information

MDPI

Article

Informing Design and Research Concerning Conversationally Explainable AI Systems by Collecting and Distilling Human Explanatory Dialogues

Alexander Berman * and Christine Howes

Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, 405 30 Gothenburg, Sweden; christine.howes@gu.se
* Correspondence: alexander.berman@gu.se

Abstract

Research into conversationally explainable artificial intelligence (CXAI) aims to emulate the interactive and co-constructive nature of explanations. From the perspective of human-centredness, previous work has shown that AI users prefer conversational explanations over static ones. Various approaches for modelling and implementing CXAI solutions have also been proposed. However, as for concrete dialogue capabilities possessed by such systems, previous approaches have not been properly grounded in analogous dialogue patterns in human–human interaction. The present study bridges this gap in previous work by experimentally collecting human dialogues revolving around AI predictions concerning personality estimation. By distilling the collected interactions into the kind of interactions that would occur if the explainer was a dialogue system, the study identifies dialogue strategies which might be important for CXAI to support. The study reveals that some of the observed strategies—explaining predictions with reference to general rules or patterns and signalling presupposition violations in questions raised by explainees—have received very limited attention in previous work on CXAI. Overall, the study contributes a methodology for empirically identifying CXAI desiderata in human dialogues as well as concrete results with implications for future work.

Keywords: explainable AI; dialogue systems; human-centred explainable AI; conversationally explainable AI

Check for updates

Academic Editors: Christos Troussas, Akrivi Krouska, Phivos Mylonas, Katerina Kabassi, Jaime Caro and Cleo Sgouropoulou

Received: 28 November 2025
Revised: 15 January 2026
Accepted: 22 January 2026
Published: 28 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

When predictions by statistical models inform high-stakes human decisions, such as in healthcare, it can be important for stakeholders to understand the basis on which such predictions are made or how the predictions might be interpreted. Research into so-called "explainable artificial intelligence" (XAI) aims to address such needs. While many efforts in XAI have focused on techniques for obtaining various kinds of explanatory information from opaque machine learning models (see, e.g., [1,2]) or on developing models that follow an interpretable logic (see, e.g., [3–6]), in recent years, more attention has been given to issues concerning how explanations are communicated in human interaction, and how knowledge about this matter can inform design and implementation of XAI systems [7–12]. The increased focus on co-constructive aspects of explanations, where the process of explaining is conceived as inherently social and communicative, manifests itself in a growing body of work which explicitly concerns the prospect of conversationally explainable AI (CXAI) (or dialogue-based XAI) [13–24].

Recent interest in CXAI is largely driven by an insight that for AI explanations to be effective and useful, they need to be aligned with the mechanisms through which humans typically manage and exchange explanations. This development resonates with a broader move in the field from a purely technical focus towards more human-centred XAI (HC-XAI), which emphasises consideration of human stakeholders' needs and perspectives as well as integration of perspectives from cognitive psychology, design theory, and related disciplines (see, e.g., [25–28]).

While the move from "non-interactive" (one-shot) XAI towards CXAI can be conceived as human-centred on an abstract conceptual level, by virtue of its focus on the co-constructive and conversational nature of explanations, this article will argue that existing approaches to designing and developing CXAI solutions lack human-centred approaches to identifying specific dialogue capabilities that it might be important for such systems to possess. The present study aims to fill this gap in previous work. In order to empirically ground CXAI systems in human conversational behaviours, the study collects and qualitatively analyses human explanatory dialogues concerning personality estimation, with the purpose of identifying capabilities that such systems need to possess in order to emulate observed human communicative strategies as well as research challenges associated with emulating human capabilities in CXAI systems. The empirical grounding in human language use is motivated by a human-centred approach to AI development which focuses on actual needs, demands and behaviours of human users [29,30]. While previous human-centred approaches to CXAI have empirically investigated users' needs for conversational AI explanations [10], the extent to which different kinds of dialogue moves contribute to explanatory success [31], as well as the kind of questions that human users might want to ask to an XAI system [32], in terms of concrete dialogue strategies, existing CXAI approaches tend to be based on the researcher's (sometimes implicit) assumptions concerning the kinds of behaviours and capabilities that it would be useful for such systems to exhibit and possess. For example, while the CXAI system TalkToModel [17] handles anaphora and ellipses, enabling the system to contextually resolve co-references in user utterances such as "What do you predict for *them*?", the developers' choice of capabilities to support seems to be based on à priori assumptions rather than explicit insights concerning human explanation strategies. An evident drawback of such an approach is that the researcher's assumptions may not properly reflect actual communicative behaviours of human explainers or explainees, which might limit the potential usefulness and value of systems designed on the basis of such assumptions. It is for these reasons that the proposed approach collects and exploratively analyses human explanatory dialogues with minimal prior assumptions, as will be further elaborated in Section 2.

The article makes the following contributions:

- A methodology is proposed for empirically grounding CXAI in human explanatory communication using dialogue distillation of experimentally collected human dialogues. This methodology extends previous human-centred approaches to CXAI by enabling identification of explanatory dialogue capabilities which might be important for CXAI systems to possess.
- A dataset of 35 collected non-expert dialogues concerning AI-based personality estimation, encompassing a total of 779 utterances, is publicly released.
- A distillation of the collected dialogues identifies 11 different dialogue capabilities used by the participants in the study. Most interestingly in relation to previous work, the study reveals that interlocutors frequently explain predictions with reference to warrants (general rules or patterns [33]), either explicitly or implicitly, a behaviour which is difficult to emulate when targeting opaque models such as deep neural nets and random forests. The study also reveals situations where utterances by explainees

presuppose false information, and where explainers signal these presupposition violations. While identification of presuppositions in user utterances has received limited attention in prior work on CXAI, the study provides empirical evidence of its function in analogous human–human interactions.

The remainder of the article is organised as follows. The experimental procedure and method of data analysis is presented in Section 2. In Section 3, the collected empirical material is analysed and organised into various observed dialogue phenomena. Finally, based on the results of the analysis, Section 4 discusses implications for future research concerning CXAI as well as limitations of the study.

2. Materials and Methods

This section describes the method for collecting and analysing human explanatory dialogues.

2.1. Experimental Setup

The data collection takes the form of a browser-based experiment where participants first listen to 30-s excerpts of 10 music tracks and rate them on a 4-point hedonic scale (dislike very much, dislike slightly, like slightly, like very much). When a participant has rated all tracks (and their ratings contain variation), they can proceed to the second part of the experiment. In the second part, participants are paired up with each other and are randomly assigned the role of either respondent or operator. They then chat with each other using an interface (see Figure 1). Operators are instructed to explain the respondent's test results, while respondents are instructed to ensure that they receive their test results and to try to understand what they are based on. Operators, but not respondents, are given access to prediction results (estimated personality traits), information about the statistical model, definitions of personality traits, local and global feature contribution plots (see Figures 2 and 3), and feature values (plots of the respondent's music preferences; see Figure 4).

Data was collected through a series of 5 trials; the last three of these trials included an additional third part where participants were paired up a second time after completing their first chat, but this time in opposite roles (potentially with another participant than in the first chat). In the last two trials, demographic information (age, gender and educational level) was collected on a voluntary basis after the two chat interactions.

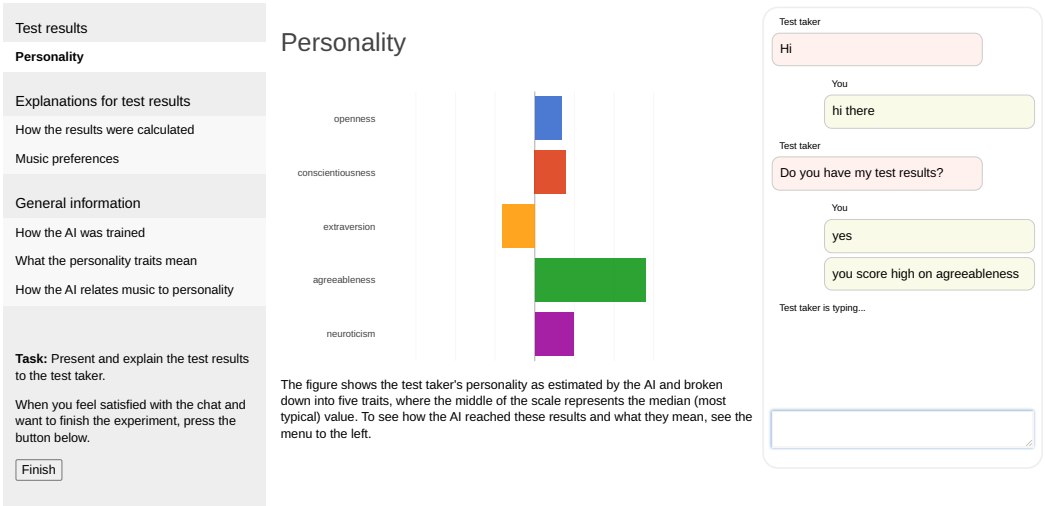


Figure 1. Screenshot of operator's main view during chat, with hypothetical prediction and chat messages. In the personality prediction plot, the X axis reflects log odds; bar lengths are proportional to the total estimated log odds for the respective traits ($\sum_i X_i \beta_{ti}$). Respondents only see a chat window (similar to right-most part of operator's view).

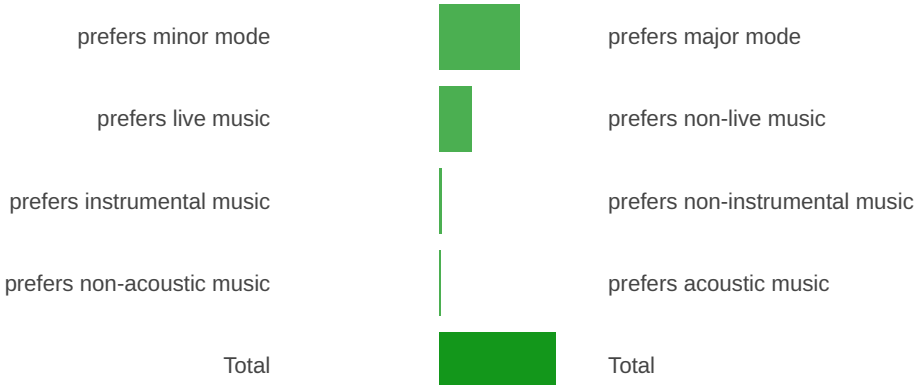


Figure 2. Example of a local feature contribution plot for agreeableness, in this case with only positive contributions. These plots are showed to operators under the section "How the results were calculated", with the following caption: *The figures above show how the AI sums up scores for different music preferences to scores for personality traits.* Contributions are sorted by decreasing importance with respect to the specific prediction. The X axis reflects log odds; for features, bar lengths are proportional to contributions ($X_i\beta_{ti}$), while for the "Total" bar, the length is proportional to the total estimated log odds ($\sum_i X_i\beta_{ti}$), as in Figure 1.

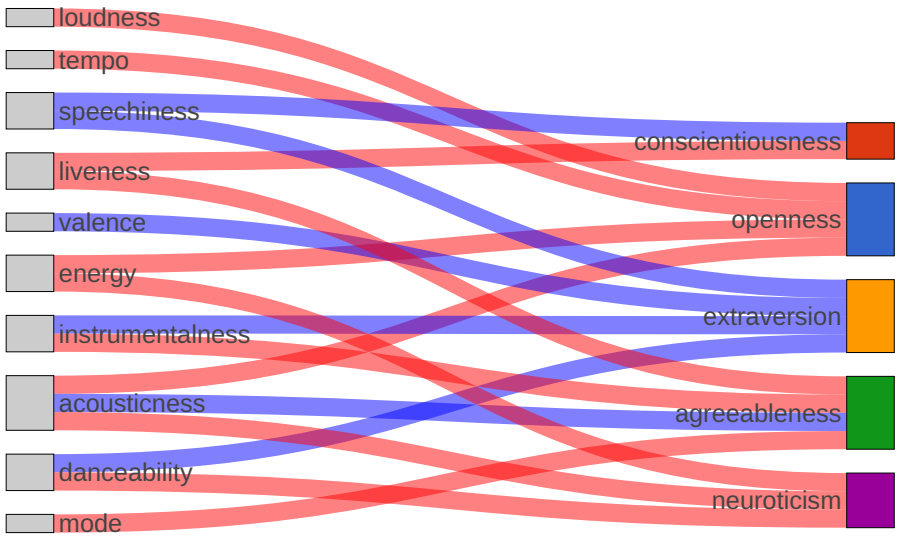


Figure 3. Global feature contribution plot, showed to operators under the section "How the AI relates music to personality", with the following caption: *The figure shows correlations between music preferences and personality traits, as learned by the AI. Blue links correspond to positive correlations, red links to negative correlations. For example, a red link from loudness to openness means that a preference for silent music correlates with openness, and conversely that a preference for loud music correlates with non-openness. (Higher valence means a stronger preference for happy/cheerful/euphoric music. Higher mode means a stronger preference for music in minor made.)* Blue links reflect positive coefficients ($\beta_{ti} > 0$) while red links reflect negative coefficients ($\beta_{ti} < 0$). Note that coefficient magnitudes are not visualised, based on the assumption that such level of detail would make the plot more difficult to interpret.

Since participants are paired up with each other, known issues of bias when using confederates [34] are avoided, enabling an open-ended investigation. When participants could not be paired up with each other, they were paired up with the experiment leader (the first author). Data collected in this manner was excluded.

For each big-five personality trait [35], a logistic regression model is trained to predict whether an individual has the trait in question (e.g., extraversion) based on the individual's music preferences and psychometric test results as ground truth. The dataset consists of music listening data and psychometric test results for 1000 users of the music website

Last.fm [36]. Specifically, for each trait, the training set is divided into two equally sized partitions: positive instances (trait value above median) and negative instances (remaining instances). In other words, the models are trained to estimate whether an individual belongs to the upper or lower half of the trait spectrum.

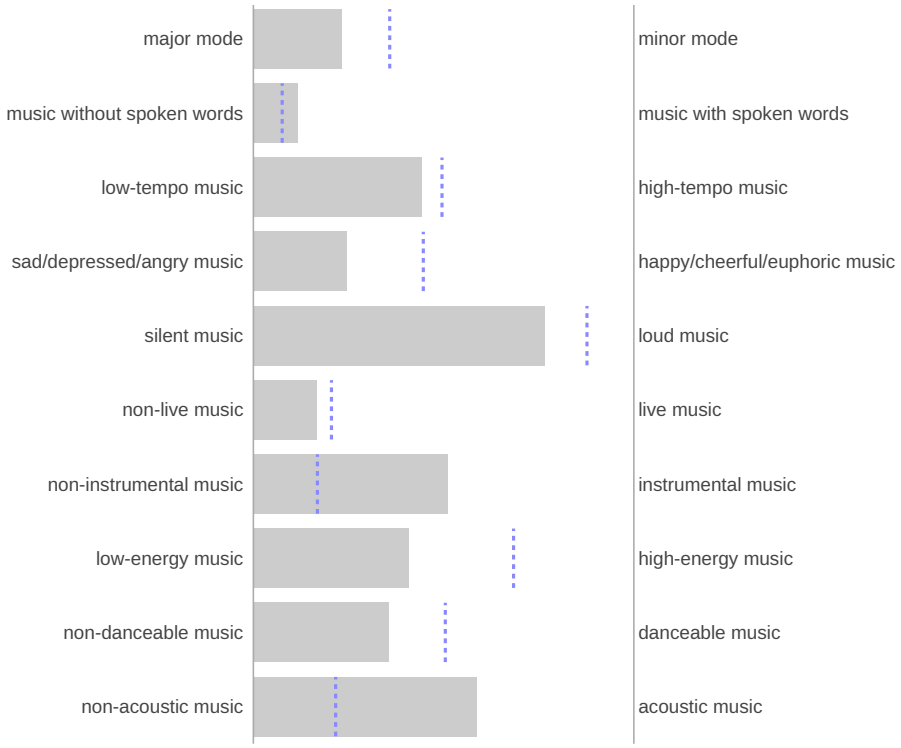


Figure 4. Plot of feature values for the respondent in relation to mean levels for the rated tracks, showed to operators under the section "Music preferences", with the following caption: *The figure shows the test taker's music preferences compared to average levels for the rated tracks (dotted blue lines), based on how the test taker rated the tracks in this experiment. Audio data for each track comes from Spotify.* The X axis spans from 0 to 1 and reflects raw feature values scaled to feature-specific min and max values; in other words, a value of 0 represents the minimum feature while a value of 1 represents the maximum feature value. Bar lengths reflect the respondent's aggregated feature values ($\tilde{X}_i$) while dotted blue lines reflect mean values for the rated tracks.

No intercept/bias terms are used. The main rationale for this choice is that the study focuses on lay explanations, while intercept terms generally make predictions more difficult to explain. Furthermore, since the training sets are balanced and features are standardised, the absence of intercepts can be assumed to not significantly degrade predictive performance.

Music preferences are encoded numerically as standardised aggregated feature values for audio properties (energy, loudness etc.):

$$X = Z_S(\tilde{X})$$

where $\tilde{X}$ is a vector of raw aggregated (non-standardised) feature values. The function Z_S standardises a vector with respect to a sample S:

$$Z_S(\tilde{X}) = \frac{\tilde{X} - \mu_S}{\sigma_S}$$

where μ_S and σ_S are vectors of mean and standard deviations for sample S. For instances in the dataset, $\tilde{X}$ contains mean values across each track that the individual has listened

to. For example, if an instance in the dataset has $\tilde{X}_i = -10$ for loudness, this means that the tracks listened to by the individual have a mean loudness value of -10. During model development, standardisation is performed with respect to the entire dataset.

Number of features range from 2 to 5 depending on trait (see Figure 3) and were selected manually as a trade-off between predictive performance and sparsity, where a small number of features was deemed desirable in order not to cause information overload. Specifically, for each trait, accuracy with n features was first measured and plotted; then, the smallest n with performance not significantly worse than the best performance was selected.

During inference, feature values are aggregated using weighted averaging across the rated tracks:

$$\tilde{X}_i = \frac{1}{\sum_{k=1}^{n} r_k} \cdot \sum_{j=1}^{n} r_j a_{ij}$$

where r_j is the respondent's rating for track j, and a_{ij} is the raw audio property value of feature i and track j. For participants, standardisation is performed with respect to the 10 rated tracks, in order to enhance variability across predictions. (If standardisation would instead be performed with respect to the entire dataset, overall prediction tendencies would be heavily influenced by the properties of the specific rated tracks. This would reduce variability across predictions, which was deemed undesirable in relation to the aim of the study.) Traits are predicted by estimating the probability that a respondent belongs to the upper half of the trait spectrum:

$$P(Y_t = 1 \mid X) = g^{-1}\left(\sum_i X_i \beta_{ti}\right)$$

where Y_t is the value of trait t and β_{ti} are the regression coefficients. The inverse link function g^{-1} for logistic regression is defined canonically:

$$g^{-1}(\eta) = \frac{1}{1 + \exp(-\eta)}$$

The tracks to be rated by participants were selected so as to maximise feature variance and thereby create ideal conditions for predictive performance. Specifically, a pool of candidate tracks was first obtained using Spotify API's `recommendations` method (https://developer.spotify.com/documentation/web-api/reference/get-recommendations, accessed on 28 April 2025). (Oral genres such as comedy and writing were excluded.) 100 random subsets, each containing 10 tracks, were then created. Finally, the subset with maximum variance (total standard deviation) across normalised feature values was selected.

It should be noted that due to the interpretability of the models, and the way in which explanatory information is obtained and visualised, the explanations in the tool accurately reflect how the models make predictions. For example, if a local feature contribution plot presents a high estimated agreeableness as mainly being composed of a preference for music in major mode (see Figure 2), this information is inherently faithful with respect to the actual workings of the model. (The specific methods for visualising explanatory information are elaborated in the captions of Figures 1–4.)

Since English proficiency levels are generally very high in the local population, and since potential participants were informed in the recruitment material that they would interact with each other in English and could choose to opt out on this basis, comprehension checks were not deemed relevant.

2.2. Recruitment of Participants

Since the experiment focuses on lay explanations, participants were recruited from a general audience where individuals cannot be expected to have expert knowledge concerning the relation between music preferences and personality or regarding how predictions from statistical models can be interpreted and explained. Specifically, in initial trials, recruitment was performed using convenience sampling of colleagues at the Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, and students attending a course on AI in society at the university; in subsequent trials, participants from the general public were recruited via information at the university's website, posters and flyers at the university campus, and marketing in social media. In the recruitment material, potential participants were invited to participate in an experiment where an AI assesses one's personality based on one's music preferences, and were informed that the assessment results would be communicated in English between participants. No monetary or other reward was offered to participants.

2.3. Collected Data

74 individuals participated in the experiment across 5 trials between June 2022 and September 2024, resulting in a corpus of 35 dialogues (referred to below as "D 1" etc.) encompassing 779 utterances (disregarding the dialogues where a participant was paired up with the experiment leader). Trial-specific information for collected data are provided in Table 1, while descriptive statistics for collected dialogues are provided in Table 2. Collected information about participant demographics (gender, age and education level) are summarised in Table 3.

Table 1. Trial-specific information (recruitment channel, date of data collection, and potential number of chats per participant) and frequencies for the collected data. "Excluded" refers to the number of dialogues where a participant was paired up with the experiment leader; these are not included in "Dialogues", which refers to the number of collected dialogues between participants. "In scope" refers to the number of dialogues that were revealed during analysis to contain explanations of AI predictions.

	Trial 1	**Trial 2**	**Trial 3**	**Trial 4**	**Trial 5**	**Total**
Channel	Colleagues	Colleagues	Students	Public	Public	
Date	Jun 2022	Jun 2022	Apr 2024	Sep 2024	Oct 2024	
Chats/participant	1	1	2	2	2	
Participants	3	4	15	46	6	74
Utterances	113	114	190	284	78	779
Dialogues	1	2	12	18	2	35
Excluded	1	0	0	5	2	8
In scope	1	1	5	4	2	13

Table 2. Descriptive statistics for collected dialogue data. Each cell contains mean value, with standard deviation in parentheses. For tokenisation, NLTK's [37] `word_tokenize` function was used.

	Operator	**Respondent**	**Total**
Utterances/dialogue	12.5 (14.7)	9.7 (10.7)	22.3 (24.6)
Tokens/utterances	9.5 (10.6)	8.6 (8.6)	9.1 (9.8)

Table 3. Participant demographics for the trials where such information was collected.

	Trial 4	Trial 5	Total
Gender			
Female	7	2	9
Male	9	2	11
Other	1	0	1
Age			
18–24 years old	1	0	1
25–34 years old	3	0	3
35–44 years old	2	1	3
45–54 years old	6	2	8
55–64 years old	3	1	4
65–74 years old	1	0	1
75 years old or older	1	0	1
Education level			
No schooling completed	0	0	0
Completed high school/gymnasium	0	0	0
Completed university degree	17	4	21

2.4. Dialogue Distillation

Building on a method for dialogue distillation proposed by Jönsson and Dahlbäck [38] and Larsson et al. [39], collected excerpts of human–human dialogues are manually rewritten into analogous human–computer interactions. The purpose of dialogue distillation is to inform dialogue system development by rewriting human–human dialogues into the kind of interactions that would occur if one of the interlocutors, in this case the explainer, was a dialogue system. In addition to a corpus of rewritten human–computer dialogues which can be used to inform design and modelling of dialogue systems, dialogue distillation also yields insights concerning the kind of dialogue capabilities that a dialogue system would need to have in order to emulate the role of the "replaced" human.

The distillation process generally follows a protocol (or set of guidelines) and builds on certain assumptions that may differ between domains and applications as well as the specific aims of the analysis. Previous work on dialogue distillation [38,39] assumes a waterfall approach where the protocol is developed before rewriting dialogues and where the dialogue system is implemented in a later stage of development to support the behaviours manifested in the rewritten dialogues. One limitation of this approach is that it can be difficult to anticipate à priori the kind of methodological choices that the protocol should inform, without first attempting to rewrite (at least some of the) dialogues. Furthermore, if dialogues are rewritten before system implementation, this can potentially result in a corpus with formally inconsistent or otherwise unimplementable dialogues. To mitigate these drawbacks, the present work adopts an iterative and incremental approach [40] where distillation is performed in parallel with modelling of dialogue management so that, at any given iteration, the system capabilities manifested by the rewritten dialogues are supported by the dialogue model. To ensure that the distilled behaviours are implementable, system behaviour is validated using automated tests [24].

The iterative process begins with minimal initial assumptions (see below) that are refined across iterations, similar to the use of open coding in grounded theory [41]. This process is steered towards gradually increasing the coverage of the dialogue management model and decreasing the differences between the rewritten and original dialogues. In the present study, this process continues until a substantial amount of the observed phenomena are deemed to be adequately modelled; at this stage, any remaining salient differences

between original and rewritten dialogues can be considered challenges for future work (as will be elaborated in Section 4).

The dialogue distillation in this study was performed by the first author, partly on the basis of feedback from the second author who reviewed in-progress distillation results.

2.4.1. Technical Assumptions

The dialogue system acting as a substitute for the human explainer is assumed to have general (domain- and language-independent) dialogue management capabilities as well as domain-specific resources for natural language understanding and generation (NLU, NLG) and for making statistical inferences using a predictive model.

The dialogue system is assumed to use some kind of information-state-based dialogue management [42–44], where the system keeps track of the state of the dialogue and selects which dialogue moves (speech acts) to perform by iteratively applying update rules. Importantly, however, no initial assumptions are made concerning the range of potential moves handled by the system, the structure of the system's information state, or the system's update rules. Instead, the elements of the system are incrementally developed during the distillation process.

2.4.2. Normative Assumptions

In addition to the theoretical goal of modelling the human–human dialogues as faithfully as possible, it is also assumed during distillation that the dialogue system should facilitate users in forming accurate mental models of how the AI works. This assumption reflects normative principles relating to AI transparency and trustworthiness [7,45] and concretely implies that explanations delivered by the dialogue system for its predictions should reflect the actual inner logic of the predictive model at hand. While this requirement may seem normatively evident, its implications for design and modelling of explanatory dialogue systems are not trivial, at least in cases where human interlocutors explain judgements using types of information that cannot straightforwardly be derived from a predictive model (as will be discussed in Sections 3.6 and 4). Furthermore, if the explainer contradicts the actual workings of the model at hand, such inaccuracies are "corrected" in the rewritten version.

2.4.3. Example

To illustrate the distillation process, we can consider the excerpt below, where the operator (O) explains a prediction in response to the respondent's (R) question:

(1) R: I really want to know what these results are based on...why am I so low on openness? kind of disagree with that
O: "openness to experience": the score was calculated on the basis that you don't prefer high-energy, loud music ... (D 20)

In early iterations, the dialogue management model only supported providing one piece of evidence or explanation at a time. Consequently, Example 1 above was rewritten as below (S = system, U = user):

U: I really want to know what the result is based on...why am I so low on openness? kind of disagree with that
S: You seem to prefer low-energy music.

In a subsequent iteration, the dialogue model was extended to support multiple pieces of evidence, causing the distilled excerpt to be refined as follows, achieving a higher degree of coverage:

U: I really want to know what the result is based on...why am I so low on openness? kind of disagree with that
S: You seem to prefer low-energy, silent music.

Later in the distillation process, the faithfulness of the operators' explanations was assessed, and potential corrections were made. At this stage, it was observed that O's stated explanation was incorrect, since R was in fact deemed to prefer high-energy, loud music. (Presumably, O had misinterpreted some of the visual information in the tool.) Consequently, the rewritten excerpt was refined as follows:

U: I really want to know what the result is based on...why am I so low on openness? kind of disagree with that
S: You seem to prefer high-energy, loud music.

The results presented in the next section pertain to the final stage of the distillation process.

3. Results

Explanations of AI predictions were addressed in 13 (37%) of the collected dialogues. A distillation of the dialogue excerpts revolving around explanations for AI predictions revealed various phenomena and challenges concerning the design and modelling of human–AI explanatory dialogue (henceforth referred to simply as dialogue modelling), as elaborated in the subsections below. Other topics in the dialogues, such as coordination of the situation as such (e.g., questions concerning whether to proceed to the next step of the experiment) and discussions concerning potential agreement or disagreement with AI predictions, were excluded from the analysis.

3.1. Types of Explananda

Two broad categories of explananda (circumstances to be explained [46]) are addressed in the dialogues: epistemic bases of predictions, and meaning/nature of target labels. Observed epistemic explanations occasionally target numerically conveyed predictions, as in the excerpt below (emphases in cited corpus excerpts have been added by the authors):

(2) O: You scored 5 on Openness. You scored 0.5 on Conscientiousness. ...
...
R: Why is it *a 5 on Openness* ... (D 2)

However, in most cases, predictions are interpreted and framed by the operator when conveyed to respondents, as in the excerpt below:

(3) O: You have been rated *the highest in openness*
R: Oh wow, why? :) (D 15)

Notably, in Excerpt 3, O's framing of the prediction involves a subjective evaluation of how the graphically presented information can/should be interpreted and contextualised in terms of how low/high a particular value is. A similar behaviour can be observed in the excerpt below, where O both frames the prediction evaluatively and elaborates the value numerically in more objective terms:

(4) R: What kind of information does the test give you?
...
O: apparently, you are *very open*
O: *almost 5* (out of −5 to 5 where 0 is the median)
R: It's interesting, I wonder what song would give *this trait* (D 13)

While most explananda target polar evaluations of predictions (i.e., that an estimated value is deemed low or high), in one case, the operator frames a prediction as neutral (ambivalent):

(5) O: on extroversion, you scored *pretty close to the median*
R: Do you know the link with music features? (D 13)

With respect to dialogue modelling, it is here assumed that the dialogue system frames predictions as positive or negative in relation to a median music listener, in a similar way to how predictions are visualised in the tool used by operators (see Figure 1). For example, if the system expresses that the user is "open", this reflects that the user's estimated score on openness is higher than for a median music listener.

Meanings of target labels are frequently targeted on a lexical (context-independent) level, as in the excerpt below, where R's question concerns the meaning of the term "agreeableness" as such:

(6) O: You have high scores for openess
...
O: Extraversion is just below the median value
O: Agreeablenes is even more below
R: *I think about the word agreeablenes, don't know what to think about that :)* (D 26)

Occasionally, the meaning of target labels is also interrogated in terms of implications of particular values. For example, one of the respondents asks: "Is a lower score on agreeableness a negative quality to have?"

3.2. Explanation Triggers and Query Types

In many cases where operators provide explanations, this is triggered by explicit queries by respondents. Among explanation requests, wh-questions are the most common. Examples include "what do you base this conclusion on", "Why do you think I'm a very aggreeable [sic] person?" and "what does agreeableness entail?"

In some cases, yes–no questions are used. In one such case, the question pragmatically serves as a confirmation request: "But the result could be because of a preference of music with lower tempo/am I correct?" In another case (Excerpt 5), the question "Do you know the link with music features?" pragmatically serves as a wh-question concerning a link between the respondent's ratings and the AI's judgement.

In two instances, literal assertions pragmatically serve as queries. In one such instance, R's utterance literally conveys an interrogative stance towards the embedded clause:

(7) R: *I wonder if music influences the personality or if it's only the other way*
O: yeah I was thinking that too (D 13)

In another instance, R's syntactically assertive utterance can be pragmatically understood as a confirmation question:

(8) O: the AI calculates the results based on a statistical model for each personality trait
...
R: hm ok, so *it takes into account many other people's statistics then*
O: a 1000 users apparently (D 19)

In Example 8, O's response seems to felicitously serve the broader information need implicated by R's indirect confirmation question.

In some cases, operators provide explanations unpromptedly. This behaviour is only observed for the meaning of target labels, as in the excerpt below:

(9) O: You scored 5 on Openness. You scored 0.5 on Conscientiousness. You scored −2.3 on Extraversion. You scored −0.7 on Agreeableness. You scored −0.5 on Neuroticism.
O: *Openness to experience describes a dimension of cognitive style that distinguishes imaginative, creative people from down-to-earth, conventional people.*
Conscientiousness concerns the way in which we control, regulate, and direct our impulses.
... (D 2)

3.3. Types of Explanantia

In response to queries concerning the epistemic basis for the AI's assessments, varying response strategies are used. One observed strategy is to provide the source of evidence, without detailing how specific types of information relate to the assessment:

(10) R: what do you base this conclusion on
O: This conclusion is based on *the score from your ratings of the music you listened to* (D 2)

A related strategy is to mention the factors on which the AI's assessment is made, without detailing how the factors contribute to the assessment:

(11) O: You scored −2.3 on Extraversion, which tells me that you are likely more introverted.
...
O: The Extraversion score *was based on Dancable/Non-Dancable music, Happy/Sad music, Intrumental/Non-instrumental music and music with/without spoken words.* (D 2)

In another case, the operator produces an algorithmic narrative conveying how the assessment was calculated. The narrative straightforwardly transforms the feature contribution plot into natural language:

(12) R: Why is it a 5 on Openness, based on what variables ...
O: *The 5 on Openness comes from the total score of four Preferences. Loud music is −5 and silent music is 5—you scored 4.5. Other preferences were acoustic music at −5 and non-acoustic on 5, where you scored −1. The high-energy music preference was −5 and low-energy music was 5—you scored 3.8. High-tempo preference was −5 and low-tempo music was 5—you scored 0,8. The total of these 4 preferences were a 5 on Openness* (D 2)

In other cases, the operator explains the AI's assessment argumentatively, i.e., as a claim supported by evidence:

(13) R: Can I get my test results? What are the scores and their interpretation?
...
O: you have scored low on "openness" and "neuroticism"
O: slightly higher on "extraversion"
...
R: I really want to know what these results are based on...why am I so low on openness? kind of disagree with that
O: "openness to experience": *the score was calculated on the basis that you don't prefer high-energy, loud music*
...
O: "Conscientiousness" *was based on the fact that you prefer non-live music and music with spoken words*
O: "Extraversion" *is based on the fact that you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrument music, non-danceable music*
(D 19)

Argumentative strategies are observed both for model-grounded explanations, as in Excerpt 13 above, and for explanations grounded in the operator's own knowledge, as in the excerpt below, where O explains the negative score for neuroticism with reference to R's preference for death metal:

(14) O: Your neuroticism was −1.7
...
O: Do you not enjoy death metal
R: not particularly
R: but slightly
O: Fair enough
O: Explains the sightly negative score (D 15)

In another instance, O speculates that not liking a particular track supports being either neurotic or non-neurotic:

(15) O: You have high on openness, and very low on neutoticism. neutral on the rest
R: oh, low on neuroticism, but I pressed so fast on the growling-dislike-button!
...
O: *Maybe if you were neurotic you would not like the growling, or you could. I can see reasons in both directions* (D 34)

Applying Toulmin's argumentation framework [33], it can be observed that operators typically support predictions with data, i.e., specific grounds supporting the prediction. In most cases, data are conveyed in a polar manner, e.g., "you don't prefer high-energy", "you showed a preference for accoustic music". (In technical terms, this corresponds to characterising a feature level as either low, typical or high.) However, in a situation where the prediction itself is neutral, data is conveyed neutrally (i.e., as neither low nor high):

(16) O: on extroversion, you scored pretty close to the median
R: Do you know the link with music features?
O: in the explanations chart there is a strong relationship between extroversion and *dancability and liveliness*
O: but if I'm reading this correctly, *your preferences didn't indicate strongly one way or another about those features* (D 13)

In another instance, a neutral claim is supported by contradictory evidence:

(17) O: you have scored low on "openness" and "neuroticism"
O: slightly higher on "extraversion"
...
R: I really want to know what these results are based on ...
...
O: "Extraversion" is based on the fact that *you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrument music, non-danceable music* (D 19)

In Excerpt 17, O explains a neutral estimate of extraversion using a contrastive while-clause whose main clause ("you prefer music with spoken words") reflects evidence for extraversion while the subordinate clause ("you disprefer sad/depressed/angry music, non-instrument music, non-danceable music") reflects evidence for introversion. (A similar behaviour can be observed later in the same dialogue in Excerpt 28, although in that case, the prediction is quite far from the median.)

While observed arguments typically highlight data, in the following exchange, O also conveys warrants (general rules or patterns, in this case relations between music preferences and personality):

(18) O: apparently, you are very open
O: almost 5 (out of −5 to 5 where 0 is the median)
R: It's interesting, I wonder what song would give this trait
O: well I actually can tell you something about that I think
O: *not which song in particular, but how openness relates to features of the music*
R: Oh great I'm interested
O: for example, *openness apparently has a strong relationship with acousticness*
O: and I think this is saying you showed a preference for accoustic music
R: Okay, yes I think I did
R: What about extraversion?
O: on extroversion, you scored pretty close to the median
R: Do you know the link with music features?
O: in the explanations chart there is a *strong relationship between extroversion and dancability and liveliness*
O: but if I'm reading this correctly, your preferences didn't indicate strongly one way or another about those features (D 13)

In Excerpt 18, O first highlights a warrant (a general relationship between openness and acousticness), and then links the warrant with the assessment (that R is very open) through a datum (R's preference for acoustic music). A similar strategy is observed in the excerpt below, where O's generic statement concerning the preferences that yield a high agreeableness score can be characterised as a warrant (or a conjunction of warrants):

(19) R: Why do you think I'm a very aggreeable person?
R: * disagreeable
O: I don't know, but the results says that *a person scores high in that category if they prefer non-live music, major mode, acoustic music, and non-instrumental music ...* (D 8)

A related behaviour can be observed in Excerpt 20 below, where O explains an assessment concerning a specific trait by elaborating different sub-traits:

(20) O: You have been rated the highest in openness
R: Oh wow, why? :)
O: Openness describes a dimension of style that is imaginative
O: creative people
O: down to earth
O: and conventional
R: are these three different grades?
O: Agreeableness is rated 0.5
O: The reasons that I gave was all for openness (D 15)

Although the conveyed sub-traits (imagination, creativity, etc.) are presented to operators in the tool as information concerning what personality traits mean, O in Excerpt 20 treats—and explicitly refers to—these sub-traits as reasons. In a possible interpretation,

O implicates that the AI generally associates openness with imagination, creativity, etc., and that the AI assesses R to be imaginative, creative, etc. Note, however, that in this case the implication is false, since the AI does not assess traits in terms of sub-traits.

As for meaning of target labels, explanantia are frequently provided as general definitions (taken from the tool), as in the first response from O below:

(21) R: ... Is a lower score on agreeableness a negative quality to have? ...
O: "*Agreeableness reflects individual differences in concern with cooperation and social harmony. Agreeable individuals value getting along with others.*"
O: *I guess it means that you are a bit less likely to like cooperation and socialising?* (D 6)

It is also common for interlocutors to elaborate on potential implications of specific values, as exemplified by O's second response in Excerpt 21. (In Larsson and Myrendal's terminology [47], these two strategies can be characterised as exemplification and explicification respectively.)

3.4. Response Strategies

In some cases, O's choice of response strategy aligns directly with R's mode of query. This is particularly evident in the case of polar/confirmation questions, as in the excerpt below:

(22) R: But the result could be because of a preference of music with lower tempo
R: am I correct? // ...
O: Yes this is correct. You preferred music with lower tempo. (D 2)

In Excerpt 22, the polar nature of R's query is explicitly reflected in O's affirmation ("Yes this is correct"), while the potential explanans embedded in R's question (a preference for music with low tempo) is explicitly confirmed by echoing it back. However, in many other cases, the degree of alignment between query type and response strategy is weak or unclear. For example, in Excerpt 11, where O mentions factors, R had previously asked, "Why is it a 5 on Openness, based on *what variables*". On the other hand, O's initial response to R's query used an algorithmic narrative (see Excerpt 12), possibly triggered by the quantitative nature of O's highlighted explanandum ("Why is a *5 on Openness*"), although such a link is far from evident. Furthermore, many queries in the corpus are constructed using open-ended wh-questions such as "*what* do you base this conclusion on" or "*Why* do you think I'm a very aggreeable person?" and thereby do not clearly indicate an expected explanans mode.

With respect to dialogue modelling, the lack of evident patterns regarding choice of response strategy for open-ended wh-questions provides limited guidance. The path taken here is to assume that in situations where the user's query does not indicate a particular kind of desired explanans, the system has some means of selecting a response strategy (e.g., a default strategy).

3.5. Argumentative Structure

As discussed above, the epistemic basis of the AI's judgements are frequently explained either via data (e.g., a preference for acoustic music), warrants (e.g., a relationship

between acousticness and openness), or a combination thereof. Argumentatively, conveying only data or warrants to support a claim (in this case concerning a person's personality) can be understood as an enthymeme [48], i.e., a truncated argument. Through this theoretical lens, a conclusion can be comprehensible if one or more premises that would make the argument more complete has been omitted, to the extent that the receiver is able to identify the missing premise(s). This can potentially explain why respondents sometimes signal understanding of operators' seemingly truncated explanations, as in the example below:

(23)
O: you have scored low on "openness" and "neuroticism"
O: slightly higher on "extraversion"
...
R: I really want to know what these results are based on ...
O: "openness to experience": the score was calculated on the basis that you don't prefer high-energy, loud music
R: well, I would not consider any of the music pieces I listened to as high-energy and loud music...but *ok, I see*
...
O: "Conscientiousness" was based on the fact that you prefer non-live music and music with spoken words
R: *mmm ic*
O: "Extraversion" is based on the fact that you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrument music, non-danceable music
R: *yes, this sounds about right*
O: "Agreeableness" is based on the fact that you prefer non-instrumental music, non-live music, and perhaps acoustic music
...
R: agreeableness's *explanation also sounds good to me*

(D 19)

In Excerpt 23, O's explanations are consistently enthymematic, in that they explain claims merely via data. Potential warrants that would make the explanations more complete, e.g., that the AI generally associates a dispreference for high-energy music with low degrees of openness, are not verbalised—even if such information is available to operators in the tool. Nevertheless, R consistently signals acceptance/positive understanding in response to O's explanations, indicating that R is able to infer implicit premises.

One factor that might explain why respondents seem to understand operators' enthymematic explanations is that data are conveyed in ways that indicate which kind of warrant might be at play. Specifically, operators use numerically vague datum expressions such as "you don't prefer high-energy music", despite having access to more detailed visualisations of the respondent's music preferences in the tool (see Figure 4). Presumably, the simple polar form of these expressions implicates a monotonic relation between the feature at hand and the target, e.g., that a preference for low-energy music is associated with low openness. If, instead, O had explained R's low estimated openness by describing R's preference for energy as "around 30 on a scale from low energy (0) to high energy (100)", this might have raised the question of whether a value of 30 specifically is indicative of low openness. A similar observation can be made in cases where operators explain predictions using their own knowledge rather than the information in the tool, e.g., when explaining the respondent's low score for neuroticism with the fact that the respondent does "not

enjoy death metal" (implicating a monotonic relation between neuroticism and preference for death metal). These findings resonate with previous work on lexical triggering of warrants [49] and the scalar nature of topoi/warrants [50].

In terms of modelling, the approach taken here is to let the system explain predictions with either data or warrants, depending on some policy. (The policy as such is not modelled.) Consequently, datum-based strategies are rewritten as follows (rewritten excerpts are denoted with the original excerpt number followed by a prime):

(13′) U: I really want to know what the result is based on...why am I so low on openness? kind of disagree with that
S: You seem to prefer low-energy, loud music.
...

Warrant-based strategies are rewritten as in the example below:

(18′) S: I think you rate high on openness.
U: It's interesting, I wonder why
S: Openness is associated with a preference for acoustic music.
...

3.6. Faithfulness and Sources of Evidence

In most cases where operators explain the AI's predictions, pieces of evidence accurately reflect the information in the tool and are thereby grounded in the actual workings of the model. However, in several instances, operators' reasoning goes beyond the information in the tool, as in Excerpt 15, where O speculates that "if you were neurotic you would not like the growling, or you could. I can see reasons in both directions". These reflections regarding the potential relation between neuroticism and preference for growling are quite evidently O's own and not grounded in the information in the tool. This can be concluded from the fact that the tool does not convey how ratings of specific tracks (such as the one with growling) influence predictions. Furthermore, operators can deduce from the information in the tool that growling is not included among the features (music preferences) considered by the model. Similarly, in Excerpt 14, O implicates a relation between neuroticism and preference for death metal which is not grounded in the model.

The inclination to use one's own knowledge and reasoning rather than merely relying on the information in the tool is particularly pertinent in relation to the meaning of target labels, as in Excerpt 21, where O uses the textual definition of agreeableness in the tool and the estimated value for agreeableness to predict R's sub-traits (cooperation and social harmony), and proposes an interpretation of those predictions: "I guess it means that you are a bit less likely to like cooperation and socialising?" Importantly, O's reasoning goes beyond the capabilities of the AI used in the experiment (or any conventional statistical model), since the model has no structured information about the meaning of the target labels. Even an AI system with access to textual definitions of target labels cannot perform this kind of reasoning unless equipped with general reasoning and linguistic capabilities.

In terms of dialogue modelling, it is here assumed appropriate for the explainer to only provide information that a CXAI system would realistically have access to. Consequently, when a user explicitly refers to a concept that goes beyond the system's information about the model or the domain, as in Excerpt 15, the system signals its inability to answer the question (see Section 3.7 below). Occurrences where an operator unpromptedly explains a prediction in a model-ungrounded way (Excerpts 14 and 21) are not rewritten, under the normative assumption that it is better to not explain a prediction unpromptedly than to provide a model-ungrounded unprompted explanation.

3.7. Answer Unavailability

When operators cannot provide an answer to an explanatory question raised by the respondent, various different strategies are used. One such strategy is to answer a related question, which can be understood as adhering to the maxim of relevance [51]. In the following excerpt, instead of answering R's question about the contribution of specific ratings, O answers a question concerning the meaning of one of the traits at hand:

(24) R: ... Among my answers, which ones tilted me the most towards being non-agreeable and slightly open?
O: *Your agreeableness reflects individual differences in concern with cooperation and social harmony.* (D 3)

Another strategy is to signal answer unavailability, which can be understood as adhering to the maxim of quality [51], as in the following excerpt:

(25) R: So, what did they base the low conscientiousnss on?
O: *hard to say, any specific music you rated high or low* (D 33)

Similarly, in Excerpt 7, O aligns with R's interrogative stance (towards the causal relationship between music listening and personality) with the feedback "yeah I was thinking that too", thereby implicating a lack of knowledge concerning the question at hand.

In some cases, answer unavailability signalling is combined with offering an answer to a related question, as in Excerpt 18, where R wonders "what song would give this trait", and O responds: "well I actually can tell you something about that I think//not which song in particular, but how openness relates to features of the music".

A similar combination of strategies can be observed in the following excerpt:

(26) R: So there were no specific answers in the test that led to these results? For instance my slight disliking of death metal.
O: *Not really, your answer for the tracks is not something i can see, What i can see is What your answers acumilated to. Then showing me What personality traits you have*
R: I see, that makes is a lot clearer. (D 3)

In Excerpt 26, R's response to O's answer unavailability feedback strongly conveys positive understanding, indicating that answer unavailability feedback can help explainees form a correct mental model of how the AI makes its predictions.

Among the observed strategies, signalling answer unavailability is here considered the most important for dialogue modelling, since it helps users form correct mental models. Assuming that the system is able to parse (at least some) questions for which answers are unavailable, exchanges can be rewritten as follows:

(24′) U: Among my answers, which ones tilted me the most towards being non-agreeable?
S: *I don't have any information about that.*

(7′) U: I wonder if a preference for acoustic music influences openness or if it's only the other way
S: *I don't have any information about that.*

The rewritten excerpts 24′ and 7′ assume that the NLU component can identify and represent relations concerning which the system lacks information. However, it is unrealistic to assume that any unanswerable question that a user might ask can be represented by the NLU using the system's internal semantics. For example, music concepts such as "death metal" and "growling", whose semantics are not related in any evident way to the features or target labels at hand, are here assumed unparsable by the NLU. Consequently, when excerpts where respondents ask questions involving such concepts are rewritten, the system gives negative understanding feedback:

(15′) S: I think you rate low on neuroticism.
U: oh, is this because I don't like growling?
S: *Sorry, I don't understand what "growling" means in this context.*

(14′) U: do you think death metal fans are 5 or −5 in neuroticism?
S: *Sorry, I don't understand what "death metal" means in this context.*

(26′) U: So there were no specific answers in the test that led to these results? For instance my slight disliking of death metal.
S: *Sorry, I don't understand what "death metal" means in this context.*

Similarly, it is assumed adequate for the system to signal negative understanding when the user asks questions concerning implications of predicted values, since the very act of formally representing such questions itself requires deep domain knowledge. Consequently, such excerpts are rewritten in the same manner:

(21′) R: Is a lower score on agreeableness a negative quality to have?
S: *Sorry, I don't understand what "negative quality" means in this context.*

3.8. Feedback and Grounding

Interlocutors (especially respondents) use various forms of feedback such as "ok" and "I see", as in the following excerpt:

(27) O: the AI calculates the results based on a statistical model for each personality trait
O: I doubt if it can discriminate different phases
R: *hm ok*, so it takes into account many other people's statistics then
O: of a particular person
R: yes, what you described is a very general picture
O: a 1000 users apparently
R: *aha I see*, quite a small set
R: ok, I am *pretty satisfied with the explanation*
O: users of the music website Last.fm
O: I am glad :) (D 19)

In Excerpt 27, O either accommodates R's acknowledgement silently, or positively affirms it. Another observed behaviour in response to acknowledgements from R is to continue to the next sub-topic or next piece of information to deliver, as in Excerpt 23, where O's incremental delivery of predictions is driven forward by affirmative feedback from R ("ok, I see", "mmm ic", "yes, this sounds about right").

Negative understanding feedback from R can also be observed, as in the following excerpt (continuation of Excerpt 23):

(28) R: I think these results kind of make sense for myself, but *I am having hard times understanding my openness raiting*. It's interesting that I am extravert but not open based on these results...don't you see a confusion there?
R: agreeableness's explanation also sounds good to me
R: i see now
O: "Neuroticism" is based on the fact that you disprefer acoustic, high-energy music while you slightly prefer non-danceable music
... (D 19)

In Excerpt 28, negative understanding feedback yields no evident response from O, possibly due to the presence of multiple parallel issues being raised. Another observed response to negative understanding feedback from R is to treat the feedback as a matter of agreement:

(29) O: Do you agree or disagree with the results
R: but *it is hard to understand* what in the music made the AI think that I am 0 in extraversion...
O: ?
O: Yeah, so disagree? (D 15)

In Excerpt 29, a possible reason for O's behaviour is that O does not know how to explain the AI's judgement. This hypothesis is supported by the fact that, throughout the interaction, this operator does not provide any specific explanations for the AI's judgement. With respect to dialogue modelling, none of the observed response behaviours above seem evidently ideal to emulate from a normative perspective. Instead, it may be more adequate to treat negative understanding feedback as requests for an explanation, and hence, to provide an explanation (if such is available).

Feedback concerning presupposition violations is also observed, as in the following excerpt, where O signals that the presupposed content embedded in R's why-question is false:

(30) R: Canbyou give me my results
O: opennesa 5
...
O: neurotisk −1.2-ish
...
R: Why am i neurotisk
O: *no, minus 1.2*
R: Which means i am not neurotisk
O: guess so (D 10)

In another case, when R asks a question which seems to presuppose that R has other music preferences than those inferred by the AI from R's ratings, which leads O to highlight the mismatch:

(31) R: ... I listen to a lot of heavy metal. Is there a correlation between loud and fast music and my score?
O: *You seem to prefer slower and less loud music*
R: haha that's the complete opposite. ... (D 8)

With respect to modelling, it is assumed the NLU component is able to identify presuppositions in user input and that the dialogue manager signals potential presupposition violations concerning, e.g., predictions or feature values embedded in questions.

3.9. Anaphora

Anaphoric references to explananda, such as "what do you base *this conclusion* on", are frequent in the analysed empirical material. During the process of distillation, anaphoric references are assumed to be contextually resolved by the NLU component.

3.10. Turn-Taking and Complex Explanantia

Interlocutors frequently combine multiple pieces of explanantia in a single utterance or sentence. For example, in Excerpt 13, O explains the AI's assessment concerning extraversion by combining multiple music preferences using a contrastive structure in a coordinating conjunction: "'Extraversion' is based on the fact that *you prefer music with spoken words while you disprefer sad/depressed/angry music, non-instrument music, non-danceable music*". Elsewhere in the same excerpt, coordination takes place on the adjectival level when the operator states that "the score was calculated on the basis that *you don't prefer high-energy, loud music*".

In contrast, in Excerpt 20, O delivers sub-explanations incrementally across multiple utterances ("Openness describes a dimension of style that is imaginative//creative people//down to earth//and conventional"); it is also worth noting that O uses the discourse marker "*and*" to explicitly mark the final utterance as a topic continuation.

In some cases, acknowledging feedback from the receiver seems to serve as a signal to provide additional information, as in Excerpt 23 (see Section 3.8).

With respect to modelling, it is here assumed that an interlocutor who delivers complex information has some means of designing a suitable turn-constructional unit [52] at any given moment in the dialogue. For example, an interlocutor may have knowledge about two explanantia but decide to initially only mention one of them, as exemplified by the rewritten excerpt below:

(13″) U: I really want to know what the result is based on...why am I so low on openness? kind of disagree with that
S: You seem to prefer low-energy music.
U: ok
S: Also, you seem to prefer loud music.
...

3.11. Ellipsis and Explanandum Co-Referencing

When operators respond to explanation requests, they occasionally omit the explanandum targeted by the respondent, presumably since it is evident from the dialogue context, as in Excerpt 30, where O responds elliptically to R's question, "Why am i neurotisk" with "no, minus 1.2", (correctly) assuming that R can infer from context that the target of the feedback is R's question about neuroticism.

In other cases, in their responses, operators restate the explanandum targeted by the respondent, as in Excerpt 13, where O responds to R's why-question concerning low openness as follows: "*"openness to experience"*: the score was calculated on the basis that you don't prefer high-energy, loud music ..."

In some cases, co-references to previously raised explananda are anaphoric rather than explicit, constituting a middle ground between ellipsis and explicit restatement, as in the excerpt below:

(32) O: ... You scored 5 in Openness, a bit under −2 in Conscientiousness, a bit over 1 in Extraversion, a bit under −4 in Agreeableness, and almost −3 in Neurotiscm
R: Why do you think I'm a very aggreeable person?
R: * disagreeable
O: I don't know, but the results says that a person scores high in *that category* if they prefer non-live music, major mode, acoustic music, and non-instrumental music. Maybe you have other preferences (D 8)

Presumably, the choice concerning whether to respond elliptically depends on the extent to which the operator assumes that the respondent can infer the explanandum from context. However, the conditions influencing this choice cannot straightforwardly be determined merely by analysing the empirical material at hand. In terms of modelling, it is hence assumed that the system has some means of choosing whether to restate the explanandum when responding to explanation requests, although the condition as such is not modelled.

The data suggests that choice of co-referring strategies is not only influenced by considerations concerning contextual salience, as discussed above, but also serves to interactively coordinate how predictions might be interpreted. In Excerpt 30, interlocutors refer to the prediction at hand both numerically ("minus 1.2") and polarly ("Which means i am not neurotisk"). A similar pattern is observed in Excerpt 32 ("a bit under −4 in Agreeableness"/"very aggreeable person"). As exemplified by Excerpt 30, an interpretation by one interlocutor can be denied/corrected by the other interlocutor. With respect to modelling, the path taken here is to let the system consistently deliver polar assertions (with potential hedging), which reduces the need to coordinate numerical interpretations. However, in future work, it might be relevant to further explore semantic coordination in relation to co-referring strategies in the context of XAI.

3.12. Reliability and Epistemic Stance

Interlocutors use different strategies for managing reliability of the AI's judgements and coordinating epistemic stance in relation to them. One such strategy focuses on highlighting the overall limitations of the AI's reasoning capabilities, as exemplified by the following excerpt:

(33) O: the AI calculates the results *based on a statistical model* for each personality trait
O: *I doubt if it can discriminate different phases*
R: hm ok, so it takes into account *many other people's statistics* then
O: *of a particular person*
R: yes, what you described is a *very general picture*
O: a *1000 users* apparently
R: aha I see, quite a *small set*
...
O: *users of the music website Last.fm*
...
R: *very specific website - most of the people right now don't really use Last.fm, maybe that would explain a lot*
...
O: I agree (D 19)

A similar strategy can be observed in the excerpt below:

(34) O: I can see that you have very high scores on openness
R: That is to be expected
R: *Selection bias for people that participate in this kind of experiment*
R: *or any experiment* (D 33)

In Toulmin's argumentation framework [33], this strategy focuses on how warrants are backed. In terms of dialogue modelling, similar behaviour can be emulated by providing the system with information concerning, e.g., model type and nature of training data.

Another strategy involves epistemic detachment from the AI which is achieved with hedging markers such as *perhaps* ("'Agreeableness' is based on the fact that you prefer non-instrumental music, non-live music, and *perhaps* acoustic music"), stance markers such as "*apparently* accousticness is positively correlated with neuroticism" and metadiscursive expressions such as "*if I'm reading this correctly*, your preferences didn't indicate strongly one way or another about those features". In Excerpt 32, O's epistemic detachment takes the form of explicit evidential marking ("*I don't know, but the results says ...*") and occurs in response to R's presupposed attribution of the AI's judgement to O.

With respect to dialogue modelling, the path taken here is to hedge the system's predictions based on confidence estimates. For example, a moderately confident prediction can be expressed as "I think you rate low on openness" or "It seems that you have an open personality", without explicitly representing propositional attitudes—in this case, that the belief concerning R's personality is ascribed to the system itself—in the internal semantics of the system. In contrast, modelling the type of interactively coordinated management of epistemic stance observed in Excerpt 32 would require propositional attitudes to be represented.

3.13. Synthesis of Findings

Building on the taxonomy proposed by Berman and Larsson [24], the findings of the analysis are synthesised into types of information which it might be important for CXAI systems to provide (Table 4) and dialogue capabilities which it might be desirable for CXAI systems to possess (Table 5).

Table 4. Types of information delivered by explainers in collected human dialogues.

Type of Information	Corpus Example(s)
Predicted value or class, e.g., that a person is extraverted or scores −2.3 on extraversion	11, 13–15, 18, 20, 23 and 32
Feature value characterisation, i.e., whether a feature value is high or low in relation to a (potentially implicit) reference value, e.g., that a person likes low-energy music	13, 14, 23 and 32
Warrant, i.e., general inference rules/patterns used by the model when making its predictions, e.g., that it associates a preference for silent music with scoring high on openness	18–20
Features based on which a prediction is made, e.g., that openness is estimated based on preferences concerning danceability, valence, instrumentalness, and speechiness	11
General definition of a term, e.g., that "openness to experience" describes a dimension of cognitive style that distinguishes imaginative, creative people from down-to-earth, conventional people	20, 21, 23 and 24
Potential implications of a specific prediction, e.g., that low agreeableness implies being less likely to cooperate and socialise	21
Inference steps or calculations on which a prediction is based	12
Model information, e.g., model type and nature of training data	33 and 34

Table 5. Explanatory dialogue capabilities identified in collected human dialogues.

Dialogue Capability	Corpus Example(s)
Question answering and information delivery	
Answer wh-question, e.g., concerning prediction outcomes, meaning of terms, or explanations for predictions	4, 6, 10, 12, 13, 18 and 20
Deliver explanation unpromptedly, assuming that the system has some means to determine whether an explanation should be provided together with the prediction	9
Select most relevant answer, e.g., datum (feature level characterisation), warrant or features, assuming that the system has some means of assessing relevance [1]	See Feature level characterisation, Warrant and Features in Table 4
Confirm/disconfirm hypothetical explanation, e.g., concerning whether a datum supports a prediction	22
Provide multiple answers, either in a single utterance and/or incrementally across multiple utterances and using continuation markers when appropriate [2]	11–13 (single utterance); 20 (incrementally)
Provide contradictory evidence contrastively, e.g., that a particular circumstance supports a prediction, while another circumstance speaks against it	17

Table 5. *Cont.*

Dialogue Capability	Corpus Example(s)
Context management	
Resolve ellipsis, e.g., implicit content of why-questions	3–5
Grounding and meta-communication	
Deliver additional information when user provides an acknowledgement, if such information is available [3]	23
Signal presupposition violation if the user's utterance presupposes that the system holds a view which it in fact does not	30 and 31
Signal answer unavailability if the user asks a question for which no answer is available	7′, 24′ and 25
Provide negative understanding feedback if a sub-phrase in the user's utterance cannot be mapped onto the system's knowledge representations	14′ and 15′

[1] Cf. previous work on informativity in relation to the user's beliefs [7,53,54]. [2] It is assumed that the system has some means of determining a suitably sized turn-constructional unit [52] (e.g., to state a maximum of three music preferences per utterance). [3] This capability requires the system to keep track of both the question that is currently being discussed [42,43], and which answers that have already been delivered.

4. Discussion

On the basis of the findings of the study, this section will discuss implications for future research concerning CXAI. Limitations of the study will also be discussed.

4.1. Implications for Future Work

One of the main findings of the study is the crucial role of arguments generally and enthymemes (truncated arguments) specifically. While a substantial body of previous work has discussed argumentation in relation to XAI and human–computer interaction on an abstract conceptual level (see, e.g., [55–59]), the present study contributes with unique empirical data and analyses concerning how arguments and enthymemes play out in concrete human dialogues revolving around AI predictions. Specifically, the study has found that explainers tend to express data (specific grounds presented in support of a claim) in ways which indicate the kind of warrant (general rule, principle or pattern) that connects the data with the claim (prediction). Importantly, this finding concerns model-grounded explanations as well as explanations that are grounded in respondents' own knowledge, indicating that the observed explanatory behaviour is not necessarily restricted to the types of models used in the present experiment. Previous work [23] suggests that argumentative information cannot generally be obtained for opaque models such as deep neural nets and random forests, even with post hoc explanation methods such as LIME [1] and SHAP [2], suggesting that the observed dialogue strategies can only be emulated when the models being explained possess certain formal properties typically associated with inherent interpretability. Consequently, an approach based on merely adding a conversational layer to post hoc explanation methods, in the vein of TalkToModel [17] and X-Agent [19], will not evidently achieve alignment with human explanation strategies. More work will be needed in the future to analyse which requirements human-aligned explanation strategies impose on models and technical explanation methods.

Another important finding concerns the handling of presupposition violations. The study revealed two scenarios in which explainees ask questions which presuppose false beliefs concerning predictions and feature values, and where explainers detect and

correct these inaccuracies. Arguably, this capability is particularly crucial in the context of CXAI, given the broad range of potential false assumptions that user utterances might presuppose. This distinguishes CXAI from menu-based interactive XAI, where, typically, users can only obtain answers to questions that have been pre-configured by the system designer. Despite this unique property of conversation, a recent review found that none of 15 studied CXAI systems were able to identify and correct false "assumptions" in the user's requests [14], while a comparison of dialogue capabilities showed that only one of the three studied CXAI systems possessed this ability [24], indicating ample room for further development in this regard.

The study also reveals that, in some cases, explainers provide explanations which are based on their own knowledge or assumptions rather than being grounded in the actual workings of the AI. While reasons for such behaviour cannot be determined, it seems plausible that the integration of other knowledge sources addresses a genuine need by mitigating limitations associated with the kind of explanatory information that can be derived from statistical models alone. In relation to CXAI, this observation highlights an important design choice concerning whether human–AI interfaces should only provide explanations grounded in the actual workings of the AI or potentially also explanations obtained from other knowledge sources. When rewriting human–human to human–AI dialogues, the path taken here has been to assume that the system only provides model-grounded explanations, in order to not lead users to form incorrect mental models. This can be contrasted with the systems proposed by Feustel et al. [60] and Schindler et al. [61], who augment model-based explanations with domain knowledge derived by other means. For example, the system proposed by Schindler et al. [61] may explain why loan duration influences a particular decision concerning credit applications by stating, "Lenders will usually feel more comfortable lending you money for a shorter period because you're more likely to be able to pay it back." Such explanations are generated on the basis of arguments automatically mined from text documents about the domain at hand, rather than being grounded in the workings of the predictive model. Hence, explanations obtained from other sources can be potentially misleading with respect to why a particular model makes a certain prediction. It is for this reason that the present study focuses on model-grounded explanations. An alternative approach for augmenting model-grounded explanations with external domain knowledge, while maintaining transparency, could be to let the system use both kinds of knowledge and state the source of evidence. For example, if a user asks why a short loan duration increases the chance of having one's credit application approved, the system can clarify that the model can only observe statistical correlations in data and not the potential causes behind these correlations. It can also offer a causal explanation for this correlation while making clear that this explanation might not reflect the actual reason for why the correlation exists in the training data.

In relation to issues concerning faithfulness and sources of evidence, the study also revealed various ways through which interlocutors signal their epistemic stance in relation to the predictions and evidence. These observations indicate design choices concerning how CXAI systems might position themselves epistemically in relation to the predictive model at hand. The approach taken here has been to manage epistemic stance on the level of surface realisation, rather than semantically. In practical terms, epistemic positioning is thereby delegated to the natural-language generation component of a dialogue system. An alternative approach could be to represent epistemic stance on the level of dialogue management. This would, in principle, enable epistemic stance to be addressed and coordinated interactively in human–computer dialogue. Such a capability can be important from the perspective of accountability and trust in critical domains such as healthcare, where transparency concerning the source of a prediction or explanation can be important.

For example, if a patient asks a conversational healthcare system why "it" predicts a certain outcome of surgery, it might be important that the system clarifies the actual source of the prediction (namely a predictive model), even if the source was not explicitly interrogated by the user (as in Example 32). Such a behaviour hinges on an ability to detect epistemic presupposed agency in user input and to highlight the actual source of a claim or evidence in the case of a presupposition violation.

Furthermore, the study has highlighted that the meaning of target labels can be an important topic in human–AI interaction. The finding suggests that it may be important for CXAI systems to engage in semantic coordination of terms or expressions used by the system [47], e.g., unpromptedly or on demand. However, such behaviour can potentially trigger false implications concerning the actual capabilities of the system and lead users to form incorrect mental models. For example, if a model predicts that a person has attention deficit hyperactivity disorder (ADHD), based on self-reports provided by the individual, and the system explains that ADHD involves executive dysfunction, then this might implicate that the model predicts that the person has executive dysfunction. However, if the model is merely trained to diagnose ADHD based on historical data (self-reports and ground-truth diagnoses), it in fact lacks knowledge about the symptoms and causes of the condition and hence cannot predict symptoms or causes. Future work will thus need to investigate how CXAI systems might engage in semantic coordination concerning target labels without producing presupposition violations that could potentially give users an inaccurate mental model of the AI's capabilities. Interesting challenges are also raised by the potential need to semantically coordinate vague and context-sensitive explanatory expressions. For example, if a system explains its prediction that a person has ADHD with reference to the fact that the person forgets appointments or obligations "relatively frequently", this may raise the question of what the system means by "relatively frequently" (in this context). Depending on the type of predictive model and technical explanation method, it may not be evident how to best answer such a question.

4.2. Validity

It can be observed that a relatively low ratio (37%) of collected dialogues was revealed to address topics of interest to the present study. Although the low ratio is not necessarily problematic by itself, combined with a relatively limited sample, it resulted in a small amount of data of interest. Due to the small sample size as well as the focus on a single domain with relatively low stakes, generalisations to other scenarios and types of situations need to be made cautiously. Potential selection biases associated with the data collection constitute an additional source of caution concerning generalisability; specifically, many of the participants in the study were recruited in an academic setting, and the education level of participants was generally high. To this end, findings should primarily be regarded as indications that certain phenomena *can* occur in human–human interactions (cf. Peräkylä [62]'s notion of possibility), even if the prevalence of these phenomena in broader settings cannot be reliably estimated merely on the basis of the data analysed in this study. Nevertheless, the study constitutes an important advancement in relation to previous work in CXAI, where the system's dialogue capabilities have been based entirely on researchers' à priori assumptions.

Like any system design method, dialogue distillation involves subjective choices shaped by technical and normative à priori assumptions. For example, if it had been deemed important to emulate respondents' use of evidence which is ungrounded in the workings of the model (see Section 3.6), some of the dialogue excerpts would be rewritten in another way, and thereby motivate other modelling choices. Even under identical initial assumptions, other coders may potentially rewrite the same dialogues differently,

depending, e.g., on modelling choices. To partly mitigate these reliability concerns, all methods and materials—dialogue distillation procedure, primary empirical sources, à priori assumptions, rewritten dialogues and derived dialogue model—have been transparently disclosed for the purposes of scrutiny and replicability, following established norms for qualitative conversation analysis [63].

4.3. Limitations

Since the models used in the experiment are employed differently in inference than in training, their predictive performance cannot straightforwardly be evaluated using conventional metrics such as accuracy. However, this limitation was deemed to have no substantive impact in relation to the focus of the study.

While English proficiency levels are generally high in the local population, potential language deficiencies might still have affected the dynamics and content of the collected dialogues, although the analysis did not reveal any evident instances to that effect.

In trials 3–5, data collection did not achieve its theoretical potential in terms of the amount of collected data. This is reflected in the frequencies for these trials in Table 1, where there is a discrepancy between, on the one hand, the number of dialogues that theoretically could have been collected given the number of participants and the number of interactions per participant and, on the other hand, the actual number of collected dialogues. For example, in trial 3, the total number of participants (15) and interactions per participant (2) could in principle enable $7 \cdot 2 = 14$ dialogues to be collected (since 7 of the 15 participants could be paired up with another participant twice); however, only 12 dialogues were collected. The reasons for this discrepancy were that three participants did not proceed through all the steps in the experiment and that one pair of participants did not exchange any chat messages. The same set of failure modes were observed for trials 4 and 5.

In some instances, it was observed that participants were not paired up as expected due to technical problems. While this explains why some participants did not complete the entire experiment, the exact reasons for each incomplete participation could not be fully determined; possibly, some participants simply aborted.

Since interlocutors only perform their respective roles once, the present study cannot investigate the extent to which conversational behaviours vary across interactions for specific interlocutors. While the present setup serves to avoid learning effects across interactions and thereby enhance the focus on lay explanations, repeated interactions could enable other kinds of research questions to be addressed.

The extent to which the identified dialogue capabilities contribute to desired outcomes when emulated in CXAI systems used by humans, such as understanding of model behaviour [64] or improved decision-making [65–67], has not been studied and constitutes an interesting direction for future research.

Author Contributions: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualisation, writing—original draft, writing—review and editing: A.B.; funding acquisition, supervision, writing—review and editing: C.H. All authors have read and agreed to the published version of the manuscript.

Funding: Alexander Berman 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. Christine Howes was supported by the ERC Starting Grant DivCon: Divergence and convergence in dialogue: The dynamic management of mismatches (101077927) and the Swedish Research Council grant (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

Institutional Review Board Statement: Participation in the study was anonymous, and no personal data, such as participants' names, were collected. Dialogue data was screened before being stored to ensure that collected utterances did not contain personal data. To reduce risk of over-reliance on AI predictions, participants were informed in a debriefing that the method used in the experiment for estimating personality from music preferences is experimental and has not been validated scientifically. Consent from participants was obtained in the form of a checkbox on the experiment's start page, through which participants agreed that their interaction would be collected, stored and made available for research purposes. The Swedish Ethical Review Authority issued an advisory opinion (case number 2022-06881-01) stating that since the study does not involve any physical or other intervention on participants, as covered by Section 4 of the Swedish Ethical Review Act, or any processing of personal data of the kind covered by Section 3 of the Ethical Review Act, the study is not covered by the provisions of Sections 3–4 of the Ethical Review Act and therefore should not undergo ethical approval. The advisory opinion also stated that the authority raises no ethical objections to the research project.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Training data for the predictive models used in the experiment is available in a public repository (https://gitlab.cp.jku.at/alessandro/pers-corr, accessed on 14 January 2025). The dialogue data collected and analysed within the study is available in the Dialogues on Music, Personality and AI repository (https://osf.io/z6d4s, accessed on 21 November 2025). All rewritten dialogues are available in the public repository for the dialogue manager, which has been developed partly on the basis of the analysis presented in this article (https://github.com/alex-berman/BKOS, accessed on 26 November 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier; KDD '16. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
2. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions; NIPS'17. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; pp. 4768–4777.
3. Lakkaraju, H.; Bach, S.H.; Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1675–1684.
4. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Stat. Surv.* **2022**, *16*, 1–85. [CrossRef]
5. Marques-Silva, J.; Ignatiev, A. No silver bullet: Interpretable ML models must be explained. *Front. Artif. Intell.* **2023**, *6*, 1128212. [CrossRef] [PubMed]
6. Molnar, C. *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*, 3rd ed.; Christoph Molnar: Munich, Germany, 2025.
7. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38. [CrossRef]
8. Rohlfing, K.J.; Cimiano, P.; Scharlau, I.; Matzner, T.; Buhl, H.M.; Buschmeier, H.; Esposito, E.; Grimminger, A.; Hammer, B.; Häb-Umbach, R.; et al. Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *13*, 717–728. [CrossRef]
9. Finke, J.; Horwath, I.; Matzner, T.; Schulz, C. (De)Coding Social Practice in the Field of XAI: Towards a Co-constructive Framework of Explanations and Understanding Between Lay Users and Algorithmic Systems. In *Proceedings of the Artificial Intelligence in HCI*, Online, 26 June–1 July 2022; Degen, H., Ntoa, S., Eds.; Springer: Cham, Switzerland, 2022; pp. 149–160.
10. Lakkaraju, H.; Slack, D.; Chen, Y.; Tan, C.; Singh, S. Rethinking Explainability as a Dialogue: A Practitioner's Perspective. *arXiv* **2022**, arXiv:2202.01875.
11. Dazeley, R.; Vamplew, P.; Foale, C.; Young, C.; Aryal, S.; Cruz, F. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artif. Intell.* **2021**, *299*, 103525. [CrossRef]
12. Mariotti, E.; Alonso, J.M.; Gatt, A. Towards Harnessing Natural Language Generation to Explain Black-box Models. In Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, Online, 18 December 2020; pp. 22–27.

13. Sokol, K.; Flach, P.A. Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, Stockholm, Sweden, 13–19 July 2018; pp. 5868–5870.
14. Mindlin, D.; Beer, F.; Sieger, L.N.; Heindorf, S.; Esposito, E.; Ngonga Ngomo, A.C.; Cimiano, P. Beyond one-shot explanations: A systematic literature review of dialogue-based xAI approaches. *Artif. Intell. Rev.* **2025**, *58*, 81. [CrossRef]
15. Werner, C. Explainable AI through Rule-based Interactive Conversation. In Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference, Copenhagen, Denmark, 30 March–2 April 2020.
16. 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 *Proceedings of the ECML PKDD 2020 Workshops, Ghent, Belgium, 14–18 September 2020*; Koprinska, I., Kamp, M., Appice, A., Loglisci, C., Antonie, L., Zimmermann, A., Guidotti, R., Özgöbek, Ö., Ribeiro, R.P., Gavaldà, R., et al., Eds.; Springer: Cham, Switzerland, 2020; pp. 447–459.
17. Slack, D.; Krishna, S.; Lakkaraju, H.; Singh, S. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nat. Mach. Intell.* **2023**, *5*, 873–883. [CrossRef]
18. Feldhus, N.; Ravichandran, A.M.; Möller, S. Mediators: Conversational Agents Explaining NLP Model Behavior. *arXiv* **2022**, arXiv:2206.06029. [CrossRef]
19. Nguyen, V.B.; Schlötterer, J.; Seifert, C. From Black Boxes to Conversations: Incorporating XAI in a Conversational Agent. In *Proceedings of the Explainable Artificial Intelligence, Lisbon, Portugal, 26–28 July 2023*; Longo, L., Ed.; Springer: Cham, Switzerland, 2023; pp. 71–96.
20. Wijekoon, A.; Corsar, D.; Wiratunga, N.; Martin, K.; Salimi, P. Tell me more: Intent Fulfilment Framework for Enhancing User Experiences in Conversational XAI. *arXiv* **2024**, arXiv:2405.10446. [CrossRef]
21. Malandri, L.; Mercorio, F.; Mezzanzanica, M.; Nobani, N. ConvXAI: A system for multimodal interaction with any black-box explainer. *Cogn. Comput.* **2023**, *15*, 613–644. [CrossRef]
22. Mindlin, D.; Robrecht, A.S.; Morasch, M.; Cimiano, P. Measuring User Understanding in Dialogue-Based XAI Systems. In Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024), Including PAIS 2024, Santiago de Compostela, Spain, 19–24 October 2024; Frontiers in Artificial Intelligence and Applications; Volume 392, pp. 1148–1155. [CrossRef]
23. Berman, A. Argumentative Dialogue As Basis For Human-AI Collaboration. In Proceedings of the HHAI 2024 Workshops, Malmö, Sweden, 10–11 June 2024.
24. Berman, A.; Larsson, S. Assessing Conversational Capabilities of Explanatory AI Interfaces. In *Proceedings of the International Conference on Human-Computer Interaction, Gothenburg, Sweden, 22–27 June 2025*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 3–21.
25. Wang, D.; Yang, Q.; Abdul, A.; Lim, B.Y. Designing theory-driven user-centric explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland, UK, 4–9 May 2019; pp. 1–15.
26. Ehsan, U.; Wintersberger, P.; Liao, Q.V.; Mara, M.; Streit, M.; Wachter, S.; Riener, A.; Riedl, M.O. Operationalizing human-centered perspectives in explainable AI. In Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, Originally, Yokohama, Japan, 8–13 May 2021; pp. 1–6.
27. Liao, Q.V.; Varshney, K.R. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv* **2022**, arXiv:2110.10790. [CrossRef]
28. Kim, J.; Maathuis, H.; Sent, D. Human-centered evaluation of explainable AI applications: A systematic review. *Front. Artif. Intell.* **2024**, *7*, 1456486. [CrossRef] [PubMed]
29. Shneiderman, B. *Human-Centered AI*; Oxford University Press: Oxford, UK, 2022.
30. Capel, T.; Brereton, M. What is human-centered about human-centered AI? A map of the research landscape. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–23.
31. Booshehri, M.; Buschmeier, H.; Cimiano, P. A Model of Factors Contributing to the Success of Dialogical Explanations; ICMI '24. In Proceedings of the 26th International Conference on Multimodal Interaction, San José, Costa Rica, 4–8 November 2024; pp. 373–381. [CrossRef]
32. Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Oahu, HI, USA, 25–30 April 2020; pp. 1–15.
33. Toulmin, S.E. *The Uses of Argument*; Cambridge University Press: Cambridge, UK, 2003.
34. Kuhlen, A.K.; Brennan, S.E. Language in dialogue: When confederates might be hazardous to your data. *Psychon. Bull. Rev.* **2013**, *20*, 54–72. [CrossRef] [PubMed]
35. John, O.P.; Srivastava, S. The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In *Handbook of Personality: Theory and Research*, 2nd ed.; Pervin, L.A., John, O.P., Eds.; Guilford Press: New York, NY, USA, 1999; pp. 102–138.
36. Melchiorre, A.B.; 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, Genoa, Italy, 12–18 July 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 313–317.

37. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media: Sebastopol, CA, USA, 2009.
38. Jönsson, A.; Dahlbäck, N. Distilling dialogues—A method using natural dialogue corpora for dialogue systems development. In *Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, WA, USA, 29 April–4 May 2000*; Association for Computational Linguistics: Stroudsburg, CA, USA, 2000; pp. 44–51.
39. Larsson, S.; Santamarta, L.; Jönsson, A. Using the process of distilling dialogues to understand dialogue systems. In Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP2000/INTERSPEECH2000), Beijing, China, 16–20 October 2000; pp. 374–377.
40. Larman, C.; Basili, V.R. Iterative and incremental developments: A brief history. *Computer* **2003**, *36*, 47–56. [CrossRef]
41. Strauss, A.; Corbin, J. *Basics of Qualitative Research*; Sage Publications: Thousand Oaks, CA, USA, 1990.
42. Larsson, S. *Issue-Based Dialogue Management*; University of Gothenburg: Gothenburg, Sweden, 2002.
43. Ginzburg, J. *The Interactive Stance*; Oxford University Press: New York, NY, USA, 2012.
44. Maraev, V.; Bernardy, J.P.; Ginzburg, J. Dialogue management with linear logic: The role of metavariables in questions and clarifications. *Trait. Autom. Des Langues* **2020**, *61*, 43–67.
45. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef]
46. Hempel, C.G.; Oppenheim, P. Studies in the Logic of Explanation. *Philos. Sci.* **1948**, *15*, 135–175. [CrossRef]
47. Larsson, S.; Myrendal, J. Dialogue acts and updates for semantic coordination. In Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue, Saarbrücken, Germany, 15–17 August 2017; pp. 59–66.
48. Breitholtz, E. *Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation*; Brill: Leiden, The Netherlands, 2020. [CrossRef]
49. Berman, A.; Gregoromichelaki, E.; Parai, C. From Interpretability to Clinically Relevant Linguistic Explanations: The Case of Spinal Surgery Decision-Support. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence—Volume 1: IAI, Porto, Portugal, 23–25 February 2025*; INSTICC: Lisbon, Portugal, 2025; pp. 909–920. [CrossRef]
50. Ducrot, O. Topoï et formes topiques. *Bull. D'études Linguist. FrançAise* **1988**, *22*, 1–14.
51. Grice, H.P. Logic and conversation. *Syntax Semant.* **1975**, *3*, 43–58.
52. Sacks, H.; Schegloff, E.A.; Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. *Language* **1974**, *50*, 696–735. [CrossRef]
53. Hosseini, S.A. Dialogues Incorporating Enthymemes and Modelling of Other Agents' Beliefs. Ph.D. Thesis, King's College, London, UK, 2016.
54. Chakraborti, T.; Kulkarni, A.; Sreedharan, S.; Smith, D.E.; Kambhampati, S. Explicability? Legibility? Predictability? Transparency? Privacy? Security? The emerging landscape of interpretable agent behavior. In Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling, Berkeley, CA, USA, 11–15 July 2019; Volume 29, pp. 86–96.
55. Bench-Capon, T.J. Specification and implementation of Toulmin dialogue game. In Proceedings of the JURIX 1998, Groningen, The Netherlands, 8–9 December 1998; Volume 98, pp. 5–20.
56. Shaheen, Q.u.a.; Toniolo, A.; Bowles, J.K.F. Dialogue Games for Explaining Medication Choices. In *Rules and Reasoning: 4th International Joint Conference, Oslo, Norway, 29 June–1 July 2020*; Gutiérrez-Basulto, V., Kliegr, T., Soylu, A., Giese, M., Roman, D., Eds.; Springer: Cham, Switzerland, 2020; pp. 97–111.
57. Prakken, H. Coherence and Flexibility in Dialogue Games for Argumentation. *J. Log. Comput.* **2005**, *15*, 1009–1040. [CrossRef]
58. Sklar, E.I.; Azhar, M.Q. Explanation through argumentation. In Proceedings of the 6th International Conference on Human-Agent Interaction, Southampton, UK, 15–18 December 2018; pp. 277–285.
59. Vassiliades, A.; Bassiliades, N.; Patkos, T. Argumentation and explainable artificial intelligence: A survey. *Knowl. Eng. Rev.* **2021**, *36*, e5. [CrossRef]
60. Feustel, I.; Rach, N.; Minker, W.; Ultes, S. Enhancing Model Transparency: A Dialogue System Approach to XAI with Domain Knowledge. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Kyoto, Japan, 18–20 September 2024*; Kawahara, T., Demberg, V., Ultes, S., Inoue, K., Mehri, S., Howcroft, D., Komatani, K., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 248–258. [CrossRef]
61. Schindler, C.; Feustel, I.; Rach, N.; Minker, W. Automatic Generation of Structured Domain Knowledge for Dialogue-based XAI Systems. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, Bilbao, Spain, 27–30 May 2025*; Torres, M.I., Matsuda, Y., Callejas, Z., del Pozo, A., D'Haro, L.F., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 1–11.
62. Peräkylä, A. Validity in Research on Naturally Occurring Social Interaction. In *Qualitative Research: Issues of Theory, Method and Practice*, 3rd ed.; Silverman, D., Ed.; Sage: London, UK, 2011; pp. 365–382.
63. Seedhouse, P. Conversation analysis as research methodology. In *Applying Conversation Analysis*; Palgrave Macmillan: London, UK, 2005; pp. 251–266.

64. Hase, P.; Bansal, M. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020*; Jurafsky, D., Chai, J., Schluter, N., Tetreault, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5540–5552. [CrossRef]
65. Kamar, E. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In Proceedings of the IJCAI, New York, NY, USA, 9–15 July 2016; pp. 4070–4073.
66. Lai, V.; Tan, C. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In Proceedings of the FAT*, Atlanta, GA, USA, 29–31 January 2019.
67. Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Krishna, R.; Gerstenberg, T.; Bernstein, M.S. When do XAI methods work? A cost-benefit approach to human-AI collaboration. In Proceedings of the CHI Workshop on Trust and Reliance in AI-Human Teams, New Orleans, LA, USA, 30 April 2022; pp. 1–15.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Reproduced with permission of copyright owner. Further reproduction prohibited without permission.