes a claim, e.g. “I think this person is introverted” or “If I had to guess, I’d say that this person is extraverted” (where the qualification reflects the model’s confidence). When the user challenges a claim, the system backs it up with its strongest datum, e.g. “The person likes high-energy music.” When the user indicates not understanding how a datum supports a claim, the system provides a warrant, e.g. “Statistically, people that like high-energy music are more likely to be introverted”. Importantly, such warrants faithfully reflect the actual reasoning process of the model and are not post-hoc approximations. The communicative capabilities are examplified by the following dialogue between user (U) and assistant (A):

A: I think this person is introverted.
U: Why?
A: The person likes high-energy music.
U: I don’t understand
A: Statistically, people that like high-energy music are more likely to be 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 fact that the person likes non-danceable music support the assessment that the person is introverted?
A: Yes.

For natural-language understanding, the system uses GPT-4 [12] to parse user utterances into a formal language for dialogue moves, while templates are used for natural-language generation. Note that while the prototype supports a particular domain (estimation of personality trait from music preferences), both the method for extracting argument structure and the dialogue system are domain-independent and can therefore be applied to any domain of choice.

6. Related Work

The present work can be situated in the context of “conversational explainable AI”, i.e. formalization, implementation and evaluation of systems that can explain AI predictions in a natural-language dialogue between system and human (see e.g. [13, 14, 15, 16, 17, 18, 19]). Typically, previous approaches do not support argumentation of the kind discussed here. For example, the system TalkToModel [18] uses explanation methods such as LIME and SHAP to explain specific predictions by listing features deemed important; for the current domain, it would amount to a phrase such as “the top 2 most important features are (1) energy (2) danceability”. It can be noted that such an explanation provides neither data (whether the person likes how- or low-energy music, etc.) or warrants (how preferences for different aspects of music correlate with extraversion).

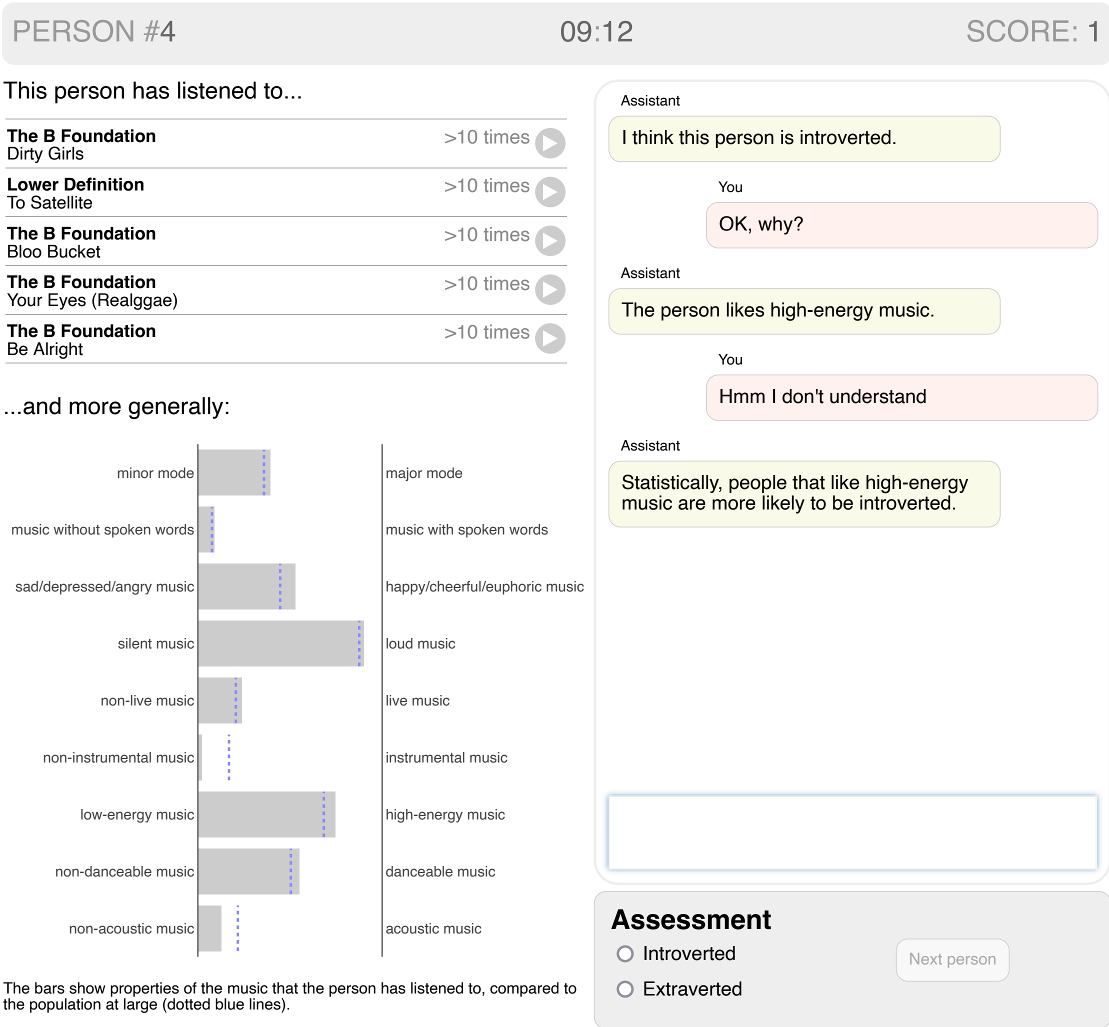

Figure 1: Screenshot of prototype. To the left, a list of tracks frequently listened to by the person whose personality is currently being assessed, followed by a bar graph visualizing audio statistics of music heard by the person. To the right, a chat window for player-AI interaction.

As for dialogue systems with argumentative capabilities, Breitholtz [23] presented a formal account of how claims can be motivated with enthymemes, which, in Tolmin's framework, corresponds to backing claims with data; Maraev et al. [20] later implemented a working prototype on this basis. In contrast to previous work, the approach presented in this paper enables arguments for ML predictions, in the form of both data and warrants, to be communicated by the system.

7. Discussion and Future Work

We have argued that black-box AI systems cannot generate warrants to support their predictions, even in tandem with popular explainability methods such as LIME or SHAP. However, a human user may still identify or produce a warrant to fill the gap. In fact, implicit premises are ubiquitous in human communication and rarely need to be made explicit.[4] Someone says "It's cold in here so let's close the window" and you immediately understand the warrant: closing the window is likely to increase the indoor temperature (assuming that it's colder outside than inside). Since the listener can identify a warrant that makes the speaker's utterance comprehensible, no warrant needs to be verbalized. From this perspective, a lack of warrant-production in human-AI collaboration does not necessarily need to be a problem. But if the purpose of the AI is to enable improved human decision-making [24], one cannot assume that an AI always "reasons" in similar ways as humans. Arguably, potential differences in reasoning between AI and human is precisely the reason why explanations and arguments are needed.

As shown in previous sections, at least some interpretable models afford argumentation. However, to our best knowledge there is no empirical data that supports the hypothesis that argumentation of the kind discussed here benefits human-AI collaboration. (A survey study found that decision-makers prefer interactive explanations in the form of natural language dialogue [25], but did not specifically investigate argumentation.) In fact, previous work suggests that explanations can cause human over-reliance on AI [26] or have no effect on accuracy [27]. However, as far as we can tell, previously evaluated human-AI interactions have not involved argumentative AI systems. We propose two mechanisms through which argumentation might benefit hybrid human-machine decision-making. First, warrants may enable users to assess whether claims are supported by reasonable generalizations. If the system argues that liking music with high energy makes it more likely that one is introverted, and the user finds this correlation generally questionable, then the user can take this into account when assessing the reliablity of the claim that the generalization is intended to support. Second, warrants can potentially make it easier to combine an AI's assessment with the user's own judgement about the case at hand. If the system supports its claim with a statistical generalization, the user can assess to what extent the generalization seems relevant for the case at hand. Sure, you may reason, people that like music with high energy may be more introverted in general, but in this case you know exactly what kind of high-energy music the person listens to, and you don't associate this music with introversion. To the extent that users successfully assess the relevance of the system's generalizations, decision-making accuracy can be improved compared to a scenario without argumentation. In future work, it would be useful to empirically study how an argumentative AI communicator affects human decision patterns in comparison with a non-argumentative interface, and thereby generate data to either support or contradict our claim that argumentation is crucial not only in human communication, but also in communication between humans and AI.

[4] This phenomenon has previously been discussed in terms of e.g. conversational implicature [21], presuppositions [22], and enthymemes [23].

Acknowledgments

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] Mercier H, Sperber D. The enigma of reason. Harvard University Press; 2017.

[2] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020;58:82-115. Available from: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[3] Amparore E, Perotti A, Bajardi P. To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods. PeerJ Computer Science. 2021;7:e479.

[4] Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys. 2022;16:1-85.

[5] Toulmin SE. The uses of argument. Cambridge university press; 2003.

[6] Van Eemeren FH, Grootendorst R, Johnson RH, Plantin C, Willard CA. Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments. Routledge; 2013.

[7] 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.

[8] 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.

[9] 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.

[10] Larsson S. Issue-Based Dialogue Management. 2002.

[11] Ginzburg J. The interactive stance: Meaning for conversation. Oxford University Press; 2012.

[12] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al.. GPT-4 Technical Report; 2024.

[13] Wijekoon A, Wiratunga N, Palihawadana C, Nkisi-Orji I, Corsar D, Martin K. iSee: Intelligent Sharing of Explanation Experience by Users for Users. In: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces. IUI '23 Companion. New York, NY, USA: Association for Computing Machinery; 2023. p. 79–82. Available from: https://doi.org/10.1145/3581754.3584137.

[14] Sokol K, Flach PA. Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant. In: IJCAI; 2018. p. 5868-70.

[15] 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.
[16] Werner C. Explainable AI through Rule-based Interactive Conversation. In: Workshop Proceedings of the EDBT/ICDT 2020 Joint Conference (March 30-April 2, 2020, Copenhagen, Denmark); 2020. .
[17] Kuźba M, Biecek P. What Would You Ask the Machine Learning Model? Identification of User Needs for Model Explanations Based on Human-Model Conversations. In: Koprinska I, Kamp M, Appice A, Loglisci C, Antonie L, Zimmermann A, et al., editors. ECML PKDD 2020 Workshops. Cham: Springer International Publishing; 2020. p. 447-59.
[18] 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.
[19] Feldhus N, Ravichandran AM, Möller S. Mediators: Conversational Agents Explaining NLP Model Behavior; 2022.
[20] Maraev V, Breitholtz E, Howes C, Bernardy JP. Why should I turn left? Towards active explainability for spoken dialogue systems. In: Proceedings of the Reasoning and Interaction Conference (ReInAct 2021); 2021. p. 58-64.
[21] Grice HP. Logic and conversation. In: Speech acts. Brill; 1975. p. 41-58.
[22] Lewis D. Scorekeeping in a language game. Journal of philosophical logic. 1979;8:339-59.
[23] 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.
[24] 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.
[25] Lakkaraju H, Slack D, Chen Y, Tan C, Singh S. Rethinking explainability as a dialogue: A practitioner's perspective. arXiv preprint arXiv:220201875. 2022.
[26] Vasconcelos H, Jörke M, Grunde-McLaughlin M, Krishna R, Gerstenberg T, Bernstein MS. When do XAI methods work? A cost-benefit approach to human-AI collaboration. In: CHI Workshop on Trust and Reliance in AI-Human Teams; 2022. p. 1-15.
[27] Alufaisan Y, Marusich LR, Bakdash JZ, Zhou Y, Kantarcioglu M. Does explainable artificial intelligence improve human decision-making? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35; 2021. p. 6618-26.